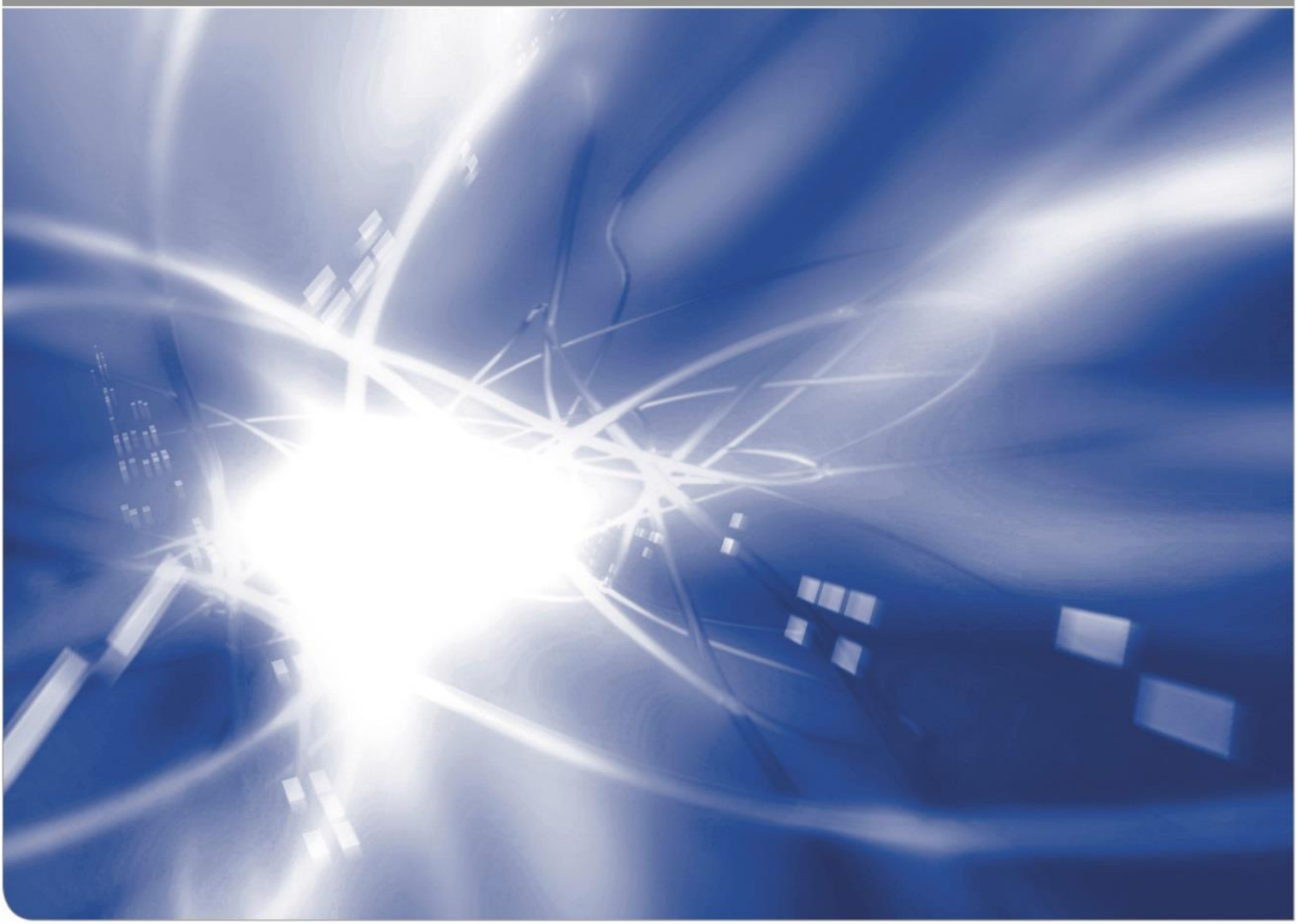
Karlsruher Institut für Technologie

Einige ethische Implikationen großer Sprachmodelle

von Reinhard Heil[1]

KIT SCIENTIFIC WORKING PAPERS 221



[1] Institut für Technikfolgenabschätzung und Systemanalyse (ITAS)

Impressum

Karlsruher Institut für Technologie (KIT)
www.kit.edu

2023

ISSN: 2194-1629

Einige ethische Implikationen großer Sprachmodelle[1]

Bevor ich, ohne Anspruch auf Vollständigkeit, auf einige ethische Implikationen der Nutzung von Anwendungen wie ChatGPT, Bard usw. eingehe, noch einige Überlegungen zu sogenannten *large language models*, also großen Sprachmodellen.

Ich werde im Folgenden **ChatGPT als pars pro toto** für alle auf *large language models* basierenden Chatbot-Anwendungen nutzen. Wichtig ist, dass Anwendungen wie ChatGPT mehr sind als das zugrundeliegende Sprachmodell. Dazu später mehr.

Kurz zu den Modellen. Das Wichtigste findet sich bereits in der Bezeichnung der Modelle: Es handelt sich um *large language models* und nicht um *large knowledge models*. Diese Unterscheidung geht in der Diskussion teils unter, sie ist aber von größter Relevanz und erklärt einige Schwächen und Stärken dieser Modelle. Sie haben wahrscheinlich schon öfters gelesen, dass sich ChatGPT auf dem Wissenstand des Jahres 2021 befände, dass es anhand von so und so vielen Milliarden Token gelernt hat usw. OpenAI macht auf seiner Webseite ähnliche Angaben. Solche Rede suggeriert, dass es sich bei den **large language models um knowledge models** handeln würde, in denen Daten oder gar Wissen direkt gespeichert sei. Als Laie gelangt man leicht zu der Überzeugung ChatGPT enthalte die Wikipedia, riesige Mengen an Artikeln, Büchern und Webseiteninhalten und nutze auf Nachfrage dieses Wissen, um eine Antwort zu geben. Das ist falsch. ChatGPT befindet sich nicht auf dem Wissensstand von 2021, sondern es bildet die statistischen Zusammenhänge der Token, also von Wörtern bzw. Wortbestandteilen, ab, mit denen es trainiert wurde. Das ist das Prinzip aller deep neural networks, das ist deren Stärke: Sie erkennen Muster in Daten. Language models erkennen Muster in den Texten mit denen sie trainiert werden und können anhand dieser Muster neue Texte produzieren. Und das können sie gut, ja sogar, seit Einführung der Transformer-Modelle, beeindruckend gut. Dass ChatGPT nicht rechnen kann oder behauptet, der Mond sei bevölkert, ist so gesehen kein Problem des Modells, sondern der Anwender, die eine, durch die Hersteller provozierte, falsche Erwartungshaltung haben. Die Qualität der Ausgabe eines language models bemisst sich vor allem daran, ob sie sprachlich korrekt ist und weniger daran, ob der produzierte Text faktisch richtig ist. Für faktisch richtige Antworten sind diese Modelle nicht der richtige Dialogpartner, eben weil sie Sprachmodelle sind und keine Wissensmodelle.

[1] Vortrag im Rahmen der Sondersitzung des HND-BW Lenkungskreises und HND-BW Expert*innenkreises »Standortbestimmung: Auswirkungen von ChatGPT auf Lehre und Studium«, Online, 24.05.2023.

Weil sie Sprachmodelle sind, sind sie prädestiniert dafür als Interfaces, als Übersetzer zwischen Mensch und Mensch, Mensch und Maschine und Maschine und Maschine eingesetzt zu werden bzw. den Dialog zwischen diesen Partnern zu ermöglichen.

Die Schnittstellenfunktion von ChatGPT kommt m.E. in der Diskussion oft zu kurz, obwohl es sich dabei um die eigentliche Stärke solcher Anwendungen handelt. Sie ermöglichen die natürlichsprachliche Kommunikation zwischen Menschen und Maschinen bzw. Anwendungen und sie können die Kommunikation zwischen Menschen unterstützen. Sie sind Interfaces, Schnittstellen, wahrscheinlich bald sogar universale Schnittstellen. Am Anfang habe ich gesagt, dass man zwischen den Sprachmodellen und den auf ihnen basierenden Anwendungen unterscheiden sollte. ChatGPT besteht vereinfacht gesagt nicht nur aus dem Sprachmodell, sondern auch aus Modulen, die die Zulässigkeit von Ein- und Ausgaben prüfen und zunehmend auch aus Verknüpfungen zu anderen Anwendungen. Die Einbindung von GPT-4 in Microsofts Suche Bing ist das prominenteste Beispiel.

“ChatGPT invented a sexual harassment scandal and named a real law prof as the accused

The AI chatbot can misrepresent key facts with great flourish, even citing a fake Washington Post article as evidence” (Verma & Oremus, 2023)

An diesem Beispiel wird schön und erschreckend deutlich, wie problematisch es ist, wenn ein *large language model* mit der Autorität eines knowledge systems versehen wird. **ChatGPT produziert, und zwar immer und ausnahmslos, fiktionale Geschichten.** Diese stimmen mal mehr mal weniger mit der Realität überein. Natürlich haben auch herkömmliche Suchmaschinen Misinformationen reproduziert, sie erzeugten sie aber nicht selbst und sie versahen sie auch nicht mit der Glaubhaftigkeit bzw. Eindeutigkeit, die mit der Textform einhergeht. Man muss also jede Ausgabe auf Richtigkeit oder zumindest Plausibilität überprüfen. Dies kann nur zum Teil automatisiert geschehen, da man dafür, und hier beißt sich die Katze in den Schwanz, entweder künstliche Intelligenzen benötigen würde, die als umfassende knowledge systems funktionieren, oder man das Wissen selbst haben muss. Mit der natürlichsprachlichen Kommunikation bzw. natürlichsprachlichen Interfaces sind dieselben Probleme verbunden wie mit der Mensch-Mensch-Kommunikation. Die Gefahr ist groß, dass es zu Missverständnissen kommt. Man sollte nicht vergessen, dass es einen Grund für formale oder formalisierte Sprachen gibt.

Ein grundlegendes ethisches Problem ist, dass die **Anbieter nicht ausreichend über die Grenzen der Anwendungen aufklären**. Es drängen sich Vergleiche zu Teslas „Autopiloten“ auf. Man sichert sich zwar mit der Zeile „Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts.” ab. Sprich man weist darauf hin, dass es sich nicht um ein verlässliches Produktivsystem handelt und an anderer Stelle wird man noch darauf hingewiesen, dass die Nutzer Versuchskaninchen sind. Das ist zwar ein übliches Verfahren, (Software gilt nicht umsonst als Bananenprodukt: Produkt reift beim Kunden), das aber bei diesen Anwendungen besonders bedenklich ist. Verantwortung lässt sich so auf die Anwender*innen abwälzen. Zudem befinden sich die großen Wettbewerber gerade in besonderer Konkurrenz und rollen deshalb beinahe ohne jegliches ethische Bedenken neue Anwendungen aus (Grant & Weise, 2023).

Es ist möglich ChatGPT mit knowledge systems zu verbinden, um so zu verhindern, dass es Misinformationen erfindet. Bekanntestes Beispiel ist die Einbindung von Wolfram Alpha in ChatGPT (Wolfram, 2023). Die Verknüpfung erlaubt es ChatGPT Fragen korrekt zu beantworten, an denen es sonst scheitert, da es nicht rechnen kann bzw. kein knowledge model ist. Dazu übersetzt es die natürlichsprachliche Eingabe in die Wolfram Alpha Language und bereitet die Ergebnisse dann in Textform auf. Natürlich gilt auch hier, dass die Übersetzung falsch sein kann, aber das Sprachmodell wird in diesem Fall korrekt als Sprachmodell genutzt.

Systeme, die auf large language models basieren werden also zunehmend als Interfaces und damit auch als Gatekeeper fungieren; ähnlich wie dies heute bereits Suchmaschinen, Empfehlungs- Priorisierung- und Filteralgorithmen tun. Wer welche Information zu sehen bekommt oder nicht, liegt heute in der sogenannten westlichen Welt zum größten Teil in der Verantwortung von privaten Unternehmen, die diese Interfaces kontrollieren. Die Systeme lassen nur bestimmte Anfragen zu, unterdrücken Informationen etc. Dies ist einerseits richtig und wichtig um ethische Standards zu gewährleisten, andererseits geschieht dies oft auf intransparente Art und Weise. Mit ChatGPT gehen alle ethischen Risiken einher, die auch von anderen Deep Neural Networks und Gatekeepersystemen bekannt sind: unter anderem Datenbias, quasi Monopole, Nutzerüberwachung, Nutzerbeeinflussung, mangelnde demokratische Kontrolle, Silencing und die Gerechtigkeitsproblematik.

Die Interface-Funktion wird sich sehr wahrscheinlich auf eine ganze Reihe von Berufen auswirken, in denen Vermittlung eine große Rolle spielt. Insbesondere Berufe, in denen die Vermittlung relativ begrenzter Informationen und/oder das Erstellen standardisierter Texte im Zentrum stehen: Rechtswesen, Bank- und Versicherungsberatung, Help-Desks, Reisebüros, viele Bürotätigkeiten. Das Erstellen von kurzen bis mittellangen Texten ist bspw. Teil der Arbeit von ca. 40 Prozent aller Beschäftigten. Ein nicht zu unterschätzendes Problem bei der automatisierten Texterstellung ist das aus der Automatisierungsforschung bekannte Problem des Übervertrauens. Je besser und zuverlässiger ein System funktioniert, desto weniger wird es kontrolliert und desto leichter werden Fehler übersehen (Merritt et al., 2019). **Im beruflichen Umfeld könnte sich in Zukunft der Schwerpunkt von der Texterzeugung auf die Textkontrolle verlagern.**

Mehr Fairness durch ChatGPT? Für alle Bildungsinstitutionen spielt Fairness eine herausragende Rolle. Zugang zu Bildung ist ein Menschenrecht und die Voraussetzung für weitere Menschenrechte, sowie für den beruflichen und finanziellen Erfolg. Das ist einer der Gründe dafür, warum im sogenannten AI-Act der Einsatz von KI im Bildungsbereich als hochriskant eingestuft wird. Begreift man ChatGPT als Lehrmittel muss dementsprechend sichergestellt sein, dass alle gleichberechtigten Zugang haben und für die Nutzung qualifiziert werden. Es gibt die Hoffnung, dass ChatGPT dazu beitragen könnte Menschen, die im bestehenden System benachteiligt sind, Bildungsressourcen zu erschließen. Dazu gehört bspw. die automatisierte Übersetzung von Texten in leichte Sprache und die Bedienung von Software in natürlicher Sprache, sowie die Unterstützung beim Lernen. Es besteht aber zugleich die Befürchtung, dass Bildungsungleichheiten sogar noch zunehmen werden, da privilegierte Schülerinnen und Schüler von neuen Medien häufig stärker profitieren, u.a. da sie oft ein größeres Vorwissen besitzen (Herzig 2014). In Deutschland korreliert der Bildungserfolg bekanntlich stark mit der sozialen Herkunft (Wößmann et al., 2023).

Englisch ist die lingua franca der akademischen Welt. Dies stellt für Menschen, die Englisch als eine additional language sprechen, also für den größten Teil der Menschheit, eine Benachteiligung gegenüber native Speakern dar. Wer in einer fremden Sprache schreibt braucht länger, kämpft mit Grammatik und Wortschatz und schafft es eventuell nicht seine Ergebnisse klar zu kommunizieren und ins richtige Licht zu rücken. Einen Artikel mit 6000 Wörter prüfen zu lassen, kostet bei SpringerNature zwischen 400$ und 700$. Sehr gutes

Englisch kann man, wenn man es sich leisten kann, kaufen. Die sprachliche Qualität eines Artikels wirkt sich auf die Rezeption aus. **Wenn man das akademische Publizieren primär als Wissenstransfer versteht, stellt sich die Frage, warum es eine Rolle spielen sollte, wie eine wissenschaftliche Erkenntnis verschriftlicht wurde**. ChatGPT ist ein Instrument, das dazu genutzt werden kann, Artikel schneller und eventuell in bessere Qualität zu erstellen. Es ergänzt damit eine ganze Reihe von Werkzeugen, die schon länger genutzt werden, wie Autokorrektur, Literaturverwaltungen, Suchmaschinen, automatische Übersetzungen und Grammatikhilfen (Staiman, 2023). Anders gelagert ist die Frage, inwiefern insbesondere in der Schule und im Studium das eigenständige Verfassen resp. Schreiben von Texten zum Verständnis- bzw. Lernprozess beiträgt (Baron, 2023; Chiang, 2023). Zum Vergleich bietet sich hier die Nutzung von Taschenrechnern im Mathematikunterricht an.

Die Diskussion im akademischen Bereich zeigt m.E. vor allem, dass Anwendungen wie ChatGPT gar nicht das eigentliche Problem sind, sondern das akademische System selbst. Man bestraft den Boten für das Überbringen der Botschaft. Missbrauch oder Fehlnutzung gedeihen immer dort besonders gut, wo es ein Anreizsystem gibt. Literatursuchmaschinen gekoppelt mit Literaturverwaltungen haben sicher dazu beigetragen, dass heute drei oder vier Quellen, die nicht gelesen wurden, als Beleg angeführt werden, statt einer einzigen. Sie haben aber das Problem nicht geschaffen. Dass es ChatGPT Papermills erlaubt ihren Ausstoß noch weiter zu erhöhen mag stimmen. Dass deren Produkte nachgefragt werden, liegt jedoch am Anreizsystem. Dass Peer-Reviews mit ChatGPT geschrieben werden und deshalb u.a. falsche Artikelhinweise enthalten, überrascht wohl auch niemanden. **Technik ist eigentlich nie alleinige Verursacherin gesellschaftlicher Probleme, genauso wenig wie sie alleine gesellschaftliche Probleme lösen kann.** Das heißt nicht, dass man sie nicht regulieren muss, man sollte aber nie die eigentlichen Probleme aus dem Auge verlieren.

Literatur

Herzig, B. (2014). Wie wirksam sind digitale Medien im Unterricht? Bielefeld: Bertelsmann Stiftung. Verfügbar unter http://www.digitalisierung-bildung.de/wp-content/uploads/2014/11/DigitaleMedienUnterricht_final.pdf

Baron, N. S. (2023, January 19). How ChatGPT robs students of motivation to write and think for themselves. *The Conversation*.

Chiang, T. (2023, February 9). ChatGPT Is a Blurry JPEG of the Web. *The New Yorker*.

Grant, N., & Weise, K. (2023, April 7). In A.I. Race, Microsoft and Google Choose Speed Over Caution - The New York Times. *New York Times*.

Herzig, B. (2014). Wie wirksam sind digitale Medien im Unterricht? Bielefeld: Bertelsmann Stiftung. Verfügbar unter http://www.digitalisierung-bildung.de/wp-content/uploads/2014/11/DigitaleMedienUnterricht_final.pdf

Merritt, S. M., Ako-Brew, A., Bryant, W., Staley, A., McKenna, M., Leone, A., Shirase, L. (2019). „Automation-Induced Complacency Potential: Development and Validation of a New Scale.“ Frontiers in Psychology 10.

Staiman, A. (2023, March 31). *Guest Post — Academic Publishers Are Missing the Point on ChatGPT*. The Scholarly Kitchen. https://scholarlykitchen.sspnet.org/2023/03/31/guest-post-academic-publishers-are-missing-the-point-on-chatgpt

Verma, P., & Oremus, W. (2023, April 5). What happens when ChatGPT lies about real people? - The Washington Post. *Washington Post*.

Wolfram, S. (2023, March 23). ChatGPT Gets Its “Wolfram Superpowers”!—Stephen Wolfram Writings. https://writings.stephenwolfram.com/2023/03/chatgpt-gets-its-wolfram-superpowers/

Wößmann, L., Schoner, F., Freundl, V., & Pfaehler, F. (2023). Der ifo-„Ein Herz für Kinder“-Chancenmonitor: Wie (un-)gerecht sind die Bildungschancen von Kindern aus verschiedenen Familien in Deutschland verteilt? *Ifo Schnelldienst*, *76*(3), 29–47. https://www.ifo.de/publikationen/2023/aufsatz-zeitschrift/der-ifo-ein-herz-fuer-kinder-chancenmonitor

KIT Scientific Working Papers
ISSN 2194-1629

www.kit.edu