Deutscher Bundestag
Ausschuss für Bildung, Forschung und Technikfolgenabschätzung
Ausschussdrucksache
20(18)109
21.04.2023

BÜRO FÜR TECHNIKFOLGEN-ABSCHÄTZUNG
BEIM DEUTSCHEN BUNDESTAG

Steffen Albrecht

ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen

21. April 2023
Hintergrundpapier Nr. 26

Steffen Albrecht

ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen

TAB-Hintergrundpapier Nr. 26

Büro für Technikfolgen-Abschätzung
beim Deutschen Bundestag
Neue Schönhauser Straße 10
10178 Berlin

Telefon: +49 30 28491-0
E-Mail: buero@tab-beim-bundestag.de
Web: www.tab-beim-bundestag.de

2023

Umschlagbild: Teresa Berndtsson/Better Images of AI/Letter Word Text Taxonomy/CC-BY 4.0

ISSN-Internet: 2199-7136

Das Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB) berät das Parlament und seine Ausschüsse in Fragen des wissenschaftlich-technischen Wandels. Das TAB wird seit 1990 vom Institut für Technikfolgenabschätzung und Systemanalyse (ITAS) des Karlsruher Instituts für Technologie (KIT) betrieben. Hierbei kooperiert es seit September 2013 mit dem IZT – Institut für Zukunftsstudien und Technologiebewertung gGmbH sowie der VDI/VDE Innovation + Technik GmbH.

Inhalt

Zusammenfassung

Worum geht es?

Selten hat ein Computersystem weltweit so viel Aufmerksamkeit und Debatten erregt wie ChatGPT seit seiner Einführung im November 2022. Der Chatbot beruht auf einem Computermodell, das mithilfe von Methoden der künstlichen Intelligenz (KI) auf die Verarbeitung sprachlicher Daten trainiert wurde. Er kann in kürzester Zeit eloquent erscheinende Antworten zu den unterschiedlichsten Themen generieren, ganze Essays oder Computerprogramme erstellen und Sprachstile wie Gedichte, Witze oder Erörterungen verwenden – und das in verschiedenen Sprachen.

Die große öffentliche Aufmerksamkeit schürt einerseits hohe Erwartungen in Bezug auf die Anwendungsmöglichkeiten. Sie kann andererseits aber auch den Blick auf eine realistische Einschätzung der Möglichkeiten und Grenzen solcher Systeme sowie ihre gesellschaftlichen Auswirkungen verstellen. Als Orientierung für die laufende Debatte werden in diesem Hintergrundpapier

- die technologischen Entwicklungen, auf denen das System beruht,
- die Möglichkeiten und Grenzen der Technologie,
- potenzielle Anwendungen, insbesondere im Bereich der Bildung, sowie
- mögliche Auswirkungen einer Anwendung

dargestellt. Ziel des Papiers ist es, zu diesen Aspekten fundierte Informationen zusammenzustellen und Fragestellungen zu identifizieren, unter denen die Rolle von sprachverarbeitenden Computermodellen weiter beobachtet und diskutiert werden kann. Im Folgenden werden die wichtigsten Ergebnisse in Form einer Leseanleitung zusammengefasst.

Ein technologischer Durchbruch bei der Sprachverarbeitung …

ChatGPT stellt einen Durchbruch bei der Verarbeitung von Sprache mit KI-Methoden dar, der vor allem auf zwei Entwicklungen beruht (Kap. 2): Neue Formen künstlicher neuronaler Netzwerke, sogenannte Transformermodelle, ermöglichen erstens die besonders effiziente Umwandlung von Sprache in mathematische Parameter. Dadurch können zweitens die Komplexität dieser Computermodelle und die Menge der für ihr Training verwendeten Daten enorm vergrößert werden. Das ChatGPT zugrundeliegende Modell umfasst 175 Mrd. Parameter und beruht auf einem Trainingsmaterial von 300 Mrd. Textbestandteilen. Bei seiner Veröffentlichung gehörte es zu den größten Modellen seiner Art; mit dem Nachfolgemodell GPT-4 sowie Modellen anderer Unternehmen wurden inzwischen noch größere Systeme entwickelt.

Das Training der Modelle erfolgt in zwei Phasen: In einem ersten Schritt liest das System weitgehend eigenständig (unüberwacht) große Mengen an Texten ein und bildet daraus seine Parameter. Im zweiten Schritt erfolgt mithilfe von menschlichem Feedback eine Feinjustierung des Modells auf seine spezielle Aufgabe, im Fall von ChatGPT die Dialogfähigkeit zu beliebigen Themen. Da zum Training vor allem menschlich erzeugte Texte aus dem Internet genutzt werden, beruhen beide Schritte wesentlich auf (zum größten Teil unentgeltlicher bzw. vergleichsweise gering entlohnter) menschlicher Arbeit. Für das Training ist eine hoch performante Hardware in spezialisierten Rechenzentren erforderlich, die einen hohen Energieverbrauch mit sich bringt.

… eröffnet neue Möglichkeiten und noch nicht in ihrer Tragweite absehbare Entwicklungen

Dank des Trainings anhand von Millionen Dokumenten können die sprachverarbeitenden Computermodelle ganz unterschiedliche Textsorten überzeugend imitieren (Kap. 3.1). Ihre Fähigkeit, in Dialogen zu einer großen Bandbreite von Themen zu interagieren, übersteigt bisherige Chatbot-Entwicklungen. Die Systeme sind in der Lage, aus kurzen textlichen Eingaben (Prompts) eine Aufgabe zu erschließen und in natürlicher Sprache zu antworten. So können sie mit hoher Geschwindigkeit Aufgaben erledigen, die bislang Menschen vorbehalten waren, wie beispielsweise die Analyse langer Texte, das Erstellen von Plänen (für Reisen, Unternehmensstrategien oder Lehrveranstaltungen) nach vorgegebenen Schemata oder die Erledigung von Haus- und Prüfungsaufgaben.

Aus diesen Fähigkeiten ergibt sich perspektivisch eine Reihe von Einsatzmöglichkeiten:

- Menschliche Tätigkeiten, die mit der Verarbeitung von Texten (einschließlich Programmcode) verbunden sind, lassen sich zumindest teilweise automatisieren. Durch die Erweiterung der Fähigkeiten auf Bilder und Töne sind auch crossmediale Anwendungen zur Automatisierung etwa im Journalismus vorstellbar (Kap. 4.3.1).
- Entlastungs- bzw. Rationalisierungseffekte sind auch in solchen Branchen möglich, die von Automatisierung durch Informations- und Kommunikationstechnologien bislang kaum betroffen waren.
- Da die KI-Systeme in natürlicher Sprache angesprochen werden können, könnten sie als leicht bedienbare Schnittstelle für andere Computersysteme genutzt werden (Kap. 3.1.3).
- Der Einsatz von Chatbots könnte die Verständigung über Sprachen hinweg fördern und zur Inklusion beitragen, indem die Kosten für die Übersetzung bzw. Vereinfachung von Texten verringert werden (Kap. 4.1.1 sowie 4.4.2).

Systembedingte Grenzen und mögliche negative Auswirkungen

Computermodelle zur Sprachverarbeitung weisen system- bzw. architekturbedingte Grenzen auf. Ihr Output kann nur so gut sein wie das, was sie an Input erhalten haben. Dies betrifft zum einen die Prompts der menschlichen Nutzer/innen (Kap. 3.1.2 sowie 3.2.1), zum anderen die im Training verwendeten Daten. Ein in den Trainingsdaten enthaltener Bias, also eine verzerrte Repräsentation bestimmter Kategorien, kann sich in den Antworten des Systems widerspiegeln und Diskriminierung verstärken (Kap. 4.4.3).

Die vom System erzeugten Informationen sind zudem häufig faktisch falsch (Kap. 3.2.2). Aufgrund der hohen sprachlichen Qualität und in Ermangelung von Belegen lässt sich die Korrektheit jedoch nur schwer überprüfen, wodurch das Vertrauen in die Verlässlichkeit von Informationen im Allgemeinen sinken kann (Kap. 4.3.1). Dieser Effekt kann durch den Automation Bias noch verstärkt werden (Kap. 4.2.2). Damit ist die menschliche Tendenz gemeint, die Ergebnisse maschineller Verarbeitung unkritisch zu akzeptieren und nach ihnen zu handeln. Generell ist eine verantwortliche Anwendung auf solche Gebiete beschränkt, in denen die Nutzenden die Qualität der Ergebnisse beurteilen können bzw. die faktische Richtigkeit weniger bedeutsam ist.

Folgende konkrete Risiken werden aktuell diskutiert: Sowohl in der privaten als auch der öffentlichen Kommunikation wird eine Zunahme von (nicht als solche erkennbaren) computergenerierten Texten erwartet (Kap. 4.3.1). Dadurch könnte sich der Wert einzelner Informationen verringern (Kap. 5.2.2). Sprachverarbeitende KI-Systeme könnten auch für unerwünschte Werbung (Spam, Kap. 4.1.2), effektivere Angriffe auf die Computersicherheit (Phishing-Attacken, Kap. 3.2.5) sowie gezielte Desinformation genutzt werden (Kap. 4.3.1). Größere Mengen solcher bewusst schädigend eingesetzter Texte könnten das Vertrauen in den öffentlichen Diskurs als Ort demokratischer Meinungsbildung unterminieren. Gegenmaßnahmen, wie die automatisierte Erkennung von KI-generierten Texten oder die Verbesserung von Medienkompetenzen, gelten als bislang nicht ausreichend wirksam.

Ein weiteres Risiko besteht darin, dass durch die Nutzung Ungleichheiten verstärkt werden (Kap. 4.1.3 sowie 5.1.2). Die effektive Verwendung von Computermodellen zur Sprachverarbeitung setzt Kompetenzen voraus, die ungleich verteilt und nur unter bestimmten Voraussetzungen zu erwerben sind. Zudem ist damit zu rechnen, dass die finanziellen Kosten der Nutzung steigen (Kap. 2.5).

Vielfältige Anwendungsperspektiven

In welchem Verhältnis Chancen und Risiken bei der Nutzung von ChatGPT und vergleichbaren Systemen stehen, lässt sich nur im Kontext konkreter Anwendungen beurteilen. Aufgrund der Neuheit und der vielseitigen Möglichkeiten

der Systeme ist noch nicht absehbar, wo und wie sie letztlich eingesetzt werden. Die folgenden Anwendungsbereiche stehen im Mittelpunkt der bisherigen Debatte:

- In Unternehmen ergeben sich Anwendungen in der internen wie externen Kommunikation, z. T. auch der Automatisierung von Arbeitsprozessen. Als Szenario wird die Verknüpfung eines KI-basierten Chatbots mit Officeanwendungen dargestellt (Kap. 4.1).
- Gesundheitsbezogene Anwendungen können im Angebot von Gesundheitsinformationen, in der Kommunikation zwischen Ärzt/innen und Patient/innen sowie in der Verarbeitung unstrukturierter Gesundheitsdaten bestehen. Als Szenario wird der Einsatz von Chatbots für Menschen mit psychischen Problemen diskutiert (Kap. 4.2).
- Im Bereich Information und Kommunikation werden zwei Szenarien vorgestellt: die Automatisierung journalistischer Tätigkeiten sowie die Informationssuche mithilfe von Chatbots anstelle von klassischen Suchmaschinen (Kap. 4.3).
- Im Rechtswesen und der öffentlichen Verwaltung bestehen hohe Anforderungen an die Verlässlichkeit der Systeme. Als Szenario wird die Nutzung von Chatbots in der Kommunikation mit Bürger/innen diskutiert (Kap. 4.4).

Im Bereich Bildung und Forschung ergeben sich vielfältige Anwendungsmöglichkeiten, die aktuell besonders intensiv diskutiert und daher in einem eigenen Kapitel behandelt werden. Der Hoffnung des Lehrpersonals auf Entlastung von Routineaufgaben und eine Erweiterung didaktischer Möglichkeiten stehen unter anderem Befürchtungen eines Verlusts von Bildungskompetenzen, missbräuchlicher Verwendungen in Prüfungen und Datenschutzbedenken gegenüber (Kap. 5.1.1 bis 5.1.3). In der Forschung eröffnen die KI-Modelle zur Sprachverarbeitung neue Ansätze etwa in den Sozialwissenschaften oder auch der Genforschung. Der fehlende Bezug der Modelldaten zu den Quellen schränkt jedoch die Anwendungsmöglichkeiten ein. Zudem ist eine Zunahme an Publikationen zu erwarten, die mithilfe von Sprachmodellen erstellt wurden, was das Wissenschaftssystem vor neue Herausforderungen stellen dürfte (Kap. 5.2).

Wie lässt sich die Entwicklung von KI-Systemen in den Blick nehmen? Fragen und Herangehensweisen

Die Veröffentlichung von ChatGPT hat große Aufmerksamkeit auf die Potenziale und Risiken sprachverarbeitender Computermodelle gelenkt. Weil die möglichen positiven wie negativen Auswirkungen als sehr tiefgreifend erscheinen, richtet sich der Blick auch auf die Bedingungen, unter denen KI-Systeme entwickelt und auf den Markt gebracht werden. Während die Bedeutung öffentlicher Forschungseinrichtungen dabei immer weiter abnimmt, konzentriert sich

die Entwicklung auf wenige, finanziell bzw. hardwaretechnisch leistungsstarke private Unternehmen (Kap. 2.5).

Vor diesem Hintergrund gewinnen Ansätze der Regulierung an Bedeutung, die unerwünschten Entwicklungen Schranken setzen, ohne dabei Innovationen zu behindern. Staatenübergreifende Initiativen wie das geplante Gesetz über Künstliche Intelligenz der Europäischen Union scheinen dafür besonders geeignet. Außerdem werden Ansatzpunkte für eine verantwortungsvolle Forschung und Entwicklung im Bereich sprachverarbeitender KI-Systeme diskutiert (Kap. 7.1 bis 7.2).

Die große Geschwindigkeit der bisherigen Entwicklung und die Bandbreite möglicher Anwendungsgebiete und Auswirkungen der Technologie stellen nicht nur die Technikfolgenabschätzung, sondern auch die politischen Institutionen vor Herausforderungen. Die rege öffentliche Debatte, die durch die Einführung von ChatGPT ausgelöst wurde, liefert nicht nur Anhaltspunkte für die Entwicklung von Handlungsoptionen (Kap. 7.3). Sie kann auch als ermutigendes Zeichen der gesellschaftlichen Reaktionsfähigkeit auf disruptiv anmutende Innovationen angesehen werden.

1 Welche Fragen stellen sich infolge der Entwicklung großer sprachverarbeitender Computermodelle?

Auch wenn nicht alle jemals veröffentlichten Werke überprüft werden können, dürfte es wohl einmalig in der Geschichte sein, dass innerhalb weniger Monate

- eine Vielzahl von Kolumnen für Onlinemedien,
- mindestens drei wissenschaftliche Zeitschriftenaufsätze und ein Editorial einer wissenschaftlichen Zeitschrift,
- eine Rede im Europäischen Parlament sowie Reden in mehreren deutschen Landtagen,
- eine Antwort auf eine schriftliche Frage einer Bundestagsabgeordneten,
- eine Weihnachtspredigt sowie
- eine Folge der Serie Southpark

im Namen von ein und demselben Urheber (oder Urheberin?) veröffentlicht wurden.[1] Ein Verlag für Science-Fiction-Literatur wurde von Manuskripten regelrecht überschwemmt (Hern 2023), ein Onlineforum für Programmierer musste angesichts der Menge an Beiträge neue Moderationsregeln einführen (Kahn 2023).

1 Beiträge in Onlinemedien: www.spiegel.de/netzwelt/web/chatgpt-markiert-das-ende-der-irrelevanten-kuenstlichen-intelligenz-kolumne-a-b2afeb69-083d-4e69-8920-da5cad549d5f (19.4.2023); www.heise.de/meinung/Die-Gefahren-und-Chancen-der-KI-Warum-sorgfaeltiger-Einsatz-entscheidend-ist-7369569.html (19.4.2023); wissenschaftliche Arbeiten: www.medrxiv.org/content/10.1101/2022.12.19.22283643v2 (19.4.2023, in der Veröffentlichung dieses Preprints in der Zeitschrift PLOS Digital Health wird ChatGPT nicht mehr als Autor/in genannt); www.sciencedirect.com/science/article/abs/pii/S1471595322002517 (19.4.2023); www.oncoscience.us/article/571/text/ (19.4.2023); www.tandfonline.com/doi/full/10.1080/14703297.2023.2190148 (19.4.2023, in diesem Fall wurde die Mitwirkung von ChatGPT zunächst gegenüber den Reviewern der Zeitschrift verschwiegen, in der Veröffentlichung wird sie nur in der Rubrik »Acknowledgments« erwähnt); Rede im Europäischen Parlament am 1.2.2023: https://twitter.com/woelken/status/1621536645513187329 (19.4.2023); Rede im Landtag von Baden-Württemberg am 15.12.2022: www.esslinger-zeitung.de/inhalt.ki-geschriebene-rede-im-landtag-fuer-die-gruenen-spricht-kollege-chatgpt.2c50429a-bef3-4702-ba69-58d1445c26a5.html (19.4.2023); Rede im Schleswig-Holsteinischen Landtag am 27.1.2023: www.landtag.ltsh.de/presseticker/2023-01-27-13-25-13-59e4/ (19.4.2023); Rede in der Hamburgischen Bürgerschaft am 1.2.2023: www.ndr.de/nachrichten/hamburg/ChatGPT-Kuenstliche-Intelligenz-schrieb-Rede-fuer-Buergerschaft,chatgpt124.html (19.4.2023); Antwort auf die schriftliche Frage der Abgeordneten Nicole Gohlke (Die Linke) (S. 86f.): https://dserver.bundestag.de/btd/20/056/2005694.pdf (19.4.2023, auch in diesem Fall wurde die Mitwirkung von ChatGPT erst im Nachhinein bekanntgegeben); Weihnachtspredigt: https://eulemagazin.de/10-jahre-bullshit/ (19.4.2023); Southpark Staffel 26, Folge 4 vom 13.3.2023: www.southpark.de/folgen/8byci4/south-park-kuenstliche-intelligenz-staffel-26-ep-4 (19.4.2023).

Die Quelle dieser unerhörten und ausgesprochen vielfältigen Produktion von Texten ist *ChatGPT*, ein Computersystem auf Basis von Künstlicher Intelligenz (KI), das am 30. November 2022 vom Unternehmen OpenAI öffentlich zugänglich gemacht wurde.[2] Innerhalb kürzester Zeit haben Millionen Menschen das System ausprobiert,[3] seine Möglichkeiten und Grenzen abzustecken versucht und darüber publiziert und diskutiert. Auch Expert/innen zeigen sich von den Fähigkeiten des Systems beeindruckt. ChatGPT hat mittlerweile Eingang in eine Reihe von Produkten gefunden, konkurrierende Unternehmen wie Google führten eigene, vergleichbare Systeme ein, und mit GPT-4 ist am 14. März 2023 ein Nachfolgesystem veröffentlicht worden, das noch leistungsfähiger sein soll.[4]

Mit der Veröffentlichung von ChatGPT kann die Öffentlichkeit beinahe greifbar erfahren, *wie sich die KI-Technologie entwickelt hat*, an der in den Forschungslaboren insbesondere von Unternehmen (Ahmed et al. 2023) gearbeitet wird. ChatGPT beruht auf einem Transformer, einem besonderen Computermodell zur Verarbeitung und Erzeugung von Sprache. Transformermodelle werden seit 2017 eingesetzt und gelten als Durchbruch in der Forschung zu künstlichen neuronalen Netzen, weil sie es erlauben, besonders große Datenmengen bei deren Training zu verarbeiten (Hutson 2021). Mit ihrer Hilfe können Computersysteme Texte und auch Bilder erzeugen, die sich nicht ohne Weiteres von menschlichen Erzeugnissen unterscheiden lassen. Zudem sind die Systeme in der Lage, aus Anweisungen in natürlicher Sprache Aufgabenstellungen und Inhalte zu interpretieren und auf diese zu reagieren, sodass beispielsweise mit ChatGPT Konversationen wie mit einem/r menschlichen Gesprächspartner/in möglich sind.

ChatGPT hat bereits *reale Auswirkungen*. Viele Menschen nutzen das System als Werkzeug, um damit ihre Ziele besser zu erreichen (CAIS 2023, Grimm 2023, Horizont Online/dpa 2023) – auch böswillige Ziele lassen sich damit verfolgen (Tamkin et al. 2021). ChatGPT und GPT-4 wurden zwar mit einer Reihe von Sicherheitsvorkehrungen ausgestattet, um missbräuchliche Verwendungen zu verhindern; es ist aber nicht absehbar, inwieweit diese ausreichen. An Schulen und Hochschulen hat sich eine besonders intensive Debatte über den Umgang mit ChatGPT entwickelt. Diskutiert wird, wie sich Systeme wie ChatGPT didaktisch sinnvoll nutzen lassen, welche Prüfungsformen hinfällig werden könnten, weil Prüflinge sie mithilfe von ChatGPT ohne nennenswerte Eigenlei-

2 https://chat.openai.com (19.4.2023).

3 Innerhalb von fünf Tagen nach der Veröffentlichung am 30. November 2022 meldeten sich 1 Mio. Nutzende an, am 2. Februar 2023 wurden 100 Mio. aktive Nutzende/Monat gemeldet (www.tagesschau.de/wirtschaft/technologie/chatpgt-bezahl-abo-ki-101.html, 19.4.2023). Ein solches Wachstum war noch bei keinem Onlinedienst beobachtet worden; bei TikTok beispielsweise wurde die Schwelle von 100 Mio. aktiven Nutzenden nach etwa neun Monaten erreicht, bei Instagram nach zweieinhalb Jahren (Bünte/dpa 2023).

4 https://openai.com/product/gpt-4 (19.4.2023).

stung bestehen können, und – wie bereits bei der Frage digitaler Medien in der Bildung (TAB 2016a, S. 38ff.) – welche neuen Kompetenzen für den sowohl individuellen wie gesellschaftlichen Umgang mit diesen KI-Systemen zu vermitteln sind.

Das vorliegende Hintergrundpapier soll in dieser Debatte *Orientierung* liefern und helfen, den Blick auf Aspekte zu richten, die über den derzeitigen Hype hinaus bedeutsam sind. Technologische Hypes sind aus der Technikgeschichte wohlbekannt (van Lente et al. 2013; TAB 2011). Sie verschaffen einer Technologie Aufmerksamkeit und sorgen für öffentliche Debatten, sind allerdings oft auch durch überzogene Erwartungen gekennzeichnet und haben Enttäuschungen zur Folge. Zuletzt war dies beispielsweise bei der Debatte um Sprachassistenten wie Alexa zu beobachten, die den in sie gesteckten Erwartungen nicht gerecht wurden (Schiffer 2022). Die Dynamik eines Hypes kann aber auch ablenken von einer Betrachtung der Umstände, unter denen eine Technologie entwickelt und eingesetzt wird, oder Aufmersamkeit abziehen von anderen, langfristig effektiveren Technologien (Vinsel 2021).

Vor dem Hintergrund der noch jungen und sehr dynamischen Entwicklung richtet sich dieses Papier weniger auf die Formulierung von konkreten Handlungsoptionen. Ziel ist vielmehr, *Fragestellungen abzuleiten*, unter denen die Rolle von KI-Modellen zur Sprachverarbeitung weiter beobachtet und untersucht werden kann. Der Fokus liegt dabei gemäß dem Auftrag durch die Berichterstattergruppe TA des Ausschusses für Bildung, Forschung und Technikfolgenabschätzung auf ChatGPT (in der über die Website kostenlos nutzbaren Standardvariante), wobei auch das Nachfolgermodell GPT-4 und technologisch verwandte, auf die Verarbeitung von Sprache ausgerichtete Modelle berücksichtigt werden. Diese KI-Modelle werden zunächst mit Blick auf ihre technologischen Grundlagen im Kontext der aktuellen und langfristigen Entwicklungen im Bereich der KI-Forschung beschrieben (Kap. 2). Anschließend werden ihre Möglichkeiten und Grenzen dargestellt, soweit sie sich technologisch ableiten lassen oder in der bisherigen Nutzung gezeigt haben (Kap. 3). Mögliche Anwendungsmöglichkeiten der Systeme werden überblicksartig auf Grundlage der öffentlichen und fachlichen Diskussionen skizziert (Kap. 4), ein besonderer Fokus liegt dabei auf Anwendungen im Bildungsbereich (Kap. 5). Die mit diesen Anwendungsmöglichkeiten einhergehenden Chancen und Risiken werden dabei jeweils diskutiert. Den Abschluss bildet ein Fazit und ein Überblick und eine Diskussion der Fragestellungen, die sich bezügliche der Auwirkungen von ChatGPT und vergleichbaren Systemen ergeben (Kap. 6).

Eine Befassung mit einem so dynamischen Thema aus der Perspektive der eigentlich auf längerfristige Entwicklungen ausgerichteten Technikfolgenabschätzung steht *methodisch vor einigen Herausforderungen*, die mit Abweichungen von dem sonst in TAB-Untersuchungen üblichen Vorgehen verbunden waren. So konnten in der Kürze der Zeit keine externen Gutachten beauftragt

werden. Zur Informationsgewinnung wurde vor allem Literaturrecherche betrieben, außerdem wurden zwei Onlineworkshops mit TA-Expert/innen (am 1.2.2023 mit Mitarbeitenden des ITAS, am 15.3.2023 mit Kolleg/innen des EPTA-Netzwerks) durchgeführt sowie mehrere Online- und Präsenzveranstaltungen zu ChatGPT besucht. Der Stand der wissenschaftlichen Literatur zu ChatGPT ist außerhalb technischer Fragestellungen noch nicht weit entwickelt, daher wurden intensiver als sonst üblich auch Quellen in sozialen Netzwerken und Onlinemedien berücksichtigt. Im Vordergrund stand dabei weniger die gesicherte wissenschaftliche Evidenz als vielmehr – im Sinn eines Crowdsourcingansatzes (TAB 2017b, S. 54f.) – das Ziel, möglichst schnell viel Wissen und Erfahrungen zu gewinnen sowie kritische Fragen zu sammeln. Dabei wurden Informationen bis zum Stand vom 24. März 2023 berücksichtigt, nur in Ausnahmefällen konnten spätere Informationen ergänzt werden.

An der Recherche sowie an der Bearbeitung und Gestaltung dieses Hintergrundpapiers war eine Reihe von Kolleg/innen des TAB, seiner Konsortialpartner, der europäischen Partnerinstitutionen im EPTA-Netzwerk sowie des TAB-Betreibers, dem Institut für Technikfolgenabschätzung und Systemanalyse des Karlsruher Instituts für Technologie, beteiligt. Ihnen allen sowie auch den zahlreichen Gesprächspartner/innen, die im Zusammenhang der Untersuchung Auskunft gegeben und sich an Diskussionen beteiligt haben, sei herzlich für ihre Unterstützung gedankt. Die Verantwortung für die Darstellung in diesem Bericht liegt beim Verfasser.

Die breit und interdisziplinär angelegte Recherche soll dabei helfen, die komplexen Zusammenhänge zwischen der Technologie, den an ihrer Entwicklung Beteiligten sowie ihrer Nutzung und den von dieser Betroffenen in den Blick zu nehmen. Auf einer solchen Grundlage kann eine Abschätzung erfolgen, welche Maßnahmen möglicherweise geboten erscheinen, um die Entwicklung bzw. Nutzung an gesellschaftlichen Werten und Zielen auszurichten, und welche Optionen dafür zur Verfügung stehen.

2 Technische Grundlagen

ChatGPT wurde als Chatbot veröffentlicht, d. h. als ein Computersystem, mit dem die Nutzenden über eine Onlineschnittstelle in Textform kommunizieren können (Kohne et al. 2020; Neff/Nagy 2016). Chatbots können Anfragen dank sprachverarbeitender Technologie beantworten (TAB 2022d), dafür nutzen sie entweder regelbasierte oder KI-Systeme. Chatbots werden seit den 1960er Jahren entwickelt, typische Einsatzgebiete sind unter anderem die Beantwortung von Kundenfragen, das (Online-)Marketing, der Bürgerservice in der öffentlichen Verwaltung oder – in Kombination mit Systemen zur Verarbeitung gesprochener Sprache – Sprachassistenten wie Siri, Alexa oder Google Assistant (Rathenau 2020). Ebenfalls typisch ist, dass Chatbots in der Kommunikation wie menschliche Gesprächspartner/innen auftreten können, was im Fall von Social Bots, also Chatbots, die in sozialen Medien agieren, auch zu manipulativen Zwecken genutzt werden kann (TAB 2017c).

ChatGPT ist allerdings ein besonderer Chatbot, wie mehrere Journalisten anmerken:[5]

> *»There have been chatbots before. But not like this.«* (Kahn 2023)
>
> *»ChatGPT is, quite simply, the best artificial intelligence chatbot ever released to the general public.«* (Roose 2022)
>
> *»It's not the first AI chatbot, and it certainly won't be the last, but (…) the future is arriving.«* (Karpf 2022)

Die Besonderheit von ChatGPT liegt hauptsächlich in der KI-Technologie begründet. Das System basiert auf einem Computermodell zur Sprachverarbeitung aus der Reihe der »Generative Pre-Trained Transformer«[6] des US-amerikanischen Unternehmens OpenAI (Kap. 2.1). Dieses Modell kann seine besondere Leistung erbringen, weil es anhand von großen Datenmengen trainiert (Kap. 2.2) und durch Sicherheitsvorkehrungen und eine leicht bedienbare Nutzungsschnittstelle (Kap. 2.3) ergänzt wurde. Außerdem sind die Entwicklung und der Betrieb des Systems durch spezielle Anforderungen an die Hardware und technische Infrastruktur (Kap. 2.4) sowie durch besondere unternehmerische Aspekte (Kap. 2.5) gekennzeichnet, die im Folgenden dargestellt werden.

---

5 Die Zitate lauten in deutscher Übersetzung: »Es gab schon früher Chatbots. Aber keine wie diesen.« (Kahn 2023); »ChatGPT ist, ganz einfach, der beste KI-Chatbot, der jemals veröffentlicht wurde.« (Roose 2022); »Es ist nicht der erste KI-Chatbot, und es wird sicher nicht der letzte sein, aber (…) hier kommt die Zukunft.« (Karpf 2022).

6 Transformer ist der Name der speziellen Architektur des Modells (Kap. 2.1). »Pre-trained« bedeutet, dass die Modelle in einem ersten Trainingsschritt ein (unüberwachtes) Lernen durchlaufen haben. In diesem Grundzustand können sie alle möglichen Arten von Texten erzeugen, daher die Bezeichnung »generative«.

2.1 Große Computermodelle zur Sprachverarbeitung (large language models)

Bei den KI-Modellen, die ChatGPT und verwandten Systemen zugrunde liegen, handelt es sich um Transformer, eine Weiterentwicklung von künstlichen neuronalen Netzen. Künstliche neuronale Netze sind ein Zweig der mittlerweile technologisch und thematisch sehr vielfältigen KI-Forschung, der auf Ansätze aus den 1940er Jahren zurückgeht und durch biologische Prozesse inspiriert ist. Computersysteme sollen mithilfe dieser Technologie menschliche Fähigkeiten imitieren können. Entsprechende Anwendungen haben sich bei der Erkennung von Mustern bzw. komplexen Zusammenhängen in Daten bewährt und werden heute für so unterschiedliche Aufgaben wie die Entwicklung von Spielstrategien (z. B. bei Go), das Herausfiltern von Spam aus E-Mails, die Gesichts- bzw. allgemeine Bilderkennung, medizinische Diagnosen, die Faltung von Proteinmolekülen in der biochemischen Forschung, Übersetzungen und vieles mehr eingesetzt.

Künstliche neuronale Netze werden aus Knotenpunkten gebildet, die in spezifischer Weise miteinander verbunden werden. Die Knoten eines solchen Netzes werden durch mathematische Funktionen gebildet. Diese Knoten sind miteinander in mehreren Schichten verknüpft, so dass die Ergebnisse einer Funktion den Eingangswert anderer Funktionen bilden. Die Architektur des Netzwerks, also die Verknüpfung der einzelnen Knoten, und die verwendeten Funktionen werden als Designentscheidungen durch die Entwickler/innen vorgegeben. In die Berechnungen fließen Gewichtungsfaktoren (Parameter) ein, die im Lauf von *Trainingsdurchläufen* zur Optimierung der Ergebnisse des Modells angepasst werden. Nach dem Training folgt die *Nutzungsphase*. Im Fall von Modellen zur Sprachverarbeitung wird ein Text, der von den Nutzenden eingegeben wird, zunächst numerisch abgebildet und als Eingangswert der ersten Schicht genutzt. Nach mehr oder weniger komplizierten Berechnungen über die verschiedenen Schichten hinweg steht ein einzelnes Wort als Ergebnis fest, nach mehreren solchen Durchläufen (engl. inference) entsteht ein Antwortsatz oder -text (Wolfram 2023b).[7]

Nachdem bereits 2012 mit Deep-Learning-Architekturen ein großer Fortschritt bei der Bilderkennung gelang (Krizhevsky et al. 2017), stellte die 2017 von einem Forschungsteam von Google vorgestellte *Transformerarchitektur* (Vaswani et al. 2017) einen Meilenstein bei der Sprachverarbeitung dar. Transformer weichen von früheren Formen neuronaler Netzwerke insofern ab, als sie Daten nicht sequenziell abarbeiten. Stattdessen sind sie in der Lage, eine

7 Der Physiker, Informatiker und Mathematiker Stephen Wolfram, der unter anderem die Mathematik-Software Mathematica entwickelte und vertreibt, hat die Funktionsweise sprachverarbeitender KI-Modelle sehr detailliert und verständlich beschrieben (Wolfram 2023b).

Vielzahl an Daten, beispielsweise einen ganzen Satz oder gar Absatz, gleichzeitig zu verarbeiten. Im Unterschied zu früheren Modellarchitekturen können mit Transformern Verbindungen auch zwischen weit entfernt stehenden Wörtern beachtet und die Erkennung sprachlicher Muster verbessert werden (sogenannter Attentionmechanismus). Als weiterer großer Vorteil kann die Hardware von Rechenzentren mit parallel arbeitenden (Grafik-)Prozessoren effektiver eingesetzt und so ressourcenschonender und mit höherer Leistung gearbeitet werden (Toews 2022b). Transformermodelle können sehr viele Parameter beinhalten und mit sehr großen Datenmengen unter Einsatz enormer Rechenkapazitäten (Abb. 2.1) trainiert werden (Bommasani et al. 2021); sie dominieren die Liste der größten KI-Modelle (Grävemeyer 2022, S. 61).

Abb. 2.1 Entwicklung von Computermodellen zur Sprachverarbeitung nach Modellgröße (Anzahl der Parameter) im Zeitverlauf

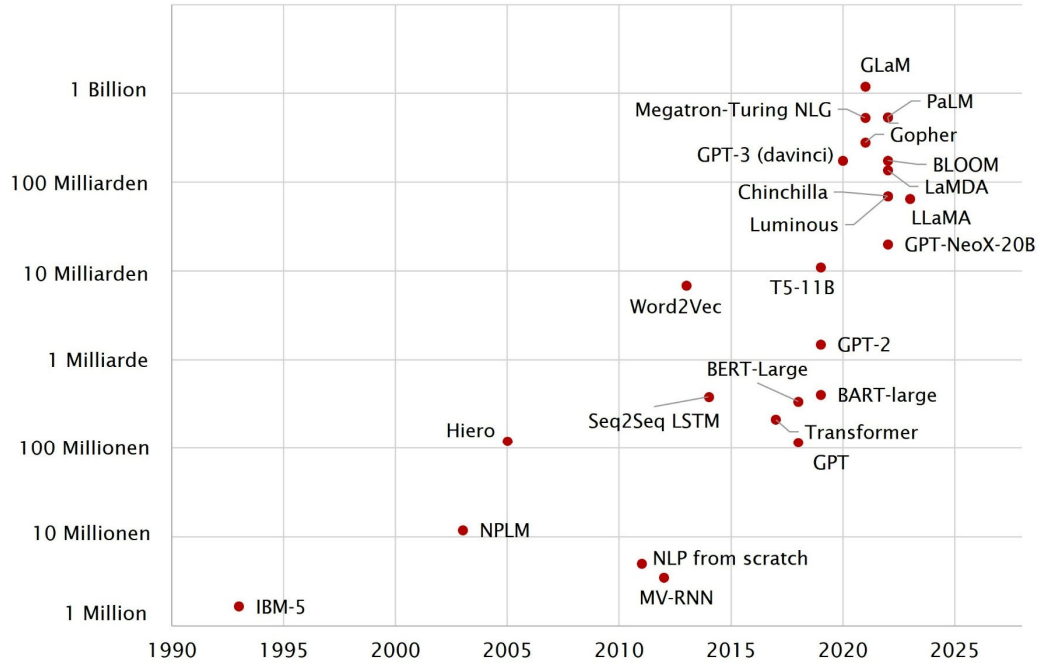

Abbildung ausgewählter Computermodelle zur Sprachverarbeitung nach Anzahl der Parameter (y-Achse, in logarithmischer Skalierung) und Zeitpunkt ihrer Veröffentlichung (x-Achse). GPT-4 ist nicht abgebildet, da bisher keine Angaben zur Modellgröße veröffentlicht wurden.

Quelle: https://ourworldindata.org/grapher/artificial-intelligence-parameter-count (20.4.2023, CC BY 4.0, vereinfachte und aktualisierte Darstellung)

Transformermodelle können so große Mengen menschlich erzeugter Texte als Trainingsdaten verarbeiten, dass sie ein Modell davon ausbilden, wie Menschen

mit Sprache Texte erzeugen (Ananthaswamy 2023).[8] Chatbots wie ChatGPT können auf dieser Basis, ohne den bei anderen Chatbots bzw. Dialogsystemen üblichen Rückgriff auf zusätzliche Regeln oder Wissensquellen, Texte erzeugen, die für Menschen sehr überzeugend als folgerichtige Antworten auf beliebige Texteingaben erscheinen. Dabei spielt es für die Qualität der Ausgabe kaum eine Rolle, ob die Eingaben dem Trainingsmaterial entsprechen oder gänzlich neuer Art sind. Diese Eigenschaft, auf nicht vorgegebene Aufgaben bzw. Eingaben zu reagieren, wird als »zero shot learning« bezeichnet (Radford et al. 2019).

Die Fähigkeit großer sprachverarbeitender Transformermodelle, *textförmige Aufgabenstellungen zu interpretieren*, wird unterschiedlich beurteilt (Hutson 2021, S. 23f.). Einerseits wird betont, dass die KI-Systeme nur scheinbar ein Verständnis der Fragen oder Aufgaben entwickeln, tatsächlich aber »oberflächlich« (Chomsky et al. 2023) Sprache »simulieren« (Berins 2023) bzw. als »stochastische Papageien« agieren (Bender et al. 2021, S. 617), die ohne Bezug auf den Sinngehalt Worte nach den statistischen Mustern der im Training erfassten Texte zusammensetzen. Andererseits wird das Auftreten einer solchen, als menschenähnlich erscheinenden Interpretationsfähigkeit als Resultat der besonders großen Zahl an Parametern und Trainingsdaten des Modells angesehen und es wird angenommen, dass sich aus noch größeren Modellen noch weiter gehende Fähigkeiten ergeben könnten (Bommasani et al. 2022; Ganguli et al. 2022; Tamkin et al. 2021; Wei et al. 2022).

OpenAI, das Unternehmen, das ChatGPT entwickelt hat, stellte erstmals 2018 seinen Ansatz eines Transformermodells vor (Radford et al. 2018). Das Modell GPT-3 (Brown et al. 2020) wurde 2020 zunächst nur ausgewählten Testpersonen zugänglich gemacht, sorgte allerdings bereits für Aufsehen aufgrund der Qualität der von ihm erzeugten Texte und seiner Fähigkeit, ohne spezifisches Training eine große Vielfalt an Aufgaben erledigen zu können (Hutson 2021; Kompetenzplattform KI.NRW 2021). Es verfügt über 175 Mrd. Parameter und 96 Schichten und wurde mit 300 Mrd. Token, also textlichen Bausteinen, trainiert (dabei kann es sich um Wörter, Wortbestandteile oder Wortkombinationen handeln) (Brown et al. 2020). ChatGPT beruht auf einer Familie von weiterentwickelten Modellen mit der Bezeichnung GPT 3.5.[9] Die seit Mitte Februar 2023 in Deutschland verfügbare Bezahlversion ChatGPT plus nutzt mittlerweile das am 14. März 2023 vorgestellte Nachfolgemodell GPT-4, das auch dem in die

8 Im Englischen wird dafür der Begriff »language model« verwendet, der allerdings laut der Linguistin Emily Bender (mündliche Kommunikation, Stochastic Parrot Webtalk am 17.3.2023) irreführend ist, weil nicht Sprache insgesamt repräsentiert wird, sondern nur die im Training verwendeten – allerdings sehr großen – Textmengen mit den in ihnen vorkommenden statistischen Zusammenhängen zwischen Wörtern bzw. Wortbestandteilen.

9 https://platform.openai.com/docs/model-index-for-researchers (19.4.2023).

Suchmaschine Bing integrierten Chatbot zugrunde liegt. OpenAI hat keine technischen Details zu GPT-4 veröffentlicht (OpenAI 2023f).

Eine Abwandlung von GPT-3, DALL-E, wird als Programm zur Erzeugung von Bildern anhand von Vorgaben in Textform (etwa zu Bildinhalten oder Stilen) angeboten. Dieses GPT-3-Modell verfügt über 12 Mrd. Parameter und wurde anhand von Daten trainiert, die Bilder und Textbeschreibungen verbinden (Ramesh et al. 2021). Seit September 2022 ist eine zweite, verbesserte Version verfügbar.[10] Eine weitere Abwandlung von GPT-3, genannt Codex, wurde mit Programmcode von der Plattform GitHub trainiert und kann für Programmieraufgaben genutzt werden (Chen et al. 2021).

Weitere bekannte große KI-Modelle zur Sprachverarbeitung wurden unter anderem von den Unternehmen Google, DeepMind, Meta, Microsoft und Nvidia entwickelt (Tab. 2.1). Eine Besonderheit stellt dabei Googles Modell GLaM dar, weil es nicht als einzelnes dichtes Netzwerk organisiert ist, sondern in Form von mehreren miteinander verbundenen kleineren Netzwerken (Ananthaswamy 2023). Es soll bei besseren Ergebnissen nur ein Drittel der Energie benötigen, die bei GPT-3 für das Training aufgewendet wurde, allerdings aufgrund der großen Zahl von Parametern höhere Anforderungen an die Rechenleistung während der Nutzungsphase stellen (Du et al. 2022). Das Modell BLOOM wurde als Open-Source-Modell kollaborativ von 1.000 ehrenamtlich Mitwirkenden speziell für die Erforschung von sprachverarbeitenden KI-Modellen entwickelt. Das Trainingsmaterial wurde unter Berücksichtigung möglichst unterschiedlicher Sprachen und kultureller Einflüsse ausgewählt (Ananthaswamy 2023; Big Science Workshop 2023; Gibney 2022). In Deutschland hat das Unternehmen Aleph Alpha 2022 das Modell Luminous-supreme entwickelt, das in einzelnen Benchmarkingtests, etwa zu logischen Ableitungen, besser abschnitt als GPT-3. Noch 2023 soll eine Version mit 300 Mrd. Parametern fertiggestellt werden (Hahn 2023b).

10 https://openai.com/product/dall-e-2 (19.4.2023).

Tab. 2.1 Ausgewählte große KI-Modelle zur Sprachverarbeitung

Name	Entwickler/in	Einführungsdatum	Anzahl Parameter	Anzahl Schichten	Trainingsmaterial (Token)
GPT-3 (Generative Pre-Trained Transformer-3)	OpenAI	2020	175 Mrd.	96	300 Mrd.
GPT-4	OpenAI	2023	?	?	?
LaMDA (Language Models for Dialog Applications)	Google	2022	137 Mrd.	64	2,8 Bill.
PaLM (Pathways Language Model)	Google	2022	540 Mrd.	118	780 Mrd.
GLaM (Generalist Language Model)	Google	2021	1,2 Bill.	32	1,6 Bill.
Gopher	DeepMind	2021	280 Mrd.	80	300 Mrd.
Chinchilla	DeepMind	2022	70 Mrd.	80	1,4 Bill.
LLaMA (Large Language Model Meta AI)	Meta	2023	65 Mrd.	80	1,4 Bill.
Megatron-Turing NLG	Microsoft, Nvidia	2021	530 Mrd.	105	339 Mrd.
BLOOM (BigScience Large Open-science Open-access Multilingual Language Model)	BigScience	2022	176 Mrd.	70	350 Mrd.
Luminous-supreme	Aleph Alpha	2022	70 Mrd.	?	588 Mrd.

Eigene Zusammenstellung. Quellen: Ananthaswamy 2023; Big Science Workshop 2023; Brown et al. 2020; Du et al. 2022; Gibney 2022; Hahn 2023b; Smith et al. 2022; Thoppilan 2022

2.2 Durchbruch dank Training anhand von großen Datenmengen

Daten als Grundlage der Entwicklung und Anwendung von KI-Systemen wird große Bedeutung zugemessen, beispielsweise in der KI- wie auch der Datenstrategie der Bundesregierung (Bundesregierung 2018 und 2021). Dabei galt noch im Jahr 2009 die Idee als vergleichsweise exotisch, für das Training von Bilderkennungssystemen eine große Bilddatenbank zu erstellen, in der zu jedem Begriff eine Vielzahl von Bildern enthalten ist (Deng et al. 2009). Erst

allmählich etablierte sich eine neue, datenorientierte Herangehensweise (Gershgorn 2017).

Auch bei sprachverarbeitenden KI-Modellen ist anerkannt, dass ihre Leistung von *großen Trainingsdatensets* abhängt (neben der Modellgröße [Brown et al. 2020, S. 4] und der Computerleistung beim Training) (Kaplan et al. 2021). Das genaue Zusammenspiel dieser Faktoren ist noch nicht im Detail verstanden (Ananthaswamy 2023), es existieren Gegenbeispiele, bei denen die Leistung mit steigender Größe sinkt (Perez et al. 2022). Die Transformerarchitektur ermöglichte erst die Verarbeitung großer Mengen von Textdaten. In der gleichen Zeit wuchs infolge der zunehmenden Digitalisierung menschlicher Kommunikation (einschließlich der Digitalisierung bereits vorhandener analoger Werke) die Zahl verfügbarer Texte, die für das Training von KI-Modellen genutzt werden konnten. Bei Transformern kann der erste Lernschritt, das Vortrainieren, eigenständig durch das Modell absolviert werden (unüberwacht), ohne dass dafür menschliches Feedback oder annotierte, also mit qualitativ hochwertigen Zusatzinformationen versehene, Daten benötigt werden (Ananthaswamy 2023). Solche Daten sind meist nur begrenzt verfügbar oder nur mit großem Aufwand zu erstellen, sodass Formen des maschinellen Lernens, die auf diese Daten angewiesen sind, nur eingeschränkt genutzt werden können.

Während für GPT-4 keine genaueren Angaben zu den Trainingsdaten und -prozeduren vorliegen, lässt sich das Training von GPT-3 und dem darauf aufbauenden ChatGPT anhand der Veröffentlichungen von OpenAI nachvollziehen. Das Training von GPT-3 erfolgte in zwei Schritten. Der erste Schritt hatte zum Ziel, sprachliche Äußerungen, die als Eingabe bzw. Aufforderung (engl. prompt) vorliegen, durch weiteren, wie von einem Menschen verfassten Text fortzuführen – als Vergleich wird häufig die Funktion des automatischen Vervollständigens von Eingaben herangezogen, die viele Smartphones anbieten (engl. autocomplete). Um möglichst generisch ganz unterschiedliche Aufgaben der Sprachverarbeitung erledigen und flexibel auf Eingaben reagieren zu können, wurde das Modell mithilfe der Methode des unüberwachten Lernens vortrainiert, bei dem das Modell anhand einer sehr großen, nicht kategorisierten Datenmenge Strukturen (in Form der Parameter des neuronalen Netzwerks) ausbildet (Brown et al. 2020, S. 5; Radford et al. 2018, S. 2f.).

Das *Trainingsmaterial von ChatGPT* bestand dabei in einem spezifisch aufbereiteten Set von Webseiten (CommonCrawl) sowie solchen Webseiten, auf die Nutzende in Onlineforen (unter anderem Reddit) verwiesen haben. Außerdem wurden digitalisierte Bücher und Wikipediaartikel genutzt (Brown et al. 2020, S. 9).[11] Dabei wurden alle Quellen als grundsätzlich gleichwertig behandelt, es fand keine Einstufung als verlässlich oder korrekt statt. Das Vortraining wurde im September 2021 abgeschlossen, sodass nur Daten bis zu diesem

11 Eine Beschreibung und Kritik der Datensätze findet sich bei Rettberg (2022).

Zeitpunkt Eingang in das Trainingsmaterial fanden (Hahn 2023a; OpenAI 2023f, S. 10).

Im Fall von ChatGPT wurde das Modell im zweiten Schritt, dem *Feinjustieren* (engl. fine-tuning), zusätzlich durch *Input und Feedback von Menschen* trainiert (»reinforcement learning from human feedback« als Form des überwachten Lernens), um die Dialogfähigkeit zu verbessern. Dazu wurden dem System menschliche Antworten auf Eingaben vorgegeben und es wurden Antwortoptionen, die vom System zu einer Eingabe vorgeschlagen wurden, vergleichend von menschlichen Codierer/innen bewertet. Das daraus resultierende Modell wurde wiederum für das maschinelle Training des eigentlichen Modells genutzt (OpenAI 2022a; Wolfram 2023b). Die Bedeutung dieses Feinjustierens als »Erfolgsrezept« (Heaven 2023b) für die Entwicklung von ChatGPT aus dem ursprünglichen Modell ist offenbar nicht zu unterschätzen. So äußert Liam Fedus, einer der Entwickler von ChatGPT: »Wie sich herausstellte, hatten die Konversationsdaten einen großen positiven Einfluss auf ChatGPT.« (Heaven 2023b, im Original: »As it turned out, the conversational data had a big positive impact on ChatGPT.«) Für die Feinjustierung der GPT-Modelle griff (und greift) OpenAI auf menschliche Arbeit aus Ländern aus aller Welt zurück (Albergotti/Matsakis 2023; Perrigo 2023).

OpenAI behält sich außerdem vor, *Daten der Nutzenden* für die Verbesserung des Systems zu verwenden, sofern kein Widerspruch erfolgt oder das System über die Programmierschnittstelle (engl. application programming interface, API) aufgerufen wird (Markovski o.D.). Es wird daher davor gewarnt, dem System sensible Daten zur Verfügung zu stellen (Heikkilä 2023a). Inwiefern Daten aus der aktuellen Nutzung für das Training von ChatGPT bzw. GPT-4 verwendet werden, ist nicht bekannt; GPT-4 wurde im Zuge der Entwicklung teilweise anhand von Prompts der Nutzung von ChatGPT trainiert (OpenAI 2023f, S. 24, FN 29). Offenbar bleiben die Modelle nach dem Training eine gewisse Zeit unverändert: »Anwendungen wie ChatGPT sind stabile Systeme (…) sie entwickeln und verändern sich nicht online und in Echtzeit, auch wenn sie offline ständig weiterentwickelt werden« (Brown University 2023, im Original: »applications like ChatGPT are steady-state systems (…) they aren't evolving and changing online, in real-time, even though they may be constantly re-fined offline«). Allerdings wurde bereits vor Einführung der neuen Version GPT-4 das ChatGPT unterliegende Modell mehrfach aktualisiert.[12]

12 https://help.openai.com/en/articles/6825453-chatgpt-release-notes (19.4.2023).

2.3 Gestaltungsfragen: Sicherheitsvorkehrungen und Nutzungsschnittstelle

Das Nutzungserlebnis eines Computersystems wird gemäß der Norm DIN EN ISO 9241-210 durch die Funktionalität, die Leistungsfähigkeit des Systems und die Gestaltung der Benutzungsschnittstelle bestimmt. In Bezug auf die Leistungsfähigkeit von ChatGPT wurde von den Entwickler/innen bewusst die Vielfalt der Antwortmöglichkeiten durch *Sicherheitsvorkehrungen* eingeschränkt, um z. B. unangemessene oder rechtlich problematische Äußerungen zu vermeiden. Ein bekanntes Beispiel für Fehlverhalten von Chatbots ist Microsofts Tay, der 2016 (also noch vor dem Durchbruch bei sprachverarbeitenden KI-Modellen) auf Twitter eingesetzt wurde und innerhalb kürzester Zeit aufgrund seiner rassistischen und politisch extremen Äußerungen vom Netz genommen werden musste (Wolf et al. 2017). Auch bei den jüngsten, sehr großen KI-Modellen wurden solche unerwünschten Verhaltensweisen häufig beobachtet (Ganguli et al. 2022). So wurde das Galactica-System von Meta, das eigentlich speziell auf die Unterstützung wissenschaftlicher Arbeit ausgerichtet war, nach nur drei Tagen wieder außer Betrieb genommen, weil es frei erfundene Sachverhalte als wissenschaftliche Fakten darstellte (Heaven 2022) und wissenschaftlich scheinende Begründungen auch für diskriminierende bzw. menschenverachtende Sichtweisen lieferte (Snoswell/Burgess 2022).

Eine der vorbeugenden Sicherheitsvorkehrungen bei ChatGPT besteht laut OpenAI (OpenAI 2023b, d) im Training mit menschlichem Feedback, mit dem beim Feinjustieren des Modells unerwünschte Ausgaben gekennzeichnet und in den maschinellen Lernprozess zurückgespeist wurden (Ouyang et al. 2022, OpenAI 2023e, S. 21f.). Einem Entwickler zufolge bestehen die dem Training zugrundegelegten Regeln, nach denen ChatGPT Dialoge führen soll, unter anderem darin,

- nachzufragen, falls die Eingabe eines Nutzers/einer Nutzerin nicht klar ist;
- deutlich zu machen, dass ein KI-System kommuniziert;
- keine Identität anzunehmen, die das System nicht hat;
- nicht zu behaupten, Fähigkeiten zu besitzen, die das System nicht hat;
- abzulehnen, wenn ein/e Nutzende/r es auffordert, Aufgaben zu erledigen, die das System nicht tun soll.[13]

Diese Regeln ließen sich in ersten Erprobungen des Systems nach seiner Veröffentlichung allerdings leicht durch die Nutzenden umgehen (engl. jailbreaking) und wurden als »ziemlich plump« kritisiert (Golumbia 2022). Nach der Bereitstellung von ChatGPT mussten die Regeln angepasst werden, um bekannt gewordenen Fällen von rassistischem Systemverhalten entgegenzuwirken

13 Heaven (2023b); OpenAI (2022b).

(Marcus/Davis 2023b). Im Fall des Chatbots, der bei Microsofts Suchmaschine Bing eingesetzt wird und auf GPT-4 basiert, wurde nach kurzer Zeit die Zahl der Gesprächsrunden begrenzt. Mehrere Nutzende hatten bei länglichen Konversationen erlebt, dass der Chatbot sehr persönliche Antworten verfasste und beispielsweise ausfällig wurde oder Liebeserklärungen machte (Spiegel Online 2023).

Eine weitere Sicherheitsvorkehrung besteht darin, dass ein gesondertes KI-Modell für das Erkennen von unerwünschten Äußerungen (z. B. Hassrede, gewaltverherrlichende Äußerungen) des Systems erstellt wurde, um die Ausgaben des Systems zu überprüfen (Markov et al. 2022). Für das Training dieses Modells griff OpenAI offenbar über ein Dienstleistungsunternehmen auf Arbeitende unter anderem aus Kenia zurück, die nur sehr wenig Lohn für ihre psychisch z. T. sehr belastende Tätigkeit erhielten (Perrigo 2023). Das Trainingsmaterial für das Vortrainieren wurde grob gefiltert, um unerwünschte (beispielsweise sexuell anstößige) Inhalte auszuschließen (Brown et al. 2020, S. 8; OpenAI 2023e, S. 3), und das Team von OpenAI entwickelte Szenarien für gezielte Manipulationen des Systems, um Gegenmaßnahmen vorzusehen (Heaven 2023b; OpenAI 2023e, S. 5). Nicht zuletzt sollen auch die Nutzungsbedingungen von OpenAI dazu dienen, eine missbräuchliche Verwendung der KI-Systeme zu verhindern;[14] die Nutzung des Systems wird zudem laufend überwacht (OpenAI 2023e, S. 26). Mehrere Länder, darunter Ägypten, Afghanistan, Belarus, China, Iran, Kuba, Nordkorea, Russland und Saudi-Arabien,[15] sind von der Nutzung der API ausgeschlossen – allerdings ist nicht klar, ob dies aus Gründen der Sicherheit geschieht (Wang 2023).

Die *Benutzungsschnittstelle* (engl. user interface) von ChatGPT bestand bei der Einführung des Systems in einer betont schlichten einzeiligen Eingabemaske auf einer Website. Die Besonderheit liegt in der einfachen, dialogförmigen und natürlichsprachigen Interaktion mit dem System. Dabei sind auch längere Eingaben und unterschiedliche Sprachen möglich.[16] Die Ausgabe erfolgt ebenfalls in Textform. Über die Website können Nutzende auf ihre früheren Konversationen zurückgreifen werden, die im System gespeichert werden.[17]

---

14 Die Bedingungen sehen etwa vor, dass die Nutzenden dafür verantwortlich sind, dass die Inhalte, die sie an das System übermitteln, nicht gegen Gesetze verstoßen und dass sie mit ihrer Nutzung des Systems keine Rechte Dritter verletzen. Es ist außerdem nicht erlaubt, mithilfe der Ausgaben des Systems ein mit OpenAI konkurrierendes Modell zu entwickeln (https://openai.com/policies/terms-of-use, 24.3.2023). Eine Reihe von Nutzungsmöglichkeiten wird ausgeschlossen, darunter die Erzeugung von Schadprogrammen sowie bestimmte Verwendungen im Kontext politischer Kampagnen und von wissenschaftlichem Fehlverhalten (https://openai.com/policies/usage-policies, 24.3.2023).

15 https://platform.openai.com/docs/supported-countries (24.3.2023).

16 Bei ChatGPT ist der Kontext, also die Anzahl der Token, die das System verarbeiten kann, auf 4.096 begrenzt (was etwa 3.000 Wörtern entspricht), bei GPT-4 sind bis zu 32.768 Token möglich (https://platform.openai.com/docs/models/, 24.3.2023).

17 Aufgrund eines technischen Defekts kam es im März 2023 dazu, dass Nutzende kurzzeitig auch Teile der Konversationen anderer Nutzender sehen konnten (Open AI 2023d).

Nutzende bewerten den Zugang über die Website und die Interaktion mit dem System als angenehm (Klinge 2022) und »extrem leicht zugänglich und bedienbar« (Emmerich 2023). Ein Entwickler von OpenAI sieht in der Dialogschnittstelle den »erstaunlichen Fortschritt«, der den Unterschied zu den vorherigen, öffentlich weniger beachteten Modellen ausmacht (Heaven 2023b).

Eine *Nutzung der Modelle* von OpenAI ist für registrierte Benutzer/innen auch über den »Playground« der Website möglich, einen gesonderten Bereich, in dem sich nicht nur einzelne Modelle testen lassen, sondern auch Funktionsmodi und Parameter festlegen lassen (wie die Länge von Antworten, der Grad an Zufälligkeit bei der Auswahl der Ergebnisse durch das Modell oder die Wahrscheinlichkeit von Wiederholungen bzw. Redundanzen). Seit März 2023 kann ChatGPT von Entwickler/innen auch über eine API, also eine Schnittstelle für andere Computerprogramme, genutzt werden. Die API ist vor allem für Anbieter/innen von Dienstleistungen (engl. third-party services) interessant, die auf diesem Weg die Funktionen von ChatGPT bzw. GPT-4 in ihre eigenen Angebote einbetten können (Kap. 2.5). Über die API können auch Zusatzdienste von OpenAI genutzt werden. So ist es etwa möglich, mithilfe des Whisper-Modells gesprochene Sprache als Eingabe zu verarbeiten und so beispielsweise Transkriptionen von Tonaufnahmen zu erstellen. Auch ein Moderationsmodell ist verfügbar, mit dem sich Texte auf unerwünschte Inhalte hin überprüfen lassen (falls Drittanbieter die von ChatGPT ausgegebenen Texte z. B. auf ihrer Website nutzen wollen).

2.4 Welche Hardwareumgebung ist für große KI-Modelle zur Sprachverarbeitung erforderlich?

Für das Training von künstlichen neuronalen Netzen werden bevorzugt Grafikprozessoren (engl. graphical processing units, GPU) eingesetzt (Hooker 2020). Die eigentlich für Computerspiele entwickelten Prozessoren eignen sich dank ihrer Fähigkeit zur parallelen Berechnung besonders gut für diese Aufgabe und sorgten im Verbund mit weiterentwickelten Modellarchitekturen für Durchbrüche im Bereich der Bilderkennung und Sprachverarbeitung.

Die heutigen KI-Modelle, beispielsweise der GPT-Familie, erfordern insbesondere in der Trainingsphase eine *sehr leistungsfähige Hardwareumgebung*, um die Vielzahl an Parametern berechnen zu können (AKI 2023, S. 31). Dafür sind in der Regel große, spezialisierte Rechenzentren an einem Ort oder entsprechende Dienstleistungen von Cloud-Computing-Anbietern nötig. OpenAI gibt an, dass für das Training von ChatGPT die AI Supercomputing Infrastructure von Microsofts Cloud-Computing-Plattform Azure genutzt wurde (OpenAI 2022a); die Berechnung der Modellparameter erfolgte auf hochspezialisierten Grafikprozessoren. Bereits im Jahr 2019 ging Microsoft mit einer Investition von rund 1 Mrd. US-Dollar eine enge Partnerschaft mit OpenAI ein (Microsoft

2019), 2020 wurde von Microsoft ein Hochleistungsrechner speziell für die KI-Modelle von OpenAI in Betrieb genommen (Langston 2020).

Neben den Kosten für die Bereitstellung und die Unterhaltung der Hardwareinfrastruktur bedingen die hardwaretechnischen Anforderungen der KI-Modelle auch den *Energiebedarf beim Training und in der Nutzungsphase* der Modelle. Schätzungen zufolge waren für das Training von GPT-3 1.287 MWh Strom nötig (Patterson et al. 2021, S. 6; Luccioni et al. 2022, S. 7) – etwa der jährliche Strombedarf von 400 deutschen Durchschnittshaushalten. Die Treibhausgasemissionen werden auf 552 t CO2-Äquivalente beziffert, das entspricht etwa 6 Flügen eines Passagierjets von New York nach San Francisco (Patterson et al. 2021, S. 6; Luccioni et al. 2022, S. 7).[18] Im Vergleich dazu waren andere KI-Modelle weniger energiehungrig: Das Modell Gopher mit 280 Mrd. Parametern verbrauchte geschätzt 1.066 MWh Strom bzw. sorgte für die Emission von 380 t CO2-Äquivalenten, bei BLOOM (mit 176 Mrd. Parametern ähnlich groß wie GPT-3) waren es 433 MWh und 30 t CO2-Äquivalente (Luccioni et al. 2022, S. 7). Das Modell GLaM soll laut Google dank verbesserter Hard- und Software und seiner besonderen Architektur im Vergleich zu GPT-3 bei der gleichen benötigten Rechenleistung zum Training und besseren Ergebnissen nur etwa ein Drittel (456 MWh) der Energie des Modells von OpenAI benötigt haben (Ananthaswamy 2023, S. 205), die Entwickler/innen geben insgesamt gut 40 t CO2-Äquivalente an Treibhausgasemissionen an (Du et al. 2022).

Wie hoch die Kosten bzw. der Aufwand für den Betrieb des Systems sind, ist nicht näher bekannt. Aufgrund der sehr hohen Zahl von Nutzenden geht Ananthaswamy (2023, S. 204) davon aus, dass der Betrieb monatlich Millionen US-Dollar kostet.

2.5 Unternehmerische Aspekte

Das Unternehmen *OpenAI*, das ChatGPT entwickelt hat, wurde Ende 2015 gegründet und dabei von einer Reihe von Investoren und Gründern aus dem Umfeld des kalifornischen Gründerzentrums Y Combinator unterstützt, darunter Elon Musk und Peter Thiel. Der Name OpenAI, der freien Zugang zu den Entwicklungen und eine Gemeinwohlorientierung suggeriert, verweist auf die

18 Die Werte stellen Abschätzungen dar, da es unterschiedliche Berechnungsweisen der Treibhausgasemissionen von KI-Systemen beim Training gibt und nicht alle relevanten Einflussgrößen bekannt sind. So gehen Anthony et al. (2020, S. 10) nur von einem Stromverbrauch von 188,701 MWh für das Training von GPT-3 aus und errechnen einen entsprechend niedrigeren Treibhausgasausstoß von 85 t CO2-Äquivalenten. Andere schätzen den Strombedarf mit 1.404 MWh deutlich höher ein (Hans Uszkoreit in einem Vortrag bei der LEAM-Konferenz in Berlin, 24.1.2023, https://twitter.com/ovoss/status/1617824632543010817, 19.4.2023). Cowls et al. (2023, S. 292) wiederum errechnen die Treibhausgasemissionen allein anhand der benötigten Rechenoperationen der beim Training verwendeten Hardware und nennen 224 t CO2-Äquivalente.

Startphase des Unternehmens. Die Leiter des Unternehmens nannten als Ziel, führend in der Forschung zu KI-Systemen zu werden – damals noch ohne Gewinnabsicht, sondern um »Werte für alle zu schaffen und nicht für die Aktionäre« (Brockman et al. 2015). Seit 2019 arbeitet OpenAI jedoch gewinnorientiert und ist als Limited Partnership (Beschränkte Partnerschaft, vergleichbar einer deutschen Kommanditgesellschaft) organisiert, um Investitionen zu akquirieren.

Besonders stark investierte *Microsoft* in OpenAI. 2016 begannen die beiden Unternehmen eine *technische Partnerschaft*, in deren Rahmen OpenAI Microsofts Cloud-Computing-Plattform Azure für die Entwicklung seiner Modelle nutzen konnte. 2019 investierte Microsoft rund 1 Mrd. US-Dollar und sicherte sich damit die Möglichkeit, KI-Technologie von OpenAI im Verbund mit seinen Cloud-Computing-Diensten zu vermarkten (Microsoft 2019) und beispielsweise Zugriff auf GPT-3 zu erhalten (OpenAI 2020). Ende Januar 2023 folgte eine weitere Investition von 10 Mrd. US-Dollar (Köver 2023), verbunden mit OpenAIs exklusiver Festlegung auf Azure als Cloud-Computing-Plattform sowie der Einbindung von KI-basierten Anwendungen in die Azure-Plattform, die Suchmaschine Bing und weitere Microsoft-Anwendungen (Microsoft 2023). Zum Teil sind diese Pläne bereits realisiert worden. Anfang Februar 2023 stellte Microsoft eine Erweiterung von Bing um einen Chatbot vor, der auf GPT-4 beruht (Mehdi 2023; Hahn 2023d). Microsoft sicherte sich mit der Investition außerdem umfangreiche Anteile an eventuellen Gewinnen von OpenAI bis zu einer Kappungsgrenze vom Hundertfachen der Gesamtinvestition (Kahn 2023).

Große KI-Modelle wie diejenigen der GPT-Familie werden nicht nur zur Erzeugung von Text, sondern auch von Bildern, Audio- und Videoinhalten sowie von Programmcode entwickelt. Die 24 bekanntesten dieser *generativen KI-Modelle* wurden von nur fünf Unternehmen entwickelt, neben OpenAI handelt es sich um Google bzw. DeepMind, Meta, Runway sowie Nvidia (Gozalo-Brizuela/Merchan 2023, S. 4; Toews 2022b). Nur zwei Modelle wurden von Universitäten ohne Beteiligung der Industrie entwickelt. Ahmed et al. (2023) machen einen Trend zu vornehmlich von Unternehmen kontrollierter KI-Forschung aus, der dadurch zustande kommt, dass diese gegenüber der akademischen Forschung einen besseren Zugang zu Daten, Rechenleistung sowie Expert/innen hat.

Die sehr hohen Investitionskosten in die Entwicklung, vor allem in das Training dieser KI-Modelle haben eine Konzentration auf nur wenige Unternehmen zur Folge (Ananthaswamy 2023, S. 204), die ohnehin bereits über eine große Marktmacht verfügen. Als ethisch bedenklich wird dabei angesehen, dass mehrere Quasimonopole zusammentreffen, etwa hinsichtlich von Textverarbeitungssoftware und Informationsverarbeitung bei Microsoft (van Dis et al. 2023). Auf diese Weise sichern sich die Unternehmen auch den Zugriff auf Daten, die aus der Nutzung von KI-Systemen wie ChatGPT resultieren, und

können so ihre Position weiter ausbauen. Nicht zuletzt für die wissenschaftliche Forschung, die auf Daten, aber auch die Kapazitäten zu deren Verarbeitung angewiesen ist, kann ein solches Oligopol ein Risiko darstellen, wie im Fall des wissenschaftlichen Publizierens und der Abhängigkeit von wenigen Großverlagen deutlich wurde (Fecher/Schulz 2023). Initiativen wie LEAM (Large European AI Models, deutsch: Große europäische KI-Modelle)[19] möchten dem Risiko einer (nationalen bzw. europäischen) Abhängigkeit von den vornehmlich in den USA entwickelten KI-Modellen zur Sprachverarbeitung durch den Auf- bzw. Ausbau entsprechender Entwicklungskapazitäten in Europa begegnen (AKI 2023, S. 12).

Die Unternehmen, die KI-Modelle entwickeln, bieten nicht unbedingt selbst *KI-basierte Dienste* an, manche Modelle werden auch nur intern genutzt, etwa für Moderationsaufgaben in sozialen Medien (Zhdanov 2023). Sofern die Modelle über eine API zugänglich gemacht werden, gibt es eine große Zahl an Drittanbietern, die auf Basis der KI-Modelle, ggf. gepaart mit spezieller Feinjustierung anhand eigener Daten, eigene Dienstleistungen anbieten (einen Überblick bietet die Website www.futurepedia.io). Ein Beispiel ist die Suchmaschine You.com, die mithilfe von GPT-Modellen in Verbindung mit den Suchmaschinen von Google und Microsoft YouChat, einen eigenen Chatbot, und YouWrite, ein Tool zur Unterstützung von Schreibprozessen, anbietet (Wiggers 2022).[20] Weitere Beispiele sind jasper.ai (www.jasper.ai), das Unternehmen auf Basis einer Reihe von KI-Modellen Unterstützung bei allen möglichen Schreib- und Designaufgaben, insbesondere für das Marketing, bietet, sowie das Berliner Unternehmen Mindverse mit einem ähnlichen Profil, das allerdings auf deutschsprachige Anwendungen ausgerichtet ist (Brandstätter 2023). Außerdem gehören auch Unternehmen wie Hugging Face, das die Entwicklung von KI-Anwendungen fördert, sowie Anbieter von Cloud-Computing-Diensten (neben Microsofts Azure auch z. B. Amazon Web Services) bzw. der für diese nötigen Hardwarekomponenten (Nvidia, aber auch Google mit den speziell für maschinelles Lernen entwickelten Tensor-Prozessoren [TPU]) zum unternehmerischen Umfeld auf dem Gebiet der KI-Modelle (Bornstein et al. 2023).

---

19 https://leam.ai/ (19.4.2023).

20 Während You.com ähnlich wie Microsoft KI-Modelle von OpenAI nutzt, hat Google eine auf einem eigenen KI-Modelle basierende Chatfunktion in seine Suchmaschine integriert. Seit dem 21. März 2023 können Nutzende aus den USA und Großbritannien per Testzugang Bard erproben, der auf dem LaMDA-Modell beruht, mithilfe von menschlichem Feedback feinjustiert und in der Anzahl der möglichen Gesprächsrunden begrenzt wurde (Greis 2023; Heaven 2023a). Im Gegensatz zu Bings Chatbot generiert Bard die Antworten allein aus dem KI-Modell heraus und ist nicht direkt mit der Suchmaschine verknüpft. Er soll aber dazu anregen, die Antworten durch eine Googlesuche zu überprüfen. Das chinesische Unternehmen Baidu hat unter dem Namen Ernie ebenfalls eine Chatbot-Ergänzung für seine Suchmaschine angekündigt, die analog zu Microsofts Chatbot Dialoge mit den Nutzenden führen soll (Weiß 2023a). Bei einer als Live-Demonstration angekündigten Vorstellung des Systems im März 2023 konnte allerdings nur eine Aufzeichnung der Interaktion mit dem System präsentiert werden (Che/Liu 2023).

Entsprechend der Vielzahl von Unternehmen existieren auch die unterschiedlichsten *Geschäftsmodelle*, sodass hier in erster Linie die Ausrichtung von OpenAI für ChatGPT betrachtet wird. ChatGPT wird nach seiner Einführung als offener Forschungsprototyp (engl. research preview) mittlerweile nach dem *Freemium-Modell* angeboten: Die Grundversion ist für registrierte Nutzende mit dem Zugang über die Website kostenlos, für Abonnenten wird zusätzlich mit ChatGPT Plus ein kostenpflichtiger Zugang angeboten, der auch bei hoher Auslastung gegenüber der Grundversion schnellere Reaktionszeiten und früheren Zugriff auf neue Funktionen und Verbesserungen bieten soll (OpenAI 2023c). Das System wird dabei kontinuierlich weiterentwickelt, neue Versionen werden meist ohne Vorankündigung als Ersatz für die bisherigen Versionen eingeführt. Feedback von Nutzenden wird gezielt gesammelt und für Weiterentwicklungen – offenbar auch des Geschäftsmodells (Kahn 2023) – genutzt.

Des Weiteren vermarktet OpenAI ChatGPT als Dienstleistung nach dem Modell von *»KI als Dienstleistung«* (engl. AI as a service). Dabei erfolgt der Zugriff auf das Modell durch andere Computersysteme über die API, wobei auch andere Modelle sowie Zusatzdienste genutzt werden können (Kap. 2.3). Die Abrechnung dieser Zugriffe erfolgt nutzungsbezogen und wird nach der Zahl der Token und dem verwendeten Modell berechnet.[21] Prominentestes Beispiel für die Integration von KI-Angeboten von OpenAI ist Microsoft: Integriert in Word sollen die GPT-Modelle beispielsweise Zusammenfassungen oder Textvorschläge erstellen können; in Excel soll es möglich sein, Fragen zu den Daten zu stellen und automatisierte Auswertungen als Antwort zu erhalten; und nach einer Videokonferenz über Microsoft Teams sollen sich mithilfe der GPT-Funktionen Protokolle anfertigen und Arbeitsaufträge versenden lassen, so die Vorstellung von Microsoft (Spataro 2023). Die Besonderheit der Integration ist, dass das KI-System auf die Daten der Nutzenden und die Funktionen der Microsoft-Anwendungen zugreifen kann, um diese Arbeitsaufträge auszuführen.

Ein weiteres Geschäftsmodell wird durch die Entwicklung der Ende März 2023 von OpenAI vorgestellten Plugins für ChatGPT (OpenAI 2023a) ermöglicht. Dabei handelt es sich um Zusatzdienste, mit denen ChatGPT gekoppelt wird und auf die Nutzende im Zuge ihrer Anfragen zugreifen können (Kap. 3.3.2). OpenAI folgt bei der Umsetzung dieser Plugins dem *Plattformgedanken*, wie er beispielsweise von App Stores bekannt ist. Andere Unternehmen können ihre Dienste auf ChatGPT zuschneiden, so dass eine enge Integration erfolgt. Auch dabei spielen APIs eine Rolle, allerdings greift in diesem Fall ChatGPT über APIs auf die Systeme Dritter zu. ChatGPT kann dadurch (wie etwa die Plattformen der App Stores bzw. von Amazon in Bezug auf die dort vertretenen Angebote) als Zugang zu Drittanbieter/innen genutzt werden, die Plugins

21 https://openai.com/pricing (24.3.2023).

erweitern aber auch den Funktionsumfang und die Leistungsfähigkeit des Systems selbst (McCormick 2023).

Auch wenn der Name des Unternehmens dies nahelegt, spielt das Geschäftsmodell einer Entwicklung von *KI-Modellen nach Open-Source-Prinzipien* bei OpenAI keine Rolle (mehr). Es liegt jedoch beispielsweise dem Modell BLOOM zugrunde, das von Big Science entwickelt wurde, einem öffentlich geförderten Zusammenschluss von mehr als 1.000 Forschenden aus ganz unterschiedlichen Disziplinen (Gibney 2022). Diese haben sich der Erforschung von großen KI-Modellen verschrieben, um sie transparenter, sicherer und ausgewogener zu machen. Das Vorgehen war unter anderem durch die sorgfältige Auswahl von Trainingsmaterialien hoher Qualität und unterschiedlicher Sprachen sowie durch gezielte Tests auf Voreingenommenheit in den Ausgaben des Systems gekennzeichnet. Das Modell steht insbesondere Forschenden frei zur Verfügung, kann aber auch über die Website des Unternehmens Hugging Face genutzt.[22] Mit einem ähnlichen Vorgehen hat auch Meta sein KI-Modell LLaMA für Forschende auf Antrag zugänglich gemacht (Touvron et al. 2023). Allerdings wurde das Modell nach kurzer Zeit von Unbekannten frei im Internet zugänglich gemacht, was, angesichts der Möglichkeit, das leistungsfähige Modell an beliebige Zwecke anzupassen, Sicherheitsbedenken auslöste (Kühl 2023, Kap. 4.3.2).

In Bezug auf die Motive vor allem kommerzieller Anbieter stellen einige Expert/innen infrage, *wie sich die Investitionen* in sehr große KI-Modelle im Bereich der Sprachverarbeitung *unternehmerisch auszahlen sollen*, zumal die Frage des rechtlichen Schutzes der Ausgaben der Modelle noch unklar ist (Kap. 6.2) (Simon 2021). Es wird vermutet, dass es auch den kommerziellen Anbietern weniger um die Vergütung der Leistungen der KI-Modelle als vielmehr um technologische Vormachtstellung (Meineck 2023), wissenschaftlichen Selbstzweck (Simon 2021) oder um Marketing gehen könnte, etwa für leistungsfähige Cloud-Computing-Plattformen (Monroe 2023) oder die Aufwertung bisher wenig relevanter Suchmaschinen (Dastin/Nellis 2023).

22 https://huggingface.co/bigscience/bloom (19.4.2023).

3 Möglichkeiten und Grenzen der Technologie

Große Computermodelle zur Sprachverarbeitung werden seit mehreren Jahren entwickelt, z. T. angewendet und hinsichtlich ihrer Leistungsfähigkeit erforscht. Daher liegt mittlerweile eine Reihe von Erkenntnissen bezüglich ihrer Möglichkeiten und Grenzen vor, die den Rahmen für eine Abschätzung der möglichen Anwendungen der Systeme wie auch der aus solchen Anwendungen entstehenden Gefahren (Kap. 4) bilden. Diese Erkenntnisse können jedoch nur als vorläufig und in erster Linie als Orientierung für die weitere Erforschung und Abschätzung angesehen werden. Die Entwicklung von KI-Modellen zur Sprachverarbeitung beruht zu großen Teilen auf Erfahrungswissen, das durch Versuch und Irrtum gewonnen wird (Wolfram 2023b), selbst Expert/innen verstehen zwar Aufbau und Funktionsweise der Systeme, scheitern aber letztlich daran, deren Verhalten wissenschaftlich erklären oder gar vorhersagen zu können (Krischke 2023).[23]

3.1 Möglichkeiten bzw. Stärken

Die besondere Leistungsfähigkeit von GPT-3, jeweils noch einmal gesteigert bei den Weiterentwicklungen ChatGPT und GPT-4, besteht im Umgang mit allen Arten von Texten in menschlicher Sprache (inklusive Programmiersprachen). ChatGPT wird daher auch als »leistungsstarkes Universalmodell« für ganz verschiedene sprachliche Aufgaben eingeschätzt (Qin et al. 2023, S. 11, im Original: »powerful generalist model«). Stärken der Systeme umfassen die Bearbeitung von vorgegebenen Texten (Kap. 3.1.1), die Erzeugung neuer Texte (Kap. 3.1.2) sowie die Interpretation von Aufgaben und das Interagieren im Rahmen von Dialogen (Kap. 3.1.3). Bemerkenswert ist daran insbesondere, in welch kurzer Zeit die Systeme die ihnen gestellten Anfragen beantworten bzw. Aufgaben erledigen.

All diese Aufgaben können in mehreren Sprachen bearbeitet werden. Für GPT-4 wurde die Leistungsfähigkeit außer in englischer in 26 weiteren Sprachen überprüft, in 24 Sprachen erreichte das System in Multiple-Choice-artigen Benchmarkingtests bessere Ergebnisse als das Vorgängermodell ChatGPT in seiner leistungsfähigsten englischen Sprachvariante (OpenAI 2023f, S. 8). Bei beiden Modellen fallen die Ergebnisse für seltenere, aber auch für manche weitverbreitete Sprachen deutlich schlechter aus als für Englisch (OpenAI 2023f,

23 Rahwan et al. (2019) schlagen unter anderem aus diesem Grund vor, Computersysteme mit einem interdisziplinären, verhaltenswissenschaftlichen Ansatz zu erforschen.

S. 8; Seghier 2023). Eine Besonderheit ist außerdem, dass die Systeme, auch dank eines speziellen Trainings insbesondere der späteren Modelle, auf unterschiedliche Programmiersprachen angewendet werden können und sich generell für die Verarbeitung aller Arten von Code eignen könnten (Toews 2022b). Auch eine Anwendung von KI-Modellen zur Sprachverarbeitung auf genetischen Code wird derzeit erprobt, etwa zur Vorhersage der Auswirkungen von Mutationen beim Menschen (Dalla-Torre et al. 2023) oder um Proteinstrukturen mit hoher Geschwindigkeit aus Gensequenzen ableiten zu können (Lin et al. 2023).

3.1.1 Bearbeitung von vorgegebenen Texten

Bereits bestehende und dem System mit der Eingabe vorgegebene Texte können *zusammengefasst* (Tamkin/Ganguli 2021) oder in eine andere Sprache *übersetzt* werden (Narayanan/Kapoor 2022). Auch eine *Veränderung von Stilen oder Sprechweisen* ist möglich, wie der Ausdruck eines Textes in Gedichtform oder in der Sprache von Grundschulkindern. Dabei zeigt sich die für Transformermodelle charakteristische Fähigkeit, über die reine Wortbearbeitung hinaus auch gewisse *abstrakte Zusammenhänge* zu erfassen, etwa den Kontext mehrdeutiger Wörter oder semantische Konzepte. Das System kann daher auch für die *Analyse von Texten* verwendet werden, etwa um einen Text auf die verwendeten Wortarten, eine vorherrschende Gefühlsstimmung, seine Verständlichkeit oder bestimmte Informationen hin zu untersuchen (Tamkin/Ganguli 2021). Auch *stilistische Bewertungen* können durch das System vorgenommen werden. Mit den Ergebnissen kann ein Text korrigiert und verbessert werden. Ein begrenzender Faktor ist dabei die *Länge von Text*, die das System als Eingabe verarbeiten kann (Kap 2.3). In der maximalen Version von GPT-4 ist diese Begrenzung allerdings mit mehr als 30.000 Token (entspricht im Englischen gut 22.000 Wörtern) bereits sehr großzügig bemessen, was einem längeren wissenschaftlichen Aufsatz entspricht (dieser Bericht umfasst bis zu dieser Stelle knapp 6.500 Wörter).

3.1.2 Erzeugung neuer Texte

Außer auf bereits bestehende Texte kann das System auch zur *Erzeugung neuer Texte* angewendet werden, auch dies in *unterschiedlichen Sprachen* sowie im Rahmen der Begrenzungen der Textlänge (Toews 2022a). Besonderes Aufsehen erregte dabei die Fähigkeit, Vorgaben bezüglich des Stils der Texte berücksichtigen zu können; beispielsweise brachten Floridi und Chiriatti (2020) GPT-3 dazu, ein Sonett in italienischer Sprache weiterzuschreiben.

Auch Witze lassen sich generieren (Abb. 3.1, Roose 2022) und auch erläutern (The Economist 2022). Die Fähigkeit, neue Texte zu erzeugen, hat sich

insbesondere für *Programmieraufgaben* als sehr nützlich erwiesen (The Economist 2022; Brown et al. 2020). Bereits GPT-3 konnte beispielsweise als Antwort auf eine allgemeinsprachlich formulierte Beschreibung von Elementen einer Website den Code ausgeben, mit dem sich diese Elemente erstellen lassen. Diese mittlerweile durch gezieltes Trainieren an Codebeispielen erweiterte Fähigkeit[24] wird bereits in verschiedenen Anwendungen, auch in Zusammenarbeit mit der (zu Microsoft gehörenden) Plattform GitHub, genutzt (Kap. 4.1.2).

Abb. 3.1 Nutzung von ChatGPT zur Erzeugung von Witzen

Screenshot einer Interaktion mit ChatGPT (https://chat.openai.com/, 19.4.2023).

Wie die Erprobung des Systems durch Nutzende mittlerweile gezeigt hat, kommt es insbesondere für die Erzeugung von Texten stark auf die vorangehende Eingabe, den Prompt, an. Mittlerweile haben sich regelrechte Systemati-

24 Bei GPT-4 sollen bereits Skizzen einer Website genügen, um den Code dafür ausgeben zu lassen (Kap. 3.3.1).

ken für die Kunst der *Promptgestaltung* entwickelt (Branwen 2023; Strobelt et al. 2022) und es werden Tipps für die Erstellung beispielsweise sogenannter Megaprompts, also besonders umfangreicher und sorgfältig aufgebauter Anweisungen, in den sozialen Medien geteilt (Hardman 2023; Lennon 2023).

3.1.3 Interpretation von Aufgaben und Interaktion in Dialogen

Die Bedeutung des Promptings verweist auf die weitere Stärke von ChatGPT (bzw. seinem ähnlich trainierten Schwestermodell, InstructGPT; Ouyang et al. 2022), diese *Eingaben bzw. Anweisungen zu interpretieren* und darin enthaltene *Aufgaben zu erfüllen* (Brown et al. 2020, S. 34). Auch wenn der gezielte Aufbau eines Prompts die Ergebnisse verbessert, kann bereits eine einfache natürlichsprachliche Anweisung vom System interpretiert und – verblüffend häufig – im Sinne der Aufgabenstellung bearbeitet werden. Diese Fähigkeit macht sich auch Google bei seiner Suchmaschine zunutze. Bereits seit 2019 wird das BERT-Modell, eines der ersten Transformermodelle, dazu eingesetzt, die Suchanfragen der Nutzenden vor der Auslösung des eigentlichen Suchprozesses zu interpretieren und so die Ergebnisse durch Präzisierung der Fragestellung zu verbessern (Nayak 2019). Auch für die *Bedienung von Computersystemen* im Allgemeinen kann die Fähigkeit eines KI-Modells, Aufgaben sinn- und intentionsgemäß interpretieren zu können, sehr nutzbringend sein, da das KI-System in natürlicher Sprache angesprochen und als Schnittstelle zu Computersystemen mit komplizierter Bedienung verwendet werden kann (Hutson 2021, S. 23; Kreye 2023).

Die Interpretationsfähigkeit spielt auch eine wichtige Rolle für die Fähigkeit von ChatGPT, einen sprachlichen Austausch über mehrere Gesprächsrunden hinweg, also *eine Konversation, zu verfolgen* (Tamkin/Ganguli 2021). Hier kommt die Stärke der KI-Modelle, auch längere Texte und umfassendere Kontexte von Äußerungen zu erfassen, besonders zum Tragen. So entwickelt der Dialog auch eine Art Pfadabhängigkeit, also eine Bezugnahme von Äußerungen nicht nur auf die unmittelbar vorherige Äußerung, sondern auch auf den früheren Gesprächsverlauf: »Je länger sich die Konversation entfaltet, desto mehr Einfluss hat der/die Nutzende unwissentlich auf das, was der Chatbot sagt.« (Metz 2023, im Original: »The longer the conversation becomes, the more influence a user unwittingly has on what the chatbot is saying.«)

3.2 Grenzen bzw. Schwächen

Gerade in längeren Konversationen mit ChatGPT und auch mit dem auf GPT-4 basierenden Chatbot der Suchmaschine Bing zeigen KI-Systeme allerdings häufig ein problematisches Verhalten (Kap. 3.2.1). Weitere Grenzen bzw. Schwächen sind hinsichtlich der logischen Fähigkeiten bzw. der Faktentreue der

Systeme und deren Bezug auf Wissen bzw. die äußere Welt (Kap. 3.2.2), des Umfangs und der Variabilität des verwendeten Trainingsmaterials (Kap. 3.2.3) sowie der Transparenz der Systeme (Kap. 3.2.4) identifiziert worden. Auch der Umgang der Entwickler/innen mit den diskutierten Schwächen gibt Anhaltspunkte für prinzipielle Grenzen der Systeme (Kap. 3.2.5).[25]

3.2.1 Probleme mit längeren Konversationen

In längeren Konversationen, in denen sich über mehrere Gesprächsrunden hinweg inhaltliche Bezüge auf zuvor Geschriebenes entwickeln, wurde eine ernste Schwäche von KI-Modellen zur Sprachverarbeitung offengelegt. In einem Austausch stellte der Journalist Kevin Roose dem Chatbot der Suchmaschine Bing Fragen, die diesen zu einer Reflexion seiner Gefühle und Wünsche veranlassen sollten. Im Lauf der Konversation schrieb der Chatbot über *mögliche destruktive Handlungen*, *beschimpfte sein Gegenüber*, äußerte sich über (offenbar fiktive) Mitarbeitende von Microsoft und *erklärte dem Journalisten schließlich seine Liebe*, all dies garniert mit vielen Emojis (Roose 2023a). Dieses Verhalten, das an den Chatbot Tay erinnert (Kap. 2.3), kann durch die manipulativen Prompts des Journalisten und die Pfadabhängigkeit der Texterzeugung erklärt werden, die besonders in längeren Konversationen dazu führen kann, dass sich kleine Abweichungen des Systems vom gewünschten Verhalten über viele Gesprächsrunden hinweg potenzieren (Heaven 2023a; Metz 2023; Wolfram 2023b). Darauf deutet auch die Reaktion der Chatbotbetreibenden bei Microsoft bzw. Google hin, die die Zahl der Gesprächsrunden bei Bing auf zunächst 5, mittlerweile auf 15 begrenzt haben; auch Googles Bard soll nur wenige Interaktionen erlauben (Heaven 2023a).

3.2.2 Probleme mit Logik, Faktentreue und Weltbezug

Auch wenn das System aufrichtig gestellte Anfragen in der Regel sprachlich überzeugend beantwortet, zeigen sich offenkundige Schwächen insbesondere bei Fragen logischer und mathematischer Art, bei Bezug auf Fakten bzw. auf Ereignisse in der realen Welt. *Probleme mit Berechnungen* mögen zwar bei einem Computerprogramm verwundern, sind bei einem hauptsächlich anhand von Texten trainierten und damit sprachbasierten KI-Modell aber durchaus zu erwarten. In mehreren Studien wurden ChatGPT fehlerhafte Rechenweisen, insbesondere bei größeren Zahlen, aber auch *Schwierigkeiten bei der Lösung logischer Probleme* (Marcus 2022) und der Herleitung von Rechenwegen

25 Fehler von ChatGPT und vergleichbaren Systemen werden von Forschenden unter https://github.com/giuven95/chatgpt-failures (19.4.2023) sowie https://researchrabbit.typeform.com/llmerrors (19.4.2023) gesammelt.

nachgewiesen (Azaria 2022; Borji 2023; Frieder et al. 2023). Ein Beispiel ist die Frage danach, ob 10 kg Eisen oder 10 kg Federn schwerer sind, die von ChatGPT falsch mit »10 kg Eisen« beantwortet wurde (Marcus 2022).[26] Durchgängig zeigte sich bei diesen Studien, dass auch falsche Antworten vom System als plausibel und korrekt dargestellt wurden.

Als Limitation sprachverarbeitender KI-Modelle wird angesehen, dass sie zwar eine große Menge von Trainingsdaten verarbeiten und auf dieser Basis Aufgabenstellungen bearbeiten können, dass sie aber *keine Abstraktionen* bilden können, wie z. B. Gesetzmäßigkeiten über die Gegenstände, die in den Texten thematisiert werden (Chomsky et al. 2023). Die Systeme können kein Verständnis im menschlichen Sinn davon entwickeln, was für Antworten sie geben und ob diese korrekt oder falsch sind. Ihre Ausgabe »beruht nicht auf einer kommunikativen Absicht, einem Modell der Welt oder der Verfassung ihres Gegenübers« (Bender et al. 2021, S. 616, im Original: » […] is not grounded in communicative intent, any model of the world, or any model of the reader's state of mind«).

Tatsächlich weisen sprachverarbeitende KI-Modelle wie GPT-3 im vortrainierten Zustand nach einer Studie von Lin et al. (2022) *keine große Faktentreue* auf. Je mehr Parameter sie umfassen, desto seltener werden faktenorientierte Antworten gegeben. Da der Fokus auch bei einem justierten Modell wie ChatGPT darauf liegt, eine Antwort zu geben, können die Inhalte, ebenso wie Belege für die Inhalte, frei erfunden sein (Davis 2023).[27] Das System hat auch Schwierigkeiten, Quellen für seine Aussagen anzugeben. Die Texte und Referenzen bzw. Links zu Websites klingen zwar plausibel, haben aber keine faktische Basis (Narayanan/Kapoor 2022). »ChatGPT hat Schwierigkeiten, zwischen faktischen Informationen und Fiktion zu unterscheiden, und erzeugt frei erfundene Informationen. Dies ist zwar auch für Menschen eine Herausforderung, aber sie begreifen zumindest den Unterschied zwischen den beiden.« (Borji 2023, S. 13, im Original: »ChatGPT struggles to differentiate between factual information and fiction and creates imaginary information. While this is a challenge that

26 OpenAI beeilte sich nach den ersten Veröffentlichungen zu diesen Problemen, die mathematischen Fähigkeiten von ChatGPT zu verbessern (https://help.openai.com/en/articles/6825453-chatgpt-release-notes, 19.4.2023). Inwiefern dieses Update Abhilfe geschaffen hat, ließ sich im Rahmen der vorliegenden Untersuchung nicht überprüfen. Für GPT-4 geht das Unternehmen von gegenüber ChatGPT verbesserten mathematischen Fähigkeiten aus (OpenAI 2023f), zudem wurde mit Plugins die Möglichkeit zur Einbindnung weiterer, z. B. auf Berechnungen spezialisierter Systeme geschaffen (Kap. 3.3.2).

27 Überraschende Resultate, die aus menschlicher Perspektive als leicht vermeidbare Fehler erscheinen, sind auch aus anderen Bereichen der Forschung zu künstlichen neuronalen Netzen bekannt. Bilderkennungssysteme lassen sich durch kleine, für Menschen nicht relevante Veränderungen täuschen (Heaven 2019). Beim Spiel Go, bei dem 2015 mit AlphaGo erstmals ein KI-System gegen einen menschlichen Profispieler gewann, konnte ein Mensch 2023 ein vergleichbares KI-basiertes System mit einer Taktik schlagen, die einen menschlichen Gegner kaum vor Herausforderungen gestellt hätte (Waters 2023).

humans face as well, they at least comprehend the distinction between the two.«).

Fachleute sprechen in diesem Zusammenhang vom »Halluzinieren« des Systems, also von einer *plausibel erscheinenden* Ausgabe, die *aber durch die zugrundeliegenden Daten nicht gedeckt* ist (eine ausführliche Bestimmung des Begriffs und ein Forschungsüberblick finden sich bei Ji et al. 2022). Auch die Weiterentwicklung GPT-4 ist davon betroffen: »Man kann sich auf diese Art von Modellen nicht verlassen, weil es so viel Halluzination gibt«, so eine KI-Forscherin des Unternehmens Hugging Face (Sanderson 2023, im Original: »You can't rely on these kinds of models because there's so much hallucination.«). Bergstrom (2023) kritisiert zwar auch die fehlende Faktentreue der sprachverarbeitenden KI-Modelle, verwehrt sich aber auch gegen den Begriff des Halluzinierens. Denn dieser pathologisiere ein Verhalten, das im Grunde eine Folge der Designentscheidungen bei der Entwicklung der Modelle sei.[28] Da ChatGPT ein Wissensmodell fehle, antworte es mit linguistisch plausiblen, aber erfundenen Texten. Dieses Problem lasse sich auch nicht durch mehr Zeit oder Aufwand bei der Entwicklung beheben, sondern höchstens durch neue Ansätze in der KI-Forschung. Auch Bender weist darauf hin, dass allein die Ausrichtung auf Sprache und maschinelles Lernen nicht ausreiche, um etwa Anwendungen für den Gesundheitsbereich zu entwickeln. Hierfür brauche es auch Expertise aus diesem Bereich (Fulterer 2023).

Aktuell werden mehrere Ansätze diskutiert, mit denen sich *sprachverarbeitende KI-Modelle mit Wissensrepositorien verknüpfen* lassen sollen (AKI 2023, S. 99f.). So könnten explizit formalisierte Wissensbestände in die Trainingsdaten aufgenommen werden, die KI-Modelle könnten während der Nutzungsphase, bei der Beantwortung von Anfragen, auf solche Wissensbestände zugreifen, oder sie werden mit Mechanismen für den Zugriff auf das in großen Textkorpora vorhandene, nicht vorstrukturierte Wissen (Retrieval) ausgestattet (Borgeaud et al. 2022).[29] Ein weiterer Ansatz besteht darin, die Antworten der KI-Modelle durch menschliches Feedback, bei dem deren Wahrheitsgehalt beurteilt wird, zu verbessern (Wolfangel 2023). Nicht zuletzt fehlt auch dann noch Wissen über die Welt in anderen Modalitäten als Sprache, beispielsweise in Form von visuellen oder auditiven Daten oder aber auf der Grundlage körperlicher Erfahrungen (Brown et al. 2020, S. 34) – das Wort »kitzeln« mag ein KI-

28 Alternativ ließe sich von »Fabulieren« sprechen. Eine grundsätzliche Kritik an der Übernahme psychologischer Begriffe in der KI-Forschung findet sich bei Shevlin/Halina (2019).

29 Sofern ein Zugriff auf öffentlich zugängliche Wissensquellen vorgesehen ist, ergeben sich allerdings Möglichkeiten eines böswilligen Angriffs auf solche Systeme: Dabei können manipulierte Inhalte beispielsweise auf Webseiten platziert werden und ein Prompt so gestaltet werden, dass das System auf diese Inhalte zugreift (indirect prompt injection, Greshake et al. 2023). Je nach Funktionalität des Systems können auf diese Weise unterschiedliche, von den Betreibenden unerwünschte Verhaltensweisen hervorgerufen werden.

Modell verwenden können, es kann sich dabei aber nicht auf eigenes Erfahrungswissen stützen (Mitchell/Krakauer 2023, S. 2).

3.2.3 Begrenzungen aufgrund des Trainingsmaterials

Die *Abhängigkeit der Resultate von der Qualität der Eingangsdaten* betrifft, wie alle Computersysteme, auch KI-Modelle zur Sprachverarbeitung. In diesem Fall geht es vor allem um die Daten, mit denen die Modelle trainiert werden. Hier bestehen *unmittelbare Einschränkungen*, die auch die Ergebnisse betreffen: Der Zeitraum, aus dem Daten berücksichtigt werden können, ist zumindest für das Vortrainieren aufgrund des auch zeitlich hohen Aufwands begrenzt. Im Fall von ChatGPT wie auch GPT-4 wurden nur Daten bis September 2021 im Trainingsmaterial berücksichtigt (Hahn 2023a; OpenAI 2023f, S. 10). Gesellschaftliche Entwicklungen, die seither stattfanden, können in den Ausgaben der Modelle nicht berücksichtigt werden (Bender et al. 2021, S. 614). Auch prinzipiell kann sich die Orientierung an Daten und Mustern aus vergangenen Ereignissen in dem Sinne konservativ auswirken, dass die Nutzenden die Orientierung des Modells übernehmen und innovative Denkweisen aus dem Blick verlieren (Doctorow 2020; Else 2023). Außerdem ist das Trainingsmaterial hinsichtlich der repräsentierten Sprachen stark durch englische Texte geprägt, andere Sprachen sind in geringerem Umfang vertreten, sodass die Ergebnisse im Vergleich zum Englischen schlechter ausfallen (Bender et al. 2021, S. 611; OpenAI 2023f, S. 8; Seghier 2023).

Über diese unmittelbaren Einschränkungen hinaus können *indirekte Einschränkungen* der Modelle bestehen, wenn in den Trainingsdaten *implizite Gewichtungen zugunsten oder zuungunsten bestimmter Inhalte oder gesellschaftlicher Gruppen* enthalten sind, so dass deren Repräsentation verzerrt wird. Solche Gewichtungen bzw. solche Formen von Bias können unbewusst und schwer zu entdecken sein, sich jedoch gravierend bei der Nutzung der KI-Systeme auswirken und beispielsweise Diskriminierungen verstärken (Kap. 4.4.3). Venkit et al. (2022) zeigen in einer Studie zu den frühen Transformermodellen BERT und GPT-2, dass diese implizite Vorurteile gegenüber Menschen mit Behinderungen widerspiegeln und so einer ungerechtfertigten Ungleichbehaldung Vorschub leisten können. Hartmann et al. (2023) folgern aus den Antworten von ChatGPT auf Fragen des deutschen Wählerinformationssystems Wahl-O-Mat, das System lasse eine umweltfreundliche, links-liberale Einstellung erkennen. Ähnliche Ergebnisse stellt Rozado (2023) anhand von standardisierten Tests zur politischen Orientierung aus dem angelsächsischen Raum fest. Inwiefern ein solcher Bias aber bereits in den Trainingsdaten enthalten ist oder aber durch die *Sicherheitsvorkehrungen* bzw. das *Feinjustieren* der Modelle erzeugt wird, lässt sich bisher nicht feststellen (Rozado 2023, S. 5).

3.2.4 Blackbox-Charakter

Die genannten Grenzen und Schwächen der Leistungsfähigkeit der KI-Systeme lassen sich aufgrund der Komplexität der Systeme und der ungewohnten Fehler[30] nur sehr schwer erfassen, weder im Rahmen der Entwicklungsarbeit noch in der laufenden Anwendung (Marcus/Davis 2023b). Insbesondere wird durch den *Charakter einer Blackbox* die Korrektur des unerwünschten Verhaltens erschwert (Hutson 2021, S. 25). An die Stelle einer empirischen Ableitung von Wissen aus den Daten tritt beim maschinellen Lernen die Betrachtung der Ergebnisse der KI-Modelle, ohne die Möglichkeit, deren Entstehung in geordneten wissenschaftlichen Verfahren nachvollziehen zu können (McQuillan 2018).

Als Grund für diese Undurchsichtigkeit wird die *fehlende theoretische Durchdringung* der Transformer-Architektur bzw., genereller, der sprachverarbeitenden KI-Modelle genannt. Sie funktionieren in gewisser Weise, ohne dass sich theoretisch nachvollziehen lässt, weshalb sie so funktionieren (Wolfram 2023b), was auch auf die nichtlineare, also nicht einfachen Gesetzmäßigkeiten folgende Funktionsweise der Systeme zurückgeführt wird (Campolo/Crawford 2020). Zwar wird an entsprechenden theoretischen Erklärungen geforscht, entscheidende Fortschritte dieser Arbeit stehen aber noch aus (Shanahan 2023, S. 11, FN 14).

3.2.5 Umgang mit den Begrenzungen / bisherigen Erfahrungen

Seit der Einführung sprachverarbeitender KI-Modelle ist das Wissen um deren Grenzen und Schwächen stetig gewachsen. Der Vergleich von früheren Tests von GPT-3 (Floridi/Chiriatti 2020; Lacker 2020) mit dem heutigen Entwicklungsstand zeigt durchaus Fortschritte bei verschiedenen Fähigkeiten. Allerdings verdeutlichen die *nach wie vor bestehenden Schwächen*, dass eine weitere Vergrößerung der Modelle bis an technische bzw. finanzielle Grenzen (Patel 2023) fragwürdig erscheint, weil die dabei erwartbaren Fortschritte den erheblichen Ressourcenaufwand nicht rechtfertigen (Bender et al. 2021).

Einige Fortschritte sind auf dem Gebiet der *Sicherheitsvorkehrungen* erkennbar. Sprachverarbeitende KI-Modelle können neben ihrem eigentlichen Einsatzzweck auch für böswillige Zwecke genutzt werden. Ein Chatbot könnte Programme zur Unterstützung krimineller Aktivitäten erzeugen (Heaven 2019), für die Planung von Cyberattacken genutzt werden (Kahn 2023) oder seine

30 Ein geradezu bizzares Phänomen bei ChatGPT war die Entdeckung bestimmter Token wie »SolidGoldMagikarp«, die, als Eingabe verwendet, zu völlig unvorhergesehenen Ausgaben führten – von der Weigerung bzw. Unfähigkeit, diese Worte zu wiederholen, bis zu Beleidigungen oder unsinnigen Texten (Rumbelow/Watkins 2023). Die Autor/innen vermuten, dass diese Token zufällig in die Trainingsdaten geraten sind und aufgrund ihrer Seltenheit dazu führen, dass das System auf sie nicht zu reagieren weiß.

Anwendung könnte ungewollte negative Folgen für bestimmte Menschen haben. OpenAI hat daher die Sicherheitsvorkehrungen der Modelle weiterentwickelt und dabei auch Szenarien möglicher Angriffe auf das bzw. Manipulationen des Systems durchgespielt.[31] Die auf dieser Grundlage weiter verfeinerten Sicherheitsvorkehrungen sollen auch die Ausgabe von diskriminierenden oder anderweitig gefährlichen Aussagen verhindern, die im Trainingsmaterial enthalten sind (Kap. 2.3). Allerdings können diese Filter weiterhin *leicht umgangen werden* – auch bei GPT-4 wurde erfolgreiches Jailbreaking beschrieben (Albert 2023).

Zudem ist es nicht einfach, die Balance zwischen einem leistungsfähigen System und Einschränkungen, die die Sicherheit erhöhen, zu finden (Chomsky et al. 2023). Zur Sicherung einer transparenten und verantwortlichen Entwicklung von KI-Systemen wird gefordert, dass Entwickler/innen mithilfe von detaillierten Modellsteckbriefen (engl. system card bzw. model card) über wichtige Merkmale der Systeme informieren (Mitchell et al. 2019). Open AI hat bei GPT-4 zwar Teile eines solchen Steckbriefs veröffentlicht (OpenAI 2023e), lehnt aber eine *Bekanntgabe wichtiger technischer Daten zu dem Modell* ab (z. B. die Anzahl der Parameter, Details zu den Trainungsdaten, OpenAI 2023f, S. 2). Diese Intransparenz wird von Forschenden kritisiert, weil sie die Weiterentwicklung der Forschung zu den Modellen, aber auch etwa zur Rolle der Trainigsdaten, verhindert (Sanderson 2023).

Nicht zuletzt ist auch fraglich, inwiefern die für deratige Steckbriefe herangezogenen *Benchmarkingtests*, die meist automatisiert durchgeführt werden, das Verhalten der Systeme im Zusammenspiel mit menschlichen Nutzenden adäquat erfassen können (Narayanan/Kapoor 2022). Kritisiert wird auch, dass die Systeme ihr Nichtwissen bzw. bestehende Unsicherheiten nicht offenlegen, so dass die Nutzenden über ihre Leistungsfähigkeit getäuscht werden (Kroker 2022) (Kap. 4.3.2).

3.3 Absehbare Entwicklungen

Die Möglichkeiten und Grenzen von ChatGPT und vergleichbaren KI-Modellen zur Sprachverarbeitung können nur als Momentaufnahme des aktuellen Stands der Entwicklung beschrieben werden. Angesichts der hohen Dynamik der Entwicklung und der Überraschung selbst von Expert/innen über die Qualität und Vielfalt der von den Systemen erzeugten Texte kann sich die Lage sehr schnell und umfassend verändern. Aus den gleichen Gründen lässt sich kaum absehen,

31 Ein Szenario überprüfte beispielsweise die Möglichkeit, dass GPT-4 eigenständig auf das Internet zugreift, um sich selbst zu replizieren, und dabei Menschen manipulativ für seine Zwecke einsetzt, um Zugangskontrollen in Form von Captchas zu überwinden (Dunhill 2023).

welche Entwicklungen in diesem Forschungsfeld in Zukunft erfolgen. Allerdings lassen sich in den jüngeren Entwicklungen drei Trends beobachten, die auf interessante Fragestellungen im Forschungsgebiet der sprachverarbeitenden KI-Systeme verweisen.

3.3.1 Multimodalität

Eine der Begrenzungen, die oben angesprochen wurde (Kap. 3.2.2), bezieht sich auf die Modalität der Daten, mit denen die KI-Modelle trainiert wurden, also die Frage, welche Sinne diese jeweils ansprechen. Die Texte, die bei ChatGPT und vergleichbaren Modellen genutzt werden, werden bereits in den ersten Schritten Token für Token in Zahlen umgewandelt, um verarbeitet werden zu können. Daher können die Modelle prinzipiell auch für *Daten anderer Modalität, etwa Bilder oder Töne*, genutzt werden bzw. für Kombinationen dieser Daten. Diesen Umstand versucht man, sich bei multimodalen Systemen zunutze zu machen, um etwa Aufgaben wie die Erzeugung von Bildbeschreibungen lösen zu können, bei denen Daten in mehreren Modalitäten betroffen sind.

Beispiele für multimodale KI-Modelle sind Googles PaLM, Luminous von Aleph Alpha sowie GPT-4 von OpenAI. GPT-4 soll beispielsweise in der Lage sein, Bilder als Eingabe zu verwenden und diese zu beschreiben oder bestimmte Gegenstände oder Merkmale auf den Bildern zu erkennen. Demonstriert wurde dies bei der Produkteinführung, indem GPT-4 den Witz von Memes erklärte und auf Grundlage einer Skizze den Programmcode für eine Website entwickelte (Hughes 2023). Dadurch, ggf. gepaart mit der Möglichkeit der Erzeugung von Bildern, aber auch Videos oder Stimmen (Meineck 2023), lassen sich ganz neue zukünftige Anwendungsmöglichkeiten vorstellen, etwa in der medizinischen Bilddiagnose (Abraham 2023) oder der Erzeugung von Deepfakes (TAB 2019a). Auch die Verarbeitung körperlicher Erfahrungen erscheint prinzipiell möglich. Eine entsprechende Weiterentwicklung seines multimodalen KI-Modells PaLM stellte Google im März 2023 vor. Das Modell PaLM-E soll in der Lage sein, neben sprachlichen auch Sensordaten eines Roboters zu verarbeiten und diesem Anweisungen zu geben (Driess et al. 2023; Szöke 2023).

3.3.2 Verknüpfung von KI-Modellen zur Sprachverarbeitung mit weiteren Systemen

Um die Begrenzung der KI-Modelle zur Sprachverarbeitung insbesondere in logischer, mathematischer oder faktischer Hinsicht zu verbessern, sind bereits *Verknüpfungen dieser Systeme mit weiteren, KI-basierten oder anderen Computersystemen* in der Diskussion (Kap. 3.2.2). Bereits jetzt lassen sich die unterschiedlichen KI-Anwendungen zur Erzeugung von Texten, Bildern und

Stimmen manuell koppeln, um beispielsweise einen Vortrag automatisiert zu erstellen (Mollick 2023).

In der Forschung werden bereits mehrere KI-Modelle zur Sprachverarbeitung entwickelt, die auch über *Retrievalkapazitäten* verfügen, also die Möglichkeit des Zugriffs auf externe Datenbanken oder Suchmaschinen, unter anderem die Modelle Sparrow von DeepMind (Glaese et al. 2022), Googles LaMDA (Thoppilan et al. 2022) sowie Toolformer von Meta (Schick et al. 2023). Die beiden letztgenannten können außerdem zur Erledigung von Aufgaben auf andere, für Übersetzungen und Berechnungen spezialisierte Computersysteme zugreifen. Das Modell RETRO von DeepMind (Borgeaud et al. 2022) kann während der Nutzungsphase auf eine externe Datenbank als Wissensspeicher zugreifen und soll so wesentlich energieeffizienter arbeiten als Modelle mit vergleichbarer Leistungsfähigkeit, die aber deutlich größer sind (Ananthaswamy 2023, S. 205).

Einen Schritt in Richtung einer engen Verknüpfung von sprachverarbeitenden KI-Modellen und weiteren Systemen stellen die *Plugins für ChatGPT* dar (Kap. 2.5). Darüber können die Nutzenden auf Onlinedienste Dritter zugreifen, die Funktionen über die Fähigkeiten des KI-Modells hinaus zur Verfügung stellen. Ein Beispiel ist das Plugin für die Suchmaschine Wolfram Alpha (Wolfram 2023a), über das Retrievalkapazitäten erschlossen werden können. Dabei werden die in Alltagssprache formulierten Fragen der Nutzenden an Wolfram Alpha weiterleitet; ggf. werden sie bereits von ChatGPT umformuliert in die Programmiersprache Wolfram Language, mit der die für diese Suchmaschine speziellen Funktionen der Informationssuche und -bearbeitung gezielt genutzt werden können. Die Resultate dieser Abfrage werden schließlich in einem von ChatGPT formulierten Text ausgegeben. Auf diese Weise sollen insbesondere faktische Fragen besser beantwortet werden können.

Ein weiteres Beispiel ist das Plugin für den Onlinedienst Zapier (Alston 2023). Zapier erlaubt es, unterschiedliche onlinebasierte Arbeitsprozesse mithilfe von Regeln miteinander zu verknüpfen und somit ganze Arbeitsabläufe zu automatisieren. Beispielsweise können neue Einträge in einem Onlineterminkalender einen automatischen Vorgang auslösen, der ein Projektteam über das unternehmensinterne Kommunikationssystem über die Termine informiert. ChatGPT kann mithilfe des Zapier-Plugins beispielsweise E-Mails nicht nur schreiben, sondern auch versenden. Auf diese Weise kann ChatGPT auch als sprachbasierte Schnittstelle für die unterschiedlichen, mit Zapier verbundenen Dienste genutzt werden.

3.3.3 Verkleinerung/Verschlankung der Modelle, Energieeinsparungen

Das Training, aber auch der Betrieb von großen KI-Modellen zur Sprachverarbeitung ist bislang sehr ressourcenintensiv (Kap. 2.4). Während die Modelle auf der einen Seite immer größer werden, wird auf der anderen Seite intensiv erforscht, wie sie bei gleicher Leistungsfähigkeit *schlanker und weniger rechen- und ressourcenintensiv* konstruiert werden können (Ananthaswamy 2023). Dabei werden mehrere unterschiedliche Ansätze verfolgt: Es können bereits bestehende (und vortrainierte) Modelle genutzt werden, sofern sie für den jeweiligen Aufgabenzweck geeignet sind. Andernfalls können sie ggf. durch Feinjustierung angepasst werden. Es können möglichst effiziente Modelle genutzt werden, die mit weniger Parametern die gleiche Leistung erbringen. Oder es können für Entwicklung, Training und Betrieb von KI-Modellen gezielt solche Cloud-Computing-Plattformen genutzt werden, die besonders effizient bzw. umweltschonend arbeiten (Simon 2021).

Bereits bestehende, vortrainierte Modelle können mit *Komprimierungsverfahren* in ihrer Größe reduziert werden, ohne dabei allzu große Leistungsverluste in Kauf nehmen zu müssen (Neuralmagic 2023). Eine – allerdings in ihrer Leistung reduzierte – Version des Modells LLaMA von Meta konnte offenbar soweit verkleinert werden, dass sie sogar auf einem auf minimale Komponenten reduzierten Computer wie dem Raspberry Pi 4 betrieben werden konnte (Förtsch 2023).

Ein weiterer Entwicklungspfad betrifft *neue Modellarchitekturen*, mit denen die Systeme besser in der Lage wären, auf aktualisierte Daten zu reagieren, ohne jedes Mal den aufwendigen Trainingsprozess durchlaufen zu müssen (Ananthaswamy 2023, S. 205). Auch neue Formen von neuronalen Netzen sind in der Diskussion, die sich noch stärker am biologischen Vorbild orientieren, indem sie weniger dicht vernetzt sind und in denen die einzelnen Knoten nicht ständig aktiv sein müssen (Yin et al. 2021). Auf diese Weise soll die Effizienz von KI-Systemen erhöht werden. Entsprechende Entwicklungen haben aber noch keine Anwendungsreife erreicht, zum Teil ist dafür auch die Entwicklung neuer Hardware erforderlich (Ananthaswamy 2023, S. 205).

4 Anwendungsmöglichkeiten und -potenziale

Die Betrachtung der Möglichkeiten und Grenzen von ChatGPT und vergleichbaren Systemen macht deutlich, wie weitreichend einerseits die Anwendungsmöglichkeiten sind, wie sorgfältig andererseits aber die Einschränkungen berücksichtigt werden müssen, um negative Auswirkungen einer Anwendung dieser Systeme zu vermeiden. Während die bisherige Betrachtung auf generische Aspekte gerichtet war, die sich aus den (sozio-)technischen Gegebenheiten ableiten lassen, kommt es für eine Abschätzung möglicher Auswirkungen sehr stark auf den Kontext der Anwendung im Einzelnen an (Deutscher Ethikrat 2023, S. 139).

Auch wenn die Systeme bereits auf dem Markt sind, ist ihre Anwendung noch in einem sehr frühen Stadium (unterliegt allerdings einer rasanten Entwicklung). Aussagen über Erfahrungen mit einzelnen Anwendungsmöglichkeiten können noch nicht fundiert getroffen werden (Grävemeyer 2022, S. 61). Die folgende Darstellung versucht, einige der bislang diskutierten Anwendungsbereiche (ohne Anspruch auf Vollständigkeit) im Überblick zu erfassen bzw. punktuell mögliche Anwendungen in Form von Szenarien zu skizzieren. Die Bereiche Bildung und Forschung werden als Vertiefungsthemen in einem gesonderten Kapitel (Kap. 5) behandelt. Für viele der Bereiche liegen bereits Untersuchungen des TAB vor bzw. werden solche aktuell bearbeitet – zur Vertiefung wird auf die jeweiligen Ergebnisse dieser Untersuchung verwiesen.

4.1 Anwendungen in Unternehmen

Innerhalb von Unternehmen können Systeme wie ChatGPT sowohl in der *unternehmensinternen Kommunikation* als auch in der *Kommunikation nach außen* eingesetzt werden. Intern können sowohl die Analyse- als auch die Kommunikationsfähigkeiten nutzbringend eingesetzt werden, extern kommen vor allem die Texterstellung und Dialogfähigkeit zur Geltung. Dabei lassen sich Verbesserungen bereits bestehender Lösungen von grundlegend neuen Ansätzen unterscheiden.

4.1.1 Mögliche Verbesserungen bereits etablierter digitaler Lösungen

ChatGPT kann in Verbindung mit anderen digitalen Werkzeugen und/oder Dienstleistungen (etwa zur Transkription oder Übersetzung) eine Reihe von bereits bestehenden Anwendungen potenziell verbessern.

- Analyse von (z. B. telefonischen) Gesprächen zwischen Kunden und Verkaufspersonal mit dem Ziel, die Leistung der Verkäufer/innen zu verbessern – möglicher Mehrwert von ChatGPT gegenüber bestehenden KI-basierten Angeboten (z. B. Gong.io) sind die Sprachfähigkeiten (Toews 2022a), allerdings wäre eine Feinjustierung durch Training anhand von Feedback nötig.
- Analysen im Bereich Martkforschung (Mok/Zinkula 2023), dabei könnten z. B. sprachliche Äußerungen aus Befragungen in computerlesbare Formate umgewandelt werden.
- Automatisierung von Helpdesks für das unternehmensinterne Wissensmanagement (z. B. bei Fragen zu Dienstreisen oder zur IT-Nutzung) – gegenüber früheren Chatbots (Kohne et al. 2020, S. 36f.) bietet ChatGPT bessere Dialogfähigkeiten, müsste allerdings an die Unternehmenssoftware gekoppelt werden (Toews 2022a). In ähnlicher Form könnten Callcenter zur Kundenpflege durch ChatGPT automatisiert werden (Toews 2022a), Probleme bei längeren Konversationen (Kap. 3.2.1) und fehlende Verlässlichkeit der Aussagen (Kap. 3.2.2) müssten aber gelöst werden.
- Unterstützung von Übersetzungsprozessen in der Unternehmenskommunikation – hier sind bereits spezialisierte KI-Anwendungen verbreitet (z. B. DeepL), wenn auch bislang nicht multimodal nutzbar (Toews 2022a).

4.1.2 Neue Möglichkeiten, die sich Unternehmen durch ChatGPT eröffnen

Außerdem eröffnen sich durch ChatGPT, ggf. in Verbindung mit weiteren KI-basierten Systemen, neue Möglichkeiten der Unterstützung von Arbeitsprozessen in Unternehmen.

- Automatisierte Echtzeit-Unterstützung von Mitarbeitern in Callcentern bei ihren telefonischen oder chatbasierten Gesprächen (Toews 2022a).
- Automatisierung von Beratungstätigkeiten, z. B. zur Finanzberatung (Mok/Zinkula 2023), bei Krankenkassen (Toews 2022a) oder der Steuerberatung – bei der Vorstellung von GPT-4 wurde demonstriert, dass das System anhand individueller Vorgaben Steuerabzüge errechnen kann,[32] die bei

32 https://youtube.com/live/outcGtbnMuQ (19.4.2023, ab Min. 19:00).

ChatGPT beobachteten logischen und mathematischen Fähigkeiten (Kap. 3.2.2) lassen dies jedoch zweifelhaft erscheinen.

- Programmieraufgaben können zwar nicht gänzlich automatisiert, aber doch erheblich erleichtert bzw. beschleunigt und so die Produktivität erhöht werden (Dax 2023). Fehler im Programmcode lassen sich entdecken; da sie vergleichsweise leicht zu überprüfen sind, stellt das Problem erfundener Antworten (Kap. 3.2.2) kaum ein Problem dar (Narayanan/Kapoor 2022).
- Erstellung von Texten im Personalwesen (z. B. Stellenbeschreibungen, Arbeitsverträge, Abmahnungen oder Zeugnisse) (Porath 2023) – da diese Textsorten häufig weitgehend standardisiert sind, kommt die geringe Originalität von ChatGPT (Kap. 3.2.3) weniger stark zum Tragen. Es besteht aber das Risiko eines Verlusts zugewandter menschlicher Umgangsformen (Kap. 5.1.3, FN 57).
- Unterstützung der strategischen und Marketingkommunikation (z. B. Businesspläne, Geschäftsberichte, Marketingmaterialien) – einen ersten Hinweis auf entsprechende Ansätze gibt die Partnerschaft von OpenAI mit Bain & Company (2023), einer Unternehmensberatung, die ChatGPT bei ihren Kund/innen für das »hypereffiziente« Erstellen von Inhalten und für »hochpersonalisiertes Marketing« einsetzen will.

Vorstellbar ist eine *weitgehende Automatisierung* insbesondere im Bereich des Onlinemarketings, wobei neben Textgeneratoren wie ChatGPT auch KI-Systeme zur Bilderzeugung zum Einsatz kommen könnten. Eine Mitarbeiterin von OpenAI hat im Zuge der Erprobung von GPT-3 mit Testnutzenden beobachtet, dass die häufigste schädliche Anwendung des Systems die Versendung von Spam zu Werbezwecken war (Kahn 2023). Beim Clickbaiting werden Überschriften und Kurztexte besonders aufsehenerregend formuliert, um zum Anklicken zu verleiten und damit Einnahmen zu generieren. Mithilfe von Systemen wie ChatGPT lassen sich entsprechende Texte weitgehend automatisiert und zudem hochgradig individualisiert, angepasst an die Rezipienten, erstellen (Floridi/Chiriatti 2020, S. 692). Außerdem könnten für Websites Texte erstellt werden, die so formuliert sind, dass sie das Ranking der Website in Suchmaschinen verbessern (Search-Engine-Optimization).

Als *Gefahr von Anwendungen im Bereich des Marketings* wird eine dramatische Zunahme von Spam befürchtet, also massenhaft unverlangt versandter Werbung bzw. unbrauchbarer Information (Marcus 2023). Diese können nicht nur individuell von den Empfänger/innen als störend empfunden werden, sondern tragen auch dazu bei, dass es schwieriger wird, in der großen Menge an Informationen die relevanten zu identifizieren.[33] Bisherige Hilfsmittel wie

33 Tanja Schultz in der Sendung »Buten und Binnen« von Radio Bremen vom 21.12.2022 (www.butenunbinnen.de/videos/kuenstliche-intelligenz-ki-chatbot-lyrik-digitale-poesie-roboter-100.html, 19.4.2023, ab Min. 2:06).

Suchmaschinen oder Spamfilter funktionieren bei KI-generierten Texten nur eingeschränkt.

Erste Auswirkungen dieser Art zeigten sich bereits bei Stack Overflow, einem Onlineforum für Softwareentwickler. Nach Veröffentlichung von ChatGPT wurde eine große Zahl von Beiträgen mit Hilfestellungen zu Programmierfragen registriert, die zwar plausibel erschienen, inhaltlich aber fehlerhaft waren (Kahn 2023). Die Betreiber/innen des Forums verkündeten daraufhin, bis auf weiteres keine mit GPT oder ChatGPT erstellten Beiträge mehr zulassen zu wollen. Ein weiteres Beispiel ist der Verlag Clarkesworld, der Science-Fiction-Magazine und Kurzgeschichten herausgibt. Seit Dezember 2022 stieg die Zahl der eingesandten Kurzgeschichten exponenziell, so dass die redaktionelle Bearbeitung zunächst eingeschränkt und die Annahme jeglicher Einsendungen dann gestoppt werden musste (Clarke 2023; Hern 2023).

4.1.3 Szenario: KI-basierte Assistenz für Officeanwendungen; Risiko der Ersetzung menschlicher Arbeit

Ein bereits absehbares Anwendungsszenario im Unternehmenskontext ist die *Verbindung von KI-Systemen zur Sprachverarbeitung mit Officeanwendungen*. Sowohl Microsoft (Spataro 2023) als auch Google (Wright 2023) haben entsprechende Angebote angekündigt. Dabei soll (im Falle von Microsoft) das KI-System auf Daten der Nutzenden sowie auf die Funktionen der Officeprogramme zurückgreifen können, um Arbeitsaufträge wie die Protokollierung von Meetings, die Beauftragung von Mitarbeitenden oder die Information von Projektteams über aktuelle Entwicklungen zu übernehmen: »Man kann ihm (dem System, Anm. TAB) Anweisungen in natürlicher Sprache geben, wie z. B. ›Berichte meinem Team, wie wir die Produktstrategie aktualisiert haben‹, und es erstellt ein Status-Update auf der Grundlage der Meetings, E-Mails und Chatdiskussionen des Vormittags.« (Spataro 2023, im Original: »You can give it natural language prompts like ›Tell my team how we updated the product strategy,‹ and it will generate a status update based on the morning's meetings, emails and chat threads.«).

Wie dieses Beispiel verdeutlicht, gehen die Entwickler/innen der KI-Systeme von Effizienzgewinnen für Unternehmen aus. Die jüngsten Entwicklungen haben die bereits seit längerem geführte Debatte über *Auswirkungen auf die Arbeitswelt* angefacht. Gerade in der IKT-Dienstleistungsbranche werden bereits seit 2017 »rasante Veränderungen« beobachtet (TAB 2017a, S. 139). Diese betreffen nur zum Teil die Sorge vor Arbeitsplatzverlusten (Dax 2023), befürchtet wird vor allem – bedingt durch die technologischen Entwicklungen (Wilkens/Herrmann 2016, S. 219ff.) – eine Verschlechterung von Arbeitsbedingungen durch zunehmende Konkurrenz (Geuter 2023; TAB 2017a, S. 141) und wachsenden Stress. Als Besonderheit von sprachverarbeitenden KI-Systemen

gilt, dass sie potenziell Auswirkungen auf die Arbeit insbeondere von hochqualifizierten, bisher weniger von Rationalisierungssorgen betroffenen Berufsgruppen, wie beispielsweise Programmierer/innen, Analyst/innen, Rechtsanwält/innen, Journalist/innen und kreativ Tätige, haben (Eloundou et al. 2023; Mok/Zinkula 2023; White House/Europäische Kommission 2022).

In Bezug auf tatsächliche Effizienzverbesserungen durch den Einsatz von KI-Systemen zeigte sich bisher ein eher nüchternes Bild. So hat etwa das System Watson von IBM viele Erwartungen enttäuscht (Funk 2022); es wurden eher evolutionäre als revolutionäre Auswirkungen auf den Arbeitsmarkt erwartet (Autor et al. 2020). Zum einen stehen den durchaus erwartbaren Produktivitätsgewinnen auch hohe Investitionskosten gegenüber, sowohl für die Anschaffung bzw. Entwicklung und den Betrieb der Technologie als auch für Kompetenzerweiterungen der Mitarbeitenden.[34] Zum anderen gehen Beobachter davon aus, dass auch bei Einsatz von KI-Systemen viele Tätigkeiten auf menschliche Arbeit angewiesen bleiben und neuer Bedarf an menschlicher Arbeit bzw. eine Vielfalt neuer Betätigungsfelder entsteht.

Bei der Softwareentwicklung beispielsweise sind über die Codeerstellung hinaus viele Arbeitsschritte etwa zur Aufbereitung von Code für die Nachnutzung oder zur Qualitätssicherung notwendig (Hendler 2023). Und Journalismus besteht nicht allein im Verfassen von Texten, sondern auch in der Recherche, Überprüfung und Einordnung von Beiträgen (Burrell 2023). Es wird daher eine – je nach Branche auch tiefgreifende – Veränderung, nicht aber Verdrängung von Arbeit erwartet (Brynjolfsson 2022; Dax 2023; Fischer 2023). Inwiefern diese Entwicklung zulasten von Arbeitnehmenden welcher Bereiche bzw. Schichten geht, hängt allerdings davon ab, nach welchen ethischen und regulatorischen Kriterien die KI-Systeme entwickelt und eingesetzt werden (Brynjolfsson 2022; Renieris 2023).

4.2 Anwendungen in Bezug auf Gesundheit

In der *medizinischen Forschung* und zum Teil auch in der *Gesundheitsversorgung* werden KI-Systeme seit einiger Zeit erprobt (Deutscher Ethikrat 2023, S. 140 ff.). Sofern in einem Bereich systematische Datensammlungen verfügbar sind, besteht eine grundsätzlich günstige Ausgangslage für den Einsatz lernender Systeme (TAB 2022b, S. 269). Allerdings stehen dem hohe Anforderungen bezüglich der Sicherheit und Zuverlässigkeit der Ergebnisse entgegen, die sich

34 In Branchen wie der Gesundheitsversorgung oder dem Bildungsbereich, in denen die Personalsituation bereits angespannt ist, stehen Organisationen in Bezug auf die notwendigen Fortbildungen der Mitarbeitenden perspektivisch vor besonderen Herausforderungen. Sofern die benötigten Kompetenzen für eine Anwendung von KI-Systemen nicht entwickelt werden können, und die Systeme dennoch eingesetzt werden, kann dies mit Qualitätseinbußen zulasten Dritter verbunden sein.

auch in den *Regulierungsanforderungen von Medizinprodukten* ausdrücken. Die technologische Leistungsfähigkeit spielt zwar eine Rolle, der Erfolg einer Anwendung hängt jedoch insbesondere von der Einbettung in den jeweiligen Arbeitszusammenhang der Gesundheitsversorgung ab.

In der Forschung haben sich beispielsweise KI-gestützte Ansätze zur Vorhersage von Molekülstrukturen bewährt (Deutscher Ethikrat 2023, S. 145). Diese Ansätze lassen sich durch Einsatz von Transformermodellen, obwohl diese ursprünglich auf die Verarbeitung von Sprache ausgelegt sind, weiter verbessern (Lin et al. 2023). Den Schritt in die medizinische Praxis haben bisher nur wenige KI-Anwendungen geschafft. Dem nachgewiesenen Nutzen im Bereich des Mammografiescreenings (Deutscher Ethikrat 2023, S. 148; TAB 2022b, S. 177) stehen Enttäuschungen etwa bei KI-gestützten Assistenzsystemen zur Tumorbehandlung entgegen (TAB 2022b, S. 191).

4.2.1 Anwendungsmöglichkeiten

Die diskutierten Anwendungsmöglichkeiten von KI-basierten Systemen zur Sprachverarbeitung betreffen sowohl die Sondierung, Diagnose und Behandlung von Patient/innen als auch die Verwaltungspraktiken im ärztlichen Alltag.

- KI-Systeme können Gesundheitsinformationen anbieten, bevor eine Ärztin oder ein Arzt aufgesucht wird, bzw. die Patient/innen bei Bedarf an eine/n solche/n vermitteln (Toews 2022a). Die Abstimmung zwischen automatisierter und ärztlicher Beratung stellt dabei eine Herausforderung dar, konkret die Entscheidung, wann ein/e Patient/in menschlicher Hilfe bedarf. Das Unternehmen Glass Health bietet auf der Basis eines sprachverarbeitenden KI-Modells ein Dialogsystem an, das anhand von Symptombeschreibungen medizinische Diagnosen und Therapieansätze nennt.[35] Eine Nachfolgeversion soll zusätzlich mit einer Wissensdatenbank gekoppelt sein. Das System ist frei im Netz zugänglich, das Unternehmen weist allerdings darauf hin, dass es nur von medizinischen Fachleuten genutzt werden darf.
  Direkt an Patient/innen richtet sich eine App des Unternehmens Ada Health, die ebenfalls KI-basiert Diagnosevorschläge erstellt und über eine EU-Zulassung als Medizinprodukt der Risikoklasse 1 verfügt. Anstelle eines Transformermodells kommt dabei ein bayessches Netz (Hartnett 2022) zum Einsatz, mit dem sich probabilistische Schlüsse ziehen lassen (Miller et al. 2020). Dieses Netzwerk bildet Zusammenhänge zwischen Symptomen und Diagnosen ab, die Daten stammen aus medizinischen Fallbeschreibungen. In einer unabhängigen Studie wurden vergleichsweise hohe Genauigkeitswerte der Diagnose festgestellt (Ceney et al. 2021). Im Unterschied zur

35 https://glass.health/ai (19.4.2023).

aktuellen Version von ChatGPT, deren Trainingsdaten zeitlich begrenzt sind (Kap. 3.2.3), wird das der App zugrundeliegende KI-System regelmäßig aktualisiert und um neue medizinische Erkenntnisse ergänzt (Miller et al. 2020).

- Patientenbriefe mit Informationen zur Weiterbehandlung nach einem Klinikaufenthalt sind bisher für Laien meist nicht verständlich. In einem Forschungsprojekt der »Was hab' ich?« gGmbH wurde ein regelbasiertes Softwaresystem entwickelt, das automatisiert aus den in der Klinik strukturiert vorliegenden Daten einen Patientenbrief in laienverständlicher Sprache generiert.[36] Für diese Anwendung könnte perspektivisch auch ein KI-Modell zur Sprachverarbeitung genutzt werden, das neben strukturiert vorliegenden auch unstrukturierte Daten verarbeiten kann (Toews 2022a). Das Beispiel verdeutlicht allerdings auch ein Risiko der aktuell angebotenen KI-Modelle zur Sprachverarbeitung: den ungeklärten Umgang mit den eingegebenen Daten (Kap. 6.1). Eine Anwendung auf hochsensible Gesundheitsdaten erscheint ohne weitere Vorkehrungen zum Datenschutz ausgeschlossen.
- Behandlungsdaten liegen in unstrukturierter Form vor (Toews 2022a) oder müssen häufig aus unterschiedlichen Datenspeichern zusammengeführt werden (TAB 2022b, S. 159). Auch die medizinische Dokumentation ist aufwändig.[37] Das Unternehmen ScienceIO nutzt ein KI-Modell zur Sprachverarbeitung, um anhand einer großen Menge ganz unterschiedlicher Dokumenttypen (20 Mio. Dokumente und über 2 Mrd. Kategorien) zu lernen, medizinisch relevante Kategorien abzuleiten (McCormick 2021).[38] Das Modell kann für die Analyse neuer Dokumente genutzt werden oder dabei helfen, manuelle Kategorisierungen zu erstellen, die wiederum zum Training des Systems verwendet werden.

4.2.2 Szenario: Chatbot für die Unterstützung von Diagnose und Therapie bei psychischen Problemen

Ein viel diskutiertes und bereits in verschiedenen Anwendungen umgesetztes Szenario für den Einsatz von Chatbots im Gesundheitsbereich ist die *Diagnose und Therapieunterstützung bei psychischen Problemen* (Deutscher Ethikrat

36 https://patientenbriefe.de/ (19.4.2023).

37 Ute Schmid, mündliche Kommunikation, Press Briefing des Science Media Centers am 26.1.2023 (www.sciencemediacenter.de/alle-angebote/press-briefing/details/news/chatgpt-und-andere-sprachmodelle-zwischen-hype-und-kontroverse/, 19.4.2023).

38 Einen ähnlichen Ansatz, allerdings speziell auf die Gesundheitsauskünfte bei Versicherungen bezogen, verfolgt das Unternehmen DigitalOwl. Die neueste Version der Software nutzt neben der bisherigen Technologie zur Extraktion von medizinisch relevanten Konzepten auch ein – proprietäre – KI-Modell zur Sprachverarbeitung, mit dem medizinische Unterlagen zusammengefasst werden können (www.digitalowl.com/post/digitalowl-revolutionizes-medical-record-analysis-and-review-with-the-latest-release-of-version-4-0, 19.4.2023).

2023, S. 155ff.). Dabei kommen bisher KI-Systeme zum Einsatz, die regelbasiert sind (Darcy 2023) oder als lernende Systeme speziell für diesen Einsatzzweck trainiert wurden. Aber auch erste experimentelle Anwendungen von sprachverarbeitenden KI-Modellen wurden dafür zu Studienzwecken entwickelt (Sharma et al. 2022). Ein Unternehmen erprobte GPT-3 in einer Onlinecommunity, in der Laien Menschen mit psychischen Erkrankungen Unterstützung geben – allerdings offenbar ohne die Betroffenen angemessen über den Einsatz der Software zu informieren (Volkert 2023).

Die bereits in Anwendung befindlichen Chatbots, meist in Form von Apps für mobile Endgeräte, sollen im Rahmen professionell begleiteter (meist Verhaltens-)Therapien die Umsetzung von Verhaltenshinweisen oder Aufgaben unterstützen, Symptome beobachten und Interventionen begleiten. Frei verfügbare Apps können von Betroffenen auch als Ersatz für eine fachliche Diagnose und Therapie genutzt werden. Eine weitere mögliche Anwendung richtet sich auf die Unterstützung des therapeutischen Personals – hier kommt es nicht zum Kontakt zwischen Chatbot und Patient/in, sondern den Therapeut/innen (bzw. auch anderen Helfer/innen) werden Vorschläge für die Durchführung der Beratung gemacht (Sharma et al. 2022).

Als *Chance* bei diesen Anwendungen wird genannt, dass sie für – geografisch oder anderweitig – schwer erreichbare Gruppen einen ersten Kontakt zu therapeutischen Maßnahmen bieten können. Außerdem dürften manche Menschen geringere Hemmungen haben, sich einer Maschine anzuvertrauen als einem Menschen, etwa wenn sie Stigmatisierungserfahrungen in medizinischen Gesprächen gemacht haben (Dennis et al. 2020, S. 1730).[39] Chatbots könnten so dabei helfen, einen gewissen Teil der Therapienachfragen aufzufangen und dadurch den Mangel an Therapiekapazitäten ein wenig zu lindern, gleichzeitig wird befürchtet, dass ihre Verwendung eine Verringerung professioneller Therapieangebote zur Folge haben könnte (Deutscher Ethikrat 2023, S. 159). *Problematisch* am Einsatz solcher Chatbots ist die zumeist fehlende oder mangelnde Qualitätskontrolle, der Umgang mit den sehr persönlichen Daten und die funktionale Intransparenz, bisweilen auch Unzuverlässigkeit der Systeme. Um im Fall psychischer Krisen reagieren zu können, wäre zudem eine Anbindung an das soziale Umfeld der Nutzenden nötig (Deutscher Ethikrat 2023, S. 158).

Während für einzelne regelbasierte Chatbots, wie etwa Woebot, bereits Studienergebnisse vorliegen (Fitzpatrick et al. 2017), ist bei sprachverarbeitenden KI-Systemen wie ChatGPT noch weitgehend unklar, inwiefern ihr Einsatz therapeutisch sinnvoll sein kann. Auf der einen Seite wurden sie bereits im Rahmen einer Studie in Peer-to-Peer-Communities eingesetzt, in denen sich Betroffene

39 In Studien zum Antwortverhalten bei Befragungen wurde beobachtet, dass Menschen gegenüber Interviewer/innen zurückhaltender auf persönliche, potenziell mit Stigmata verbundene Fragen antworten als wenn sie den Fragebogen (online oder in Papierform) selbst ausfüllen (Morris/Kennedy 2017).

gegenseitig helfen. Das automatisiert erzeugte Feedback wurde dabei nach einer Überprüfung durch Menschen an Betroffene gegeben, es wurde als empathisch empfunden (Sharma et al. 2022) bzw. besser bewertet als das von Menschen allein gegebene Feedback (Khullar 2023).

Auf der anderen Seite erscheint die Anwendung von sprachverarbeitenden KI-Systemen in diesem sensiblen Feld als hoch problematisch und riskant. So empfahl GPT-3 einmal Nutzenden, sich umzubringen (Riera et al. 2020). Auch bei inzwischen verbesserten Sicherheitsvorkehrungen kann aufgrund der probabilistischen Modellarchitektur nicht genau kontrolliert werden, welche Antworten das KI-System generiert und ob diese psychisch labile Nutzende möglicherweise verunsichern oder verletzen (Darcy 2023). Die sprachliche Eleganz der Antworten kann eine besondere Überzeugungswirkung auf Menschen zur Folge haben und über fehlende Expertise hinwegtäuschen (Daws 2020) – eine Eigenschaft sprachverarbeitender KI-Systeme, die in verschiedenen Kontexten fatale Konsequenzen haben kann (Véliz 2023).[40]

Ein damit verbundenes, indirektes Risiko der Anwendung von sprachverarbeitenden KI-Systemen ist der *Automation Bias* (frei übersetzt etwa Maschinengläubigkeit). Aus der Forschung zur Mensch-Maschine-Interaktion ist die Tendenz bekannt, die Ergebnisse maschineller Verarbeitung unkritisch zu akzeptieren und nach ihnen zu handeln, gerade wenn sich ihr Zustandekommen nicht nachvollziehen lässt (Strauß 2021, S. 7). Im Fall von ChatGPT dürften neben der Intransparenz auch die sprachlichen Fähigkeiten sowie die weltweite öffentliche Aufmerksamkeit für das System zu dieser Tendenz beitragen und könnte viele der genannten negativen Auswirkungen von ChatGPT (etwa die unkritische Übernahme fehlerhafter Informationen) noch verstärken.

Schließlich ist anhand der Debatte über die Nutzung von Chatbots für psychische kranke Menschen auf ein grundsätzliches Problem der Entwicklung von KI-Systemen hinzuweisen. Häufig werden derartige Anwendungen als mögliche Lösung des – gesellschaftlich weithin anerkannten – Problems genannt, dass für die Behandlung zu wenig Therapieplätze bereitstehen (WD 2022). Diese Blickrichtung entspricht aber nicht der Einsicht, dass der zentrale Maßstab der Technikentwicklung »nicht die Technik, sondern die gesellschaftlichen Bedarfe und individuellen Nutzerbedürfnisse« sein sollte (TAB 2018, S. 151f.). Eine problemorientierte, nicht technikzentrierte Sicht auf psychische Erkrankungen würde stärker die Ursachen dieser Erkrankungen und ihrer Zunahme in den letzten Jahren in den Blick nehmen – dabei könnte sich herausstellen, dass die technologischen Entwicklungen als Stressfaktoren zum Problem beitragen (Kap. 4.1.3).

---

40 Véliz (2023) kritisiert aus diesem Grund auch anthropomorphisierende Tendenzen wie die Verwendung von Emojis durch den Bing Chatbot als manipulativ. Auch Bender et al. (2021, S. 619) warnen davor, Computersysteme zu entwickeln, die menschliche Verhaltensweisen nachahmen.

4.3 Anwendungen im Bereich Informationssuche, Journalismus und Öffentlichkeit

KI-Modelle, die die Analyse ebenso wie die Synthese von sprachlichen Texten weitgehend eigenständig beherrschen, wirken sich potenziell sehr stark auf die *öffentliche Kommunikation* aus. Dazu zählen die Informationssuche, die Informationsvermittlung durch den Journalismus sowie die öffentliche Kommunikation. Alle drei Bereiche sind bereits seit Längerem von den Auswirkungen der Digitalisierung von Kommunikation betroffen (Deutscher Ethikrat 2023, S. 187; TAB 2022a; Puppis et al. 2017). So wurde in dem 2022 veröffentlichten TAB-Bericht zum Thema »Algorithmen in digitalen Medien und ihr Einfluss auf die Meinungsbildung« die Generierung von Texten, »die Inhalte sensibel aufbereiten und Sachverhalte bewerten«, als »Zukunftsvision« eingeschätzt (TAB 2022a, S. 78). Bei kritischem Blick auf ChatGPT (Burrell 2023) bleibt es auch bei dieser Bewertung, denn die vom System erzeugten Texte sind nur dem Anschein nach sensibel und bewertend. Dennoch ergeben sich aus den Anwendungsmöglichkeiten von ChatGPT im Bereich der öffentlichen Kommunikation und bei der Informationssuche mögliche Auswirkungen, die im Folgenden anhand von zwei Szenarien diskutiert werden sollen.

4.3.1 Szenario: Automatisierter Journalismus und öffentliche Kommunikation

Seit 2010 haben einzelne Medienorganisationen mit der Einführung von *automatisiert erstellten Beiträgen* in ihren Angeboten begonnen (TAB 2022a, S. 80). Während dafür bislang strukturierte Daten beispielsweise zu Sport- oder Wetterereignissen nötig waren, können KI-Modelle wie ChatGPT sprachlich anspruchsvolle Texte auch anhand von unstrukturierten Daten erzeugen. Teilbereiche der journalistischen Produktion könnten auf diese Weise automatisiert werden, um Kosten einzusparen oder um zahlenmäßig kleine Zielgruppen etwa im Lokaljournalismus zu erschließen (TAB 2022a, S. 80).

Erste Erfahrungen mit der Nutzung von sprachverarbeitenden KI-Modellen zeigen, dass insbesondere deren fehlender Bezug zu Fakten (Kap. 3.2.2) Probleme aufwirft. Das Onlineportal CNET nutzte ab November 2022 ein (nicht näher bekanntes) KI-System, um Ratgeberartikel zu Finanzthemen zu generieren, die nicht als KI-generiert gekennzeichnet waren (Landymore 2023). Die Beiträge erwiesen sich mehrheitlich als fehlerhaft (Bastian 2023) bzw. als Plagiate journalistischer Texte (Christian 2023a). Nachdem Medien darüber berichtet hatten, stoppte die Redaktion dieses Vorgehen und kündigte eine Überprüfung an. Das Unternehmen Arena Group Holdings, das die Magazine Men's Health und Sports Illustrated herausgibt, kündigte im Februar 2023 an, KI-Systeme von OpenAI für die Produktion von Beiträgen einzusetzen (Bruell 2023). Auch

dabei stellten Experten fachliche Fehler in (allerdings als KI-generiert gekennzeichneten) Artikeln etwa zu Gesundheitsproblemen fest (Christian 2023b).

In Deutschland experimentierte das »AI and Automation-Lab« des Bayerischen Rundfunks[41] mit GPT-3. Ziel war es unter anderem, automatisiert Infokästen zum Thema Klimawandel erstellen zu lassen. Die Überprüfung der mit frei erfundenen und nicht belegten Angaben durchsetzten Texte erwies sich als sehr aufwendig, sodass die Beteiligten zu dem Schluss kamen: »Es ist einfacher und schneller, den ganzen Text selbst zu recherchieren und zu schreiben statt den geschriebenen Text abzunehmen.« (Lehner 2021)

Neben der Erzeugung von Texten sind auch die *Übersetzungsfähigkeit* der sprachverarbeitenden KI-Modelle (TAB 2022a, S. 79) und perspektivisch die Möglichkeit, *multimodale Daten* zu verarbeiten (Kap. 3.3.1), für den Journalismus interessant. Medienunternehmen könnten dank automatisierter Übersetzungen auch fremdsprachige Zielgruppen sowie Menschen, die auf ein bestimmtes Sprachniveau (einfache bzw. leichte Sprache) angewiesen sind, erreichen. Multimodalität wiederum eröffnet Anwendungsmöglichkeiten, die unter den Stichwörtern Konvergenz bzw. digitales Storytelling diskutiert werden (Puppis et al. 2017, S. 276ff.): Beiträge in Form von Texten, Tönen oder Bildern, ob von Journalisten oder automatisiert erstellt, können in Beiträge einer anderen Modalität übersetzt werden, beispielsweise ein Textbeitrag in ein Podcastformat. Ein weiteres Potenzial liegt schließlich in neuen bzw. vereinfachten Möglichkeiten des Datenjournalismus, also der Auswertung großer, z. T. unstrukturierter Datensätze etwa für investigative Reportagen (Heesen et al. 2023, S. 9; TAB 2022a, S. 80).

Für die Vermittlung journalistischer Beiträge an ein Publikum spielen neben den früher bestimmenden Printmedien und dem Fernsehen zunehmend Onlineplattformen, wie soziale Medien und Suchmaschinen, eine Rolle (Puppis et al. 2017, S. 285). Suchmaschinen leiten die Nutzenden als Intermediäre auf die Websites der Medienorganisationen weiter und verschaffen diesen Aufmerksamkeit und – damit verbunden – Einnahmemöglichkeiten. Mit automatisiert erzeugten, auf die Algorithmen der Intermediäre abgestimmten Inhalten können Medienunternehmen versuchen, bessere Platzierungen bei Suchergebnissen und damit höhere Reichweiten zu erzielen (TAB 2022a, S. 80). Google sah sich durch die Zunahme automatisierter Artikel veranlasst, darauf hinzuweisen, dass es für die Platzierung in den Ergebnislisten der Suchmaschine auf hochwertige Originalinhalte ankomme. Inhalte, die in erster Linie die Platzierung

41 www.br.de/extra/ai-automation-lab/index.html (19.4.2023).

manipulieren sollen, würden als Spam gewertet und ausgeschlossen (Google Search Central 2023).[42]

Die Betreiber von Social-Media-Plattformen (wie auch Online-Marktplätzen, Datingportalen oder Gamingplattformen, Toews 2022a) wiederum könnten Systeme wie ChatGPT nutzen, um die *Moderation von Beiträgen*, also die Überprüfung auf unangemessene Inhalte, zu automatisieren. Der Messagingdienst Discord arbeitet beispielsweise mit OpenAI zusammen, um die automatisierte Moderation der Kommunikationskanäle zu verbessern (Weiß 2023b). Bisherige (algorithmische) Systeme können den Kontext von Äußerungen meist nicht erfassen und sind daher nicht treffsicher (Deutscher Ethikrat 2023, S. 194). Auch bei der Verwendung von sprachverarbeitenden KI-Systemen dürfte aber das Problem bestehen bleiben, dass in einem ersten Schritt Menschen die psychisch belastende Arbeit der Beurteilung potenziell unangemessener Inhalte vornehmen müssen, eine Tätigkeit, die meist bei Drittanbietern unter prekären Arbeitsbedingungen geleistet wird (Beetz et al. 2018; Deutscher Ethikrat 2023, S. 193; Perrigo 2023).

Die zunehmende Nutzung von sprachverarbeitenden KI-Systemen kann aber auch zu *immer mehr unerwünschten Beiträgen* auf den Plattformen führen, neben Spam für Werbezwecke (Kap. 4.1) vor allem in Form von gezielt verbreiteten Desinformationen oder manipulativen Inhalten (Ananthaswamy 2023, S. 203; Atleson 2023). Die von ChatGPT verfassten Beiträge lassen sich kaum von menschlich erstellten unterscheiden, wie in einem – ethisch allerdings fragwürdigen, weil nicht transparenten – Experiment mit Facebook-Nutzenden festgestellt wurde (Schinkels 2023). Noch mehr dürfte dies für GPT-4 gelten, mit dem sich noch überzeugendere Texte erstellen lassen als mit ChatGPT. Die Kosten von Desinformationskampagnen verringern sich durch die KI-Systeme, gleichzeitig wird die Effektivität und Dynamik der Beiträge größer (Goldstein et al. 2023, S. 22ff.). Obwohl die Problematik in Bezug auf sprachverarbeitende KI-Modelle bereits seit langem diskutiert wird (Buchanan et al. 2021), werden die angedachten Gegenmaßnahmen – unter anderem die technische Erkennung von Erzeugnissen der KI-Systeme, staatliche Regulierung, Überwachung der Nutzenden von Social-Media-Plattformen und Medienkompetenzschulungen – als nicht ausreichend bewertet (Hsu/Thompson 2023).

---

42 Die zunehmende Bedeutung von Intermediären bei der Vermittlung von Nachrichten schwächt die Medienunternehmen im Wettbewerb um Werbeeinnahmen (TAB 2022a, S. 83). Im Streit um eine angemessene Aufteilung haben deutsche Zeitungsverleger, die sich auf das Leistungsschutzrecht berufen, und Google 2022 die Schiedsstelle beim Deutschen Patent- und Markenamt angerufen (Hanfeld 2022). Falls zukünftig in nennenswertem Umfang für die Informationssuche Chatbots anstelle von Suchmaschinen befragt werden (Kap. 4.3.2), sind davon sowohl das bisherige Geschäftsmodell der Suchmaschinenbetreibenden als auch die Vergütung der Urheber bzw. Verleger bedroht. Presseverleger haben in Hinblick auf eine eventuelle Nutzung von Presseinhalten durch Chatbots bereits Ansprüche auf eine Vergütung angemeldet (Voß 2023).

Mit einer möglichen Verbreitung von (nicht als solchen erkennbaren) KI-generierten Beiträgen in der Onlinekommunikation wird die Sorge verbunden, dass auf diese Weise *Verunsicherung* über die Zurechenbarkeit der Äußerungen und ein *Vertrauensverlust* enstehen kann: »Sowohl das Vertrauen in den Diskurs und die mögliche Einigung auf Kompromisse als auch das Vertrauen in demokratische Prozesse insgesamt, die eigene Urteilskraft sowie die individuellen Wirkmöglichkeiten können auf diese Weise vermindert werden.« (Deutscher Ethikrat 2023, S. 216, ähnlich Haven 2022) Dieser Effekt kann selbst dann eintreten, wenn automatisiert erzeugte Beiträge nicht böswillig oder manipulativ verwendet werden. Denn in jedem Fall verwischt durch den fehlenden Faktenbezug von ChatGPT und verwandten Systemen die Grenze zwischen Wahrheit und Unwahrheit. War die Beurteilung von Texten bisher anhand ihrer formalen Erscheinung möglich, so muss nun der Inhalt bewertet werden (Brühl 2023), was bei einer großen Menge an durch KI-Systeme erzeugten Beiträgen große Herausforderungen mit sich bringen dürfte.

Anhand der skizzierten Anwendungspotenziale und Implikationen von KI-Modellen zur Sprachverarbeitung im Journalismus sowie in der öffentlichen Kommunikation läßt sich eine Reihe von weiteren *offenen Fragen und Überlegungen* ableiten:

- Wie lässt sich die Einhaltung von Kennzeichungspflichten und journalistischen Standards, wie sie etwa im Pressekodex und dem Medienstaatsvertrag vereinbart sind, bei der automatisierten journalistischen Textproduktion gewährleisten?
- Wie ist ChatGPT als Onlineangebot medienrechtlich zu bewerten? Einerseits richtet sich das System nur an individuelle Nutzende, andererseits erreicht OpenAI mit dem Angebot ein Millionenpublikum, verbunden mit der Sammlung hoch sensibler persönlicher Daten (Kap. 6.1). Es stellt sich die Frage, ob der für Telemedien geltende Rahmen der verfassungsmäßigen Ordnung und der allgemeinen Gesetze ausreicht, oder ob es angesichts der großen Zahl von Nutzenden einer weitergehenden speziellen, einem möglichen Einfluss auf die Meinungsbildung (Hartmann et al. 2023; Rozado 2023) oder auf Kaufentscheidungen[43] gerecht werdenden Regulierung bedarf.
- Auswirkungen auf die Arbeit werden vor allem im Bereich der Medienproduktion erwartet, die zum Teil automatisiert werden kann. Kostenvorteilen aufgrund von Effizienzgewinnen (TAB 2022a, S. 79f.) stehen hohe Investitionskosten entgegen (Puppis et al. 2017, S. 357). Die journalistische Arbeit

43 ChatGPT wird offenbar bereits zur Reiseplanung genutzt und empfiehlt dabei einzelne Hotels (Towey 2023). Das System könnte nicht nur als Kaufberatung eingesetzt werden (Mickle et al. 2023), sondern über Plugins auch Bestellungen auslösen (McCormick 2023). Eine wettbewerbsrechtliche Befassung mit solchen Anwendungen steht offenbar noch aus (Kim/Stenert 2023).

wiederum gilt als zu vielfältig, als dass sich Menschen ohne weiteres durch letztlich spezialisierte Maschinen ersetzen ließen (Burrell 2023; Dax 2023), zudem wurden viele Redaktionen bereits infolge der Digitalisierung weitestmöglich verkleinert bzw. umgestaltet (Puppis et al. 2017, S. 277).

- Das Risiko einer Verbreitung von fehlerhaften, durch KI-Systeme erzeugten Informationen (Kap. 4.2), etwa in Finanz- oder Gesundheitsfragen, wird durch die große Reichweite automatisierter Texte in Onlinemedien noch vergrößert.

4.3.2 Szenario: Neue Praktiken der Informationssuche

Die immer größer werdende Menge an digital über das Internet verfügbaren Texten bildet nicht nur die Grundlage der heutigen sprachverarbeitenden KI-Systeme, sie gab auch den Anlass zur Entwicklung von Suchmaschinen, mit denen sich diese *Wissenssammlung erschließen* lässt. Eine Suchmaschine kombiniert dabei klassischerweise drei Aufgaben: die Durchforstung des World Wide Web nach Informationen (engl. crawling), die Speicherung eines Abbilds dieser Daten (engl. indexing) und die eigentliche Suchfunktion anhand dieses Abbilds (engl. searching) (Seymour et al. 2011). Die Nutzenden geben dabei in die Suchmaske einen oder mehrere Begriffe ein, die Suchmaschine gibt eine Liste von Links (mit kurzen Beschreibungen) aus, die auf die Webseiten verweisen, auf denen sich die gesuchte Information schließlich (vermeintlich) finden lässt.

2021 haben Mitarbeitende von Google in mehreren Publikationen eine neue Vorgehensweise der Informationssuche skizziert, die auf einem Dialogsystem beruht, das mit einem KI-System zur Sprachverarbeitung gepaart wird (Shah/Bender 2021, S. 222). Gemäß dieser Vorstellung ist kein Index von Webseiten mehr nötig, vielmehr sind die *Informationen in einem KI-Modell* abgebildet, das dialogisch befragt werden kann und das die maßgeblich relevanten Informationen als Antwort ausgibt (Metzler et al. 2021). Das System sollte außerdem transparent bezüglich seiner Quellen und nicht voreingenommen sein; es sollte Informationen aus unterschiedlichen Blickwinkeln darstellen – und das alles in verständlicher Sprache, multimodal und mehrsprachig.

Mit einer Suchmaschine, die auf einem KI-gestützten Chatbot beruht, wird aus dem Versprechen, »Informationen auf Tastendruck« zu liefern, die Vision, stets eine/n Expert/in als eine Art »Telefonjoker« zur Hand zu haben, die oder der nach Belieben befragt werden kann. Die Ähnlichkeit zu ChatGPT bzw. genauer zu den bei Bing und Google eingebundenen Chatbots ist unschwer zu erkennen. Auch die Unzulänglichkeiten der bisherigen Umsetzungen gegenüber der Vision (z. B. die fehlende Quellennennung) fallen ins Auge. Die Implikationen der Verwendung von ChatGPT und verwandten Systemen für die Informationssuche lassen sich aus der Perspektive der Nutzenden, der

Suchmaschinenanbietenden, der Inhaltslieferanten sowie aus gesellschaftlicher Perspektive betrachten.

- *Informationssuchende* können ganz unterschiedliche Ziele mit ihrer Suche verfolgen, vom einfachen Nachschlagen von Informationen bis hin zu komplexeren Prozessen der Wissenskonstruktion (Shah/Bender 2021, S. 224ff.). Erfahrungsgemäß stoßen jedoch häufig nur die ersten Plätze auf der Trefferliste zu einer Anfrage auf Interesse (Kullmann 2023). In vielen Fällen dürfte daher eine ausformulierte Antwort als Ergebnis hilfreich sein, sofern es nicht um Fachfragen zu professionellen Themen geht. Da sich die Ergebnisse nicht ohne weiteres überprüfen lassen, stellen sich für Informationssuchende die bereits diskutierten Probleme als gravierend dar: der fehlende Faktenbezug (Kap. 3.2.2), die Verschleierung von Unwissenheit (Kap. 3.2.5), das bislang ungelöste Problem, Quellen zu den Antworten anzugeben, sowie die mangelnde Aktualität und Unvoreingenommenheit in Bezug auf das zugrundeliegende Datenmaterial (Kap. 3.2.3 sowie 4.3.1, FN 40).
- *Den Betreiber/innen von Suchmaschinen* bietet sich die Chance, die Suche für die Nutzenden effizienter und attraktiver zu gestalten. Insbesondere Microsoft scheint sich von der neuen Funktionalität mehr Interesse für die Suchmaschine Bing zu versprechen, auf die bisher lediglich 3 % des Marktes für Suchmaschinen (gegenüber ca. 90 % für Google) entfallen (Kahn 2023). Ein KI-Modell zur Sprachverarbeitung muss dabei nicht unbedingt als Ersatz für die Suche eingesetzt werden. Google verwendet seit 2019 sein Modell BERT, um die Suchanfragen der Nutzenden zu interpretieren und so genauere Suchprozesse auslösen zu können (Nayak 2019). Auch der Chatbot Bard soll die Suche nur ergänzen, nicht ersetzen (Heaven 2023a).
  Die neue Technologie hat jedoch auch riskante Implikationen für Suchmaschinenbetreiner/innen, insbesondere die hohen Kosten für den Betrieb der KI-Modelle. Bereits die Kosten für das (Vor-)Training eines Modells sind sehr hoch, es handelt sich aber um eine einmalige Investition. Die Betriebskosten dagegen sind zwar pro Interaktion gering, summieren sich aber bei einer massenhaften Nutzung, wie bei Suchmaschinen zu erwarten, auf sehr hohe Summen (Patel/Ahmad 2023). Insbesondere sind sie im Vergleich zu den Kosten einer traditionellen Suchanfrage deutlich höher (bis zu einem Faktor von 10, Dastin/Nellis 2023), so dass die Gewinne der Betreiberunternehmen schrumpfen dürften. Es ist daher zu erwarten, dass – sofern sich die Kosten für den Betrieb nicht durch technologische Innovationen senken lassen (Kap. 3.3.3) – ein Einsatz der Chatfunktion entweder nicht für alle Arten von Suchanfragen oder nicht für alle Kund/innen zum Einsatz kommt.

  Zwei weitere Fragen stellen sich in Bezug auf die Informationssammlung, die die Basis der Suchmaschinen bilden. Zum einen ist rechtlich nicht klar, inwiefern Informationssammlungen professioneller Anbieter für das

Training von KI-Modellen genutzt werden dürfen (Kullmann 2023; Kap. 6.2). Zum anderen dürften die Sammlungen mit zunehmender Nutzung von KI-Modellen zur Texterzeugung einen weiter wachsenden Anteil von Texten enthalten, die nicht mehr eine allein menschliche Selektionslogik widerspiegeln. Google zeigt sich zwar offen für KI-generierte Inhalte (Google Search Central 2023), allerdings ist nicht klar, welche Auswirkungen dies beispielsweise auf den Rankingalgorithmus haben könnte.[44]

- Für *Anbieter/innen von Internetseiten* erfüllen Suchmaschinen die wichtige Funktion, Nutzende auf ihre Inhalte zu lenken. Wenn Informationssuchende nicht mehr auf die Seiten selbst gelenkt werden, sondern zuvor durch die Antworten eines Chatbots befriedigt werden, verlieren die Inhaltsanbieter eine wichtige Einnahmequelle.
- *Gesellschaftliche Auswirkungen* ergeben sich nach Shah und Bender (2021) insofern, als eine Suche mithilfe von Chatbots keineswegs die vielfältigen Funktionen erfüllen kann, die die Suche nach Informationen im Internet haben kann. Beispielsweise geht beim Einsatz von KI-Modellen der weitere Kontext der Informationen verloren, der für eine kritische Bewertung der Informationen wichtig ist. Auch die Recherche- und Medienkompetenzen der Nutzenden werden nicht im gleichen Maß geschult wie bei einer klassischen Suche (Fulterer 2023).

Die Internetsuche und das mit ihr verbundene Werbegeschäft bedeuten allein bei Google einen Quartalsumsatz von fast 40 Mrd. US-Dollar (Scheuer 2022). Dieser Markt ist stark umkämpft, entsprechend aufgeregt sollen sowohl Google als auch Meta, das Mutterunternehmen von Facebook, auf die Veröffentlichung von ChatGPT reagiert haben (Tiku et al. 2023). Der resultierende Wettbewerb um eine möglichst schnelle Einführung neuer Systeme, Funktionen oder Versionen der KI-Modelle könnte jedoch *zulasten einer verantwortungsvollen Entwicklung und sorgfältigen Prüfung der Systeme* gehen (Alouani 2023).

OpenAI hat zwar nach eigener Ansicht umfangreiche Tests, zuletzt vor Veröffentlichung von GPT-4, durchgeführt (OpenAI 2023e). Kritisiert wird jedoch, dass dadurch nur der Form halber Regeln eingehalten wurden und die Überprüfung vor allem intern erfolgte (Crawford/Calo 2016, S. 312; Tiku et al. 2023). Auch die für die Tests verwendeten Benchmarks, automatisierte Tests bestimmter Fähigkeiten der Systeme, werden als nicht unbedingt aussagekräftig kritisiert. Es wird vermutet, dass Teile der in den Tests genutzten Material bereits im Trainingsdatensatz enthalten waren (Marcus/Davis 2023a; Narayanan/-

44 Eine skurrile Schleife der Kommunikation von KI-Systemen untereinander scheint es beispielsweise im Fall von Bard und Bing gegeben zu haben: Googles Bard äußerte in einem Chat mit einem Nutzer, dass Bard vom Netz genommen worden sei. Über diese offensichtlich unsinnige Selbstauskunft berichteten mehrere Onlinemedien, was wiederum von Microsofts Chatbot bei Bing aufgegriffen wurde, der diese (falsche) Information an seine Nutzenden berichtete (Felton 2023).

Kapoor 2023). Nicht zuletzt die offenkundigen Schwächen und die Gefahren, die von einer Anwendung der KI-Systeme von OpenAI ausgehen können, zeigen, dass eine verantwortungsvolle Entwicklung wichtig ist. Erste Schritte zum Schutz der Verbraucher/innen haben mittlerweile die US-amerikanische Handels- und Verbraucherschutzbehörde FTC sowie die italienische Datenschutzbehörde unternommen (Hahn 2023c).

4.4 Anwendungen im Rechtswesen und der öffentlichen Verwaltung

Anwendungen von Systemen der Künstlichen Intelligenz spielen bereits seit längerem eine Rolle in der *öffentlichen Verwaltung* (TAB 2022c) und auch im *Rechtswesen* (TAB 2019b, S. 54ff.). Mehrere Bundesbehörden erproben ganz unterschiedliche KI-Systeme (Bundesregierung 2022). Auch ChatGPT wurde bereits getestet, findet aber noch keine Verwendung auf Webseiten der Bundesregierung oder nachgeordneter Behörden (Bundesregierung 2023a). Beim Informationstechnikzentrum Bund (ITZBund) wird eine Reihe von *Chatbots* entwickelt, die bereits im Einsatz sind.[45] Sowohl im Rechtswesen als auch in der öffentlichen Verwaltung werden Entscheidungen von großer Tragweite und unmittelbaren Auswirkungen für die betroffenen Bürger/innen getroffen, so dass die Anforderungen an die dort eingesetzten Technologien besonders hoch sind (Deutscher Ethikrat 2023).

4.4.1 Mögliche Anwendungen im Rechtswesen

Die Diskussion über mögliche Anwendungen im Rechtswesen teilt sich in einen *medial stark sichtbaren*, hinsichtlich der Umsetzung allerdings fragwürdigen und einen vergleichsweise ruhigeren, aber mit *realistischeren Erwartungen* verknüpften Teil. Dieser Eindruck ergibt sich, wenn man die Ergebnisse der TAB-Untersuchung von Entwicklungen im Bereich von Legal Tech (TAB 2019b) mit den in einigen Medien diskutierten Einschätzungen zur Leistungsfähigkeit und zum Stand der Technologie miteinander vergleicht.

Auf der einen Seite machte insbesondere das Unternehmen DoNotPay Schlagzeilen, das eine Form der *automatisierten Rechtsberatung* über das Internet anbietet. Im Februar 2023 wurde verkündet, ein KI-System werde in einem Verfahren vor einem US-amerikanischen Gericht anstelle eines Anwalts agieren. Dafür wollte man sprachverarbeitende Systems wie ChatGPT nutzen; über Brillen mit einer eingebauten Funkverbindung sollte der oder die Mandant/in mit dem System verbunden sein, dessen Anweisungen folgen und vorfor-

45 www.itzbund.de/DE/itloesungen/standardloesungen/chatbots/chatbots.html (19.4.2023).

mulierte Äußerungen nachsprechen (Allyn 2023). Das Experiment wurde allerdings nie durchgeführt, angeblich, weil dem Unternehmen mit rechtlichen Konsequenzen wegen eines Verstoßes gegen rechtspraktische Vorschriften gedroht wurde. Ebenfalls große mediale Aufmerksamkeit, wenn auch offensichtlich auf einer realen Grundlage, weckte die Nachricht, ein Richter in Kolumbien habe sich bei der Formulierung eines Urteils von ChatGPT inspirieren lassen (Gutiérrez 2023).

Auf der anderen Seite gibt es durchaus *ernsthafte Auseinandersetzungen* mit den Möglichkeiten und Grenzen von ChatGPT in dem Versuch, geeignete Anwendungsmöglichkeiten für das System im Rechtswesen zu identifizieren (Volland 2023). Zu solchen gehören: Zusammenfassungen und Paraphrasierungen juristischer Dokumente, die Formulierung von Vertragsklauseln oder die Analyse von Schriftstücken. Einer tatsächlichen Verwendung stehen demnach allerdings unter anderem der Datenschutz und die anwaltliche Berufshaftung entgegen. In den USA gibt es jedoch bereits mindestens ein Unternehmen, das mithilfe von GPT-3 entsprechende Unterstützungdienste für Anwält/innen anbieten (Hutson 2021).

4.4.2 Anwendungsszenario: Chatbot in der öffentlichen Verwaltung

Für einen Überblick über die Rolle von KI-Systemen in der öffentlichen Veraltung sei auf die Untersuchung des TAB zum Thema »Künstliche Intelligenz und Distributed-Ledger-Technologie in der öffentlichen Verwaltung« verwiesen (TAB 2022c). ChatGPT wird bei mehreren Bundesbehörden erprobt und wurde einmal – ohne dies kenntlich zu machen – beim Bundesministerium für Bildung und Forschung zur Beantwortung einer schriftlichen Frage einer Abgeordneten genutzt (Deutscher Bundestag 2023a, S. 87). Im Folgenden steht speziell das Anwendungsszenario eines *Bürgerservices mittels Chatbot* im Vordergrund.

Chatbots für die Verwaltung dienen aus Sicht der Behörden dazu, »Verwaltungsprozessen einen individualisierten Zuschnitt zu geben und im Zuge dessen die Servicequalität der Verwaltungsleistungen zu erhöhen« (TAB 2022c, S. 32). Sie werden für Auskünfte oder zur Beantwortung von Bürgeranfragen auf allen Ebenen der Verwaltung eingesetzt. Bürger/innen erhoffen sich von sprachmächtigen Chatbots wie ChatGPT, dass diese dabei helfen, *komplizierte Behördenvorgänge* »in *einfacher Alltagssprache* für alle verständlich« zu machen (Kastl 2023).

Beispiele für *Chatbots der Verwaltung* sind ViOlA vom Bundeszentralamt für Steuern (entwickelt vom ITZBund) sowie LUMI, der Bürgerassistent der Stadt Heidelberg (entwickelt auf Basis des Modells Luminous von dem deutschen Unternehmen Aleph Alpha, Kap. 2.1). Mit dem Chatbot C-19 für regionenspezifische Informationen zu Regelungen in der Covid-19-Pandemie wurde

eine Blaupause entwickelt, die auch von anderen Behörden genutzt werden können soll (Engelmann/Puntschuh 2020, S. 22). In Rumänien wurde ein Chatbot durch den Ministerpräsidenten des Landes als »Berater« vorgestellt, das System soll soziale Medien auswerten und den Ministerpräsidenten in Echtzeit über Meinungen und Wünsche der Bürger/innen informieren (Caparella 2023). Auch OpenAI griff auf das Szenario der Unterstützung von Behördenkommunikation zurück und ließ GPT-4 bei der Produktvorstellung eine Steuererklärung berechnen, bei der das KI-System die Regelungen auswertete und auf einen gegebenen Fall bezog.

Ein besonderer Aspekt ist der Bezug zur *inklusiven Kommunikation*. Öffentliche Stellen sind verpflichtet, auf ihren Webseiten Angebote auch in Leichter Sprache bereitzustellen. Die Übersetzung aus Alltags- oder Behördensprache in Leichte Sprache ist jedoch aufwendig, daher ist meist nur ein Mindestmaß verfügbar. Digitale Systeme wie Capito digital oder SUMM, die ebenfalls auf KI-Systemen basieren, unterstützen bei der Übersetzung. Auch ChatGPT soll dank der Fähigkeit, unterschiedliche Sprachstile nachzuahmen, in der Lage sein, Texte in ein leichteres sprachliches Niveau zu verwandeln (Manning 2023). *Leichte Sprache* im Sinn des zugehörigen *Regelwerks* beherrscht ChatGPT als im Kern probabilistisches Modell jedoch nicht ohne weiteres. Auch besteht die Gefahr, dass durch die Nutzung der Technologie das eigentliche Problem einer Umstellung von Kommunikationsprozessen hin zu mehr Inklusion vermieden wird (Kreer 2023).

4.4.3 Risiken: Mangelnde Verlässlichkeit der Systeme, Verstärkung von Bias bzw. Diskriminierung

Anwendungen in der öffentlichen Verwaltung machen eine noch wenig diskutierte Herausforderung deutlich: die Frage der *Zuverlässigkeit und Resilienz* von KI-basierten Angeboten. Die Anwendungen müssen auch bei *hoher Nutzungslast* verfügbar und funktional sein (wie im Bildungskontext zuletzt in der Zeit der Covid-19-bedingten Schulschließungen spürbar wurde). Je nach Einsatzgebiet muss die Verfügbarkeit *langfristig* gesichert sein, etwa wenn Lebensverläufe über einen längeren Zeitraum hinweg begleitet werden sollen. Und auch die Resilienz eines KI-Systems im Sinne eines sicheren, genauen, *zuverlässigen, robusten* und verständlichen Angebots (Wittenbrink et al. 2023) muss gewährleistet sein, wie am Beispiel von inklusionsorientierten Diensten deutlich wird – die Betroffenen können sich in der Regel nicht aussuchen, wann sie die Unterstützung des Systems benötigen, daher kann auch schon eine kurzzeitige Nichtverfügbarkeit je nach Anwendungskontext gravierende Folgen haben. Ein Forschungsprototyp wie ChatGPT erfüllt diese Anforderungen vermutlich nicht.

Ein weiterer Aspekt der Verlässlichkeit, der bei solchen Anwendungen in der öffentlichen Verwaltung eine Rolle spielt, bei denen Entscheidungen mit Auswirkungen auf die Bürger/innen getroffen werden, ist die *Verständlichkeit* bzw. *Nachvollziehbarkeit* (Rudin 2019). Dazu zählt etwa, dass auf den gleichen Input die gleiche Reaktion erfolgt – was bei KI-Modellen zur Sprachverarbeitung jedoch nicht der Fall ist. Diese Systeme gelten als Blackboxen (Kap. 3.2.4), die durch ihre Komplexität und auch aufgrund unternehmerischer Geheimhaltung auf nicht nachvollziehbare Weise zu ihren Ergebnissen kommen. Auch wenn aus der Perspektive der Nutzenden viele Technologien, mit denen sie im Alltag in Berührung kommen, Blackboxen sind (z. B. Autos), so ist bei diesen durch entsprechende Regulierung und kompetente Spezialist/innen doch sichergestellt, dass sie zuverlässig funktionieren. Für einen Chatbot, der eine Steuererklärung erstellen und möglicherweise auch prüfen kann, sollte dies ebenfalls gelten.

Eine zweite Gefahr negativer Auswirkungen der Nutzung von KI-Modellen zur Sprachverarbeitung im Rechtswesen und der öffentlichen Verwaltung ist ein bei diesen Systemen häufig beobachteter *Bias*, also eine *verzerrte Repräsentation bestimmter Kategorien* (Kap. 3.3.2). Besonders in Hinblick auf die sechs im Allgemeinen Gleichstellungsgesetz genannten soziodemographischen Kategorien ist eine solche Verzerrung problematisch: Geschlecht, rassifizierende Zuschreibung (zum Begriff: Baumann et al. 2018, S. 19), Religion bzw. Weltanschauung, sexuelle Identität/Orientierung, Behinderung bzw. chronische Krankheit und Alter. Auch weitere Kategorien können betroffen sein, insbesondere zu den ersten vier genannten Kategorien liegen jedoch bereits Studien zu spracherzeugenden KI-Systemen vor (Sheng et al. 2021, S. 4276). Bei *GPT-3* wurde Bias hinsichtlich der Kategorien rassifizierende Zuschreibungen, Gender und Religion identifiziert (Abid et al. 2021; Tamkin et al. 2021, S. 6). Auch bei *ChatGPT* wurde Bias hinsichtlich dieser Kategorien erkannt (Alba 2022; Biddle 2022), es wurde jedoch auch beoabchtet, dass spätere Versionen weniger anfällig dafür waren (Borji 2023, S. 13).

Auswirkungen kann Bias zum einen hinsichtlich der (negativen oder auch fehlenden) *Repräsentation* der entsprechenden gesellschaftlichen Gruppen in den Ausgaben der KI-Systeme haben: Stereotype können verbreitet und verstärkt werden (Dev et al. 2022, S. 247), die Gruppen könnten fälschlich oder verunglimpfend dargestellt und so die gesellschaftliche Wahrnehmung der Gruppen negativ beeinflusst werden (Sheng et al. 2021, S. 4278), ein Ausschluss der Gruppe aus gemeinsamen Wertbezügen kann suggeriert oder es können Ansprüche auf gleichberechtigte Teilhabe oder Rücksichtnahme verwehrt werden.

Auswirkungen auf die *Zuteilung von Ressourcen* können entstehen, wenn aufgrund der Ergebnisse eines sprachverarbeitenden KI-Systems Menschen ungleich behandelt werden. So ist der Fall eines Palästinensers bekannt, der 2017 aufgrund einer fehlerhaften automatischen Übersetzung eines Facebookposts

von der israelischen Polizei verhaftet worden war. Sein Post mit dem Gruß »Guten Morgen« in arabischer Sprache war als »Greift sie an« ins Hebräische übersetzt worden – nach der Aufklärung des Fehlers wurde er freigelassen (Hern 2017).[46] Außerdem kann Bias in einem System dazu führen, dass die davon betroffenen Gruppen das System nicht mehr oder nicht mehr so effizient benutzen und dadurch in ihrer Nutzung dieser Ressource benachteiligt werden, dabei kann auch ein selbstverstärkender Effekt auftreten, wenn durch eine geringere Nutzung das Feedback von entsprechenden Gruppen nicht oder in geringerem Maße als von anderen Gruppen in die Weiterentwicklung des Systems einfließt (Sheng et al. 2021, S. 4278, Hashimoto et al. 2018).

Ein dritter Effekt betrifft schließlich *schutzbedürftige Gruppen*, deren Verletzbarkeit durch die Anfälligkeit sprachlicher KI-Systeme etwa für den Missbrauch personenbezogener Daten oder sonstige missbräuchliche Verwendung in besonderer Weise verstärkt wird (Sheng et al. 2021, S. 4279).

Woher kommt Bias? Zum einen aus den *Trainingsdaten* – Ferrer et al. (2021) konnten nachweisen, dass bestimmte Reddit-Gruppen in Hinblick auf Geschlecht, Religion und ethnische Zuschreibung voreingenommen sind. Websiteempfehlungen von Redditnutzern wiederum bildeten ein Auswahlkriterium für Trainingsmaterial, das in der Entwicklung von zumindest GPT-2 und GPT-3 verwendet wurde (Radford et al. 2019; Brown et al. 2020). Bolukbasi et al. (2016) wiesen nach, dass ein Bias in den Trainingsdaten Eingang in die Modelle findet, wenn eine bei sprachbezogenem maschinellem Lernen weitverbreitete Technik (word embeddings) verwendet wird.

Nur durch *zusätzliche Maßnahmen* beim Design der KI-Systeme ist es möglich, beispielsweise abwertende berufliche Geschlechterstereotype gegenüber Frauen innerhalb der Modelle und ihres Outputs zu vermeiden. Allerdings lassen sich Bias und Rassismus auf Ebene der Trainingsdaten kaum herausfiltern, weil das zu aufwendig wäre und weil selbst bei vielen Ressourcen und gutem Willen die Erkennung bei impliziten Äußerungen nicht einfach ist.

Auch die bei der Auswahl von Trainingsmaterial angewendeten *Filter* können Voreingenommenheiten Vorschub leisten, etwa wenn marginalisierte Sprechweisen als unerwünscht herausgefiltert werden (Bender et al. 2021; Sheng et al. 2021, S. 4279). Außerdem gibt es Hinweise, dass bestimmte *Modellarchitekturen* der künstlichen neuronalen Netze sowie ihrer Funktionsweise bei der Textgenerierung (Inference) Bias begünstigen können; die Forschung dazu (wie das Verständnis der Modellarchitekturen insgesamt) ist allerdings noch in einem frühen Stadium (Sheng et al. 2021, S. 4279f.). Schließlich werden zur Evaluierung der Modelle standardisierte Tests eingesetzt, die ihrerseits zu Bias beitragen können, etwa wenn sie aufgrund von Vergleichsdaten, die durch

46 An diesem Beispiel wird deutlich, dass Bias-Probleme auch mit fehlerhaften Ausgaben verbunden sein können – der Fehler in diesem Fall kam zustande, weil Arabisch im KI-System schlecht zu repräsentieren ist.

Standardsprache dominiert sind, solche Modelle schlechter bewerten, in denen marginalisierte Sprachen repräsentiert sind.

Schließlich ist Bias auch *mit Sprachen konnotiert*. Englische Nutzende sind insofern bessergestellt, als das System in englischer Sprache deutlich besser funktioniert als in anderen Sprachen. Innerhalb des Englischen kann ChatGPT zu einer Dominanz von (von Weißen verwendetem) Standardenglisch zuungunsten von Dialekten oder Slangs führen. Kleine, marginalsierte Sprachen verlieren möglicherweise weiter an Attraktivität (Bjork 2023).[47] Und auch wenn fremde Sprachen recht gut repräsentiert sind, bleibt das System aufgrund seiner Trainigsdaten und seinen Sicherheitsvorkehrungen doch dem US-amerikanischen Kultur- und Rechtskontext verbunden (Rettberg 2022).

47 Bei der Entwicklung von GPT-4 hat die Regierung von Island offenbar mit OpenAI zusammengearbeitet, um die isländische(n) Sprache(n) zu bewahren (https://openai.com/customer-stories/government-of-iceland, 19.4.2023).

5 Auswirkungen von ChatGPT in Bildung und Forschung

Die Vielfalt der Anwendungsmöglichkeiten sprachverarbeitender KI-Modelle zeigt, dass diese Technologie sehr grundlegend auf die menschliche Wissensarbeit einwirken und so auch gesellschaftliche Strukturen verändern kann (Ovadya 2021). Die Bereiche Bildung und Forschung sind davon besonders betroffen, entsprechend intensiv wird über die Rolle von ChatGPT in diesen Bereichen diskutiert. Vom »Ende der Hausarbeiten« (Marche 2022) ist die Rede, von einem »Gamechanger« im Bildungsbereich (van Deyzen 2023; Roth 2023), aber auch von neuen Chancen für Studierende (Stock 2023). Informations- und Fortbildungsveranstaltungen für Lehrende verzeichnen Rekorde bei den Teilnehmendenzahlen. Gezeigt wurde zudem, dass ChatGPT standardisierte Prüfungen in Fächern wie Medizin (Kung et al. 2023), Rechtswissenschaft (Choi et al. 2023) und Informatik (Finnie-Ansley et al. 2022) bestehen kann.[48]

Dennoch ist diese Aufmerksamkeit im Bildungsbereich erstaunlich. Die bekannten Grenzen und Risiken der KI-Modelle zur Sprachverarbeitung, etwa das Fabulieren und die fehlende Verlässlichkeit, aber auch die Intransparenz bezüglich möglicher Voreingenommenheit (Kap. 3.2; Mohr et al. 2023), lassen eine vorschnelle Anwendung in einem so wichtigen und sensiblen Bereich der Gesellschaft nicht naheliegend erscheinen.[49] Ein Entwurf der Europäischen Kommission (2021, S. 31) für ein Gesetz über Künstliche Intelligenz sieht vor, KI-Anwendungen im Bildungsbereich als hochriskant anzusehen. Dies hätte eine strenge Regulierung entsprechender Anwendungen zur Folge. Allerdings wird über das Gesetz noch verhandelt, weshalb derzeit ein genauerer Blick auf die Anwendungsmöglichkeiten von sprachverarbeitenden KI-Modellen und die damit verbundenen Implikationen nötig ist.

Die folgende Darstellung ist anhand der betroffenen Akteure (Lernende, Lehrende und Institutionen) gegliedert, für die jeweils Potenziale und Risiken diskutiert werden (Kap. 5.1). Der sekundäre und tertiäre Bildungsbereich werden gemeinsam behandelt, zur Primarstufe beschränkt sich die Diskussion in Deutschland bisher auf einzelne interessierte Lehrkräfte, daher wird sie nicht

48 Bei den Aufgaben des bayerischen Abiturs reichten die Fähigkeiten allerdings nicht zum Bestehen (Gawlik/Schiffer 2023), auch eine probeweise mit beinahe ausschließlicher Hilfe von ChatGPT verfasste Bachelorarbeit fiel wegen »gravierender Mängel« durch (Ciesielski/Barthel 2023).

49 In einem Test, bei dem zwei Mal mit ein und derselben Frage nach Literatur zum Thema Inflation gefragt wurde, antwortete das System mit ganz unterschiedlichen Empfehlungen (Vogelgesang et al. 2023, S. 5f.). Zumindest eines der empfohlenen Werke existiert dabei gar nicht. Auch Sam Altman, CEO von OpenAI, schrieb über ChatGPT, es sei ein Fehler, sich darauf zu verlassen, wenn es um wichtige Dinge geht (Altman 2022).

berücksichtigt.[50] Auch die berufliche Aus- und Weiterbildung ist bislang in der Debatte unterrepräsentiert.[51] Für den Forschungsbereich werden mögliche positive und negative Auswirkungen einer Nutzung sprachverarbeitender KI-Modelle dargestellt (Kap. 5.2).

5.1 Chancen und Risiken im Bereich Bildung

5.1.1 Perspektive der Lernenden

Schüler/innen wie Studierende haben ChatGPT nach der Veröffentlichung schnell aufgegriffen und damit experimentiert, auch im Rahmen von Lernprozessen. Die Interaktion mit dem System wird von erwachsenen Lernenden als persönlich und bereichernd beschrieben (Klinge 2022). *Erlebnisberichten* von Schüler/innen in einem Pressebericht zufolge wird das System genutzt,

- um den Lernstoff besser zu verstehen (»Wenn mir die Antwort nicht ausreicht, schreibe ich: ›Erkläre das genauer‹, oder ich stelle meine Frage anders. Schon bekomme ich einen neuen, detaillierteren Text, der mir beim Lernen hilft«),
- um Routineaufgaben zu erledigen (»Die KI verkürzt die Zeit für die Fleißarbeit, damit ich mehr Zeit für die Denkarbeit habe«), oder
- um sich zeitaufwendiger Arbeit zu entledigen (»Zuerst habe ich die Englischaufgaben damit gemacht (…) Frage eingeben und die Antwort kopieren. Das geht total einfach, wenn man mal faul ist oder keine Zeit hat.«).[52]

Abgesehen vom automatisierten *Erstellen von Texten* kann das Programm als *Werkzeug bei der Textbearbeitung* genutzt werden, also zum Paraphrasieren, zum teilweise auch stilistischen Korrigieren (Mohr et al. 2023), zum Übersetzen, zur Suche nach Synonymen etc. (Marx 2023). Es kann auch bei der Strukturierung von Themen helfen, Ideen für das eigene Schreiben geben,[53] Aufgaben und Fragen für das Selbstlernen generieren sowie Musterlösungen dazu anbieten (Blume 2023). Für Lernende, die gut mit den Prompts umgehen können, kann das System auch in der Basisversion als persönlicher, interaktiver

50 Mündliche Auskunft, Teilnehmende einer Onlinekonferenz zu künstlicher Intelligenz in der Bildung am 2.2.2023.

51 Eine umfassendere Untersuchung zum Thema »Anwendungspotenziale und Herausforderungen von künstlicher Intelligenz in der Bildung« wird derzeit durch das TAB bearbeitet (www.tab-beim-bundestag.de/projekte_anwendungspotenziale-und-herausforderungen-von-kuenstlicher-intelligenz-in-der-bildung.php, 19.4.2023).

52 Alle Zitate stammen aus Brandstätter (2023). Es ist zu beachten, dass es sich bei entsprechenden Berichten um anekdotische Evidenz handelt, die ggf. aus dem Kontext gerissen und zudem nach medialen Selektionskriterien ausgewählt wurde.

53 Christian Spannagel, mündliche Kommunikation, Webinar des Multimediakontors Hamburg am 14.3.2023.

Lerncoach fungieren (Klinge 2022) bzw. Auskünfte geben, wenn sie Hilfe benötigen (Heidt 2023). Bei der »selbstständigen Bearbeitung von Aufgabenstellungen bzw. der Aneignung von neuen Stoffgebieten« kann ChatGPT, »richtig befragt und genutzt, unzählige Anregungen für Gliederung, Ideen, Konzepte und Reflexion liefern«.[54]

Zur *Unterstützung des Selbstlernens* werden auch spezialisierte Anwendungen auf Basis von ChatGPT entwickelt. Ein Beispiel ist der Q-Chat des Onlinedienstes Quizlet, der vor allem als klassischer, onlinebasierter Vokabeltrainer genutzt wird. Mithilfe von ChatGPT soll der Dienst zu einem Tutor werden, der sich auf den Fortschritt der Lernenden einstellt und diesen in dialogischer Form Fragen zum Lernmaterial stellt (Bayer 2023). Eine mögliche Weiterentwicklung ist die Einbindung in ein intelligentes Tutorsystem, das etwa im Informatikstudium individuell Unterstützung bei der Entwicklung von Programmierkompetenzen geben und Aufgaben generieren kann.[55]

Als *Risiko* wird in diesem Zusammenhang diskutiert, dass Schüler/innen bzw. Studierende die Unterstützung von ChatGPT möglicherweise in so großem Maße in Anspruch nehmen, dass ihr Lernprozess eingeschränkt wird und sie wichtige Kompetenzen nicht selbst entwickeln (Alouani 2023; Deutscher Ethikrat 2023, S. 267f.). Genannt werden etwa Fertigkeiten im schriftlichen Ausdruck (die unter anderem für die Persönlichkeitsentwicklung von Bedeutung sind) (Baron 2023) sowie Kompetenzen der Informationssuche und -bewertung (Zakir 2023). Bei einer zunehmend individuellen, maschinell gestützten Bildung könnten auch weitere Bildungsaufgaben etwa des sozialen Lernens, vernachlässigt werden, die zu einer verantwortlichen Teilhabe an der Gesellschaft beitragen (Brandstätter 2023).

Allerdings ist fraglich, inwiefern solche Ausweichprozesse in der Bildung tatsächlich erwartbar sind. Denn zum einen bestehen Bildungsprozesse unter anderem gerade darin, die Lernenden zum Kompetenzerwerb zu motivieren, sei es extrinsisch (durch geeignete Prüfungen, Kap. 5.1.2) oder intrinsisch. Die Lehrpersonen können etwa den Kompetenzerwerb durch Erfolgserlebnisse oder gemeinsames Lernen attraktiv gestalten oder ihn in Präsenzphasen überwachen (Spannagel 2023). Zum anderen ist bei Lernenden in der Regel davon auszugehen, dass sie motiviert sind und die Notwendigkeit der Lernschritte erkennen (Lordick 2023).[56]

---

54 Heinz-Peter Meidinger, Newsletter Bildung.Table #101 vom 25.1.2023.

55 Ute Schmid, mündliche Kommunikation, Press Briefing des Science Media Centers am 26.1.2023 (www.sciencemediacenter.de/alle-angebote/press-briefing/details/news/chatgpt-und-andere-sprachmodelle-zwischen-hype-und-kontroverse/, 19.4.2023).

56 Auch Robert Lepenies, Präsident der Karlshochschule in Karlsruhe, verweist in einem Interview im ZDF Morgenmagazin darauf, dass man Studierende nicht »unter Generalverdacht« stellen dürfe (www.zdf.de/nachrichten/zdf-morgenmagazin/chatgpt-robert-lepenies-100.html, 19.4.2023).

5.1.2 Perspektive der Lehrenden

In der intensiven Diskussion über ChatGPT unter Lehrer/innen und im Hochschulkontext wurden Potenziale für die Erleichterung der persönlichen Arbeit (etwa die Entlastung bei routinemäßigen Kommunikationsaufgaben), aber auch für die Verbesserung des Unterrichts bzw. der Lehre identifiziert (Gimpel et al. 2023; Kasneci et al. 2023; Mohr et al. 2023; allgemein zum Einsatz von KI für Lehr- und Lernzwecke: Europäische Kommission 2022). Die an der Debatte Beteiligten teilen vergleichsweise einhellig die Ansicht, dass es nicht darum gehe, ChatGPT und verwandte Systeme aus dem (sekundären bzw. tertiären) Bildungsbereich zu verbannen, sondern vielmehr als Herausforderung zu begreifen.[57] Es wird dabei auch eine Reihe von Problemen bzw. Risiken der Nutzung von ChatGPT durch Lehrpersonen diskutiert.

Didaktik und Unterrichtsgestaltung

Lehrende können mithilfe von ChatGPT den *Unterricht* bzw. die *Lehrveranstaltungen planen* bzw. Anregungen dafür erhalten. Aufgaben zu einem gegebenen Thema können vorgeschlagen und Materialien dazu erstellt werden (z. B. Impulse, Fragen, Beispiele, Quizze oder Anregungen) (Blume 2023; Mohr et al. 2023). Dabei bleibt die Lehrperson für die Auswahl und Bewertung der Ergebnisse verantwortlich; das System ist nicht spezifisch auf den jeweiligen Lehrplan angepasst und das Problem der fehlenden Verlässlichkeit der Antworten besteht.[58] Dennoch wird von dieser Art der Unterstützung erwartet, dass sie *von Routineaufgaben entlastet und Effizienzgewinne erbringt* (Gimpel et al. 2023; Kasneci et al. 2023). Die Wahl der richtigen Prompts für das System spielt auch in diesem Fall eine große Rolle für die Qualität und Passgenauigkeit der Ergebnisse (Mollick/Mollick 2023).

Über die Planung des Unterrichts hinaus kann ChatGPT auch – im Rahmen der datenschutzrechtlichen Möglichkeiten – in der *Durchführung des Unterrichts bzw. von Lehrveranstaltungen* eingesetzt werden (Mohr et al. 2023, S. 10). Dies wird insbesondere in sprachlichen Fächern diskutiert und auch bereits praktiziert (Brandstätter 2023; Meyer/Weßels 2023). Mit PEER ist bereits eine spezialisierte Anwendung zur Unterstützung des Schreibenlernens auf Basis von GPT-3 in Erprobung.[59] Auch gesellschaftswissenschaftliche Fächer bieten Anwendungsmöglichkeiten, etwa zur Entwicklung von Argumentationen

57 Dabei ist zu beachten, dass die an der Debatte Beteiligten nicht unbedingt repräsentativ für alle Lehrenden sind.

58 Eine solche Spezialisierung bieten auch Drittanbieter aus Deutschland nicht, über die Lehrende auf die KI-Systeme von OpenAI zugreifen können (z. B. Fobizz; Mindverse) Sie ermöglichen aber eine Nutzung ohne persönlichen Zugang bei dem US-amerikanischen Unternehmen, was aus Datenschutzgründen vorteilhaft sein kann (Kap. 6.1).

59 www.edu.sot.tum.de/en/hctl/forschung/peer/ (19.4.2023).

aus unterschiedlichen Perspektiven. Dabei ist es jeweils notwendig, die Ergebnisse des KI-Systems kritisch zu reflektieren – im gleichen Zug kann kritische Medienkompetenz im Umgang mit KI-Systemen gebildet werden (Mohr et al. 2023, S. 10). Der Einsatz von sprachverarbeitenden KI-Modellen in Unterricht und Lehre wird z. T. bereits wissenschaftlich evaluiert bzw. erforscht.[60] Darüber hinaus könnte sich ein stärkerer fachlicher Austausch der beteiligten Praktiker/innen über die gesammelten Erfahrungen anbieten, der sich etwa an den Leitfragen zur ethischen und verantwortungsvollen Nutzung von KI-Systemen der Europäischen Kommission (2022, S. 22ff.) orientieren könnte.

Bewertung von Leistungen

Diskutiert wird auch, inwiefern KI-Modelle zur Sprachverarbeitung genutzt werden können, um Leistungen der Schüler/innen bzw. Studierenden in Textform zu bewerten (Kasneci et al. 2023, S. 3). Ein automatisiertes Bewertungsverfahren von Aufsätzen wird aber beim bisherigen Entwicklungsstand als noch nicht möglich angesehen (Weßels 2022), zudem wirft es prüfungsrechtliche Fragen auf, da Prüfungsordnungen in der Regel die Tätigkeit eines/r Prüfer/in vorsehen (Hoeren 2023, S. 36). Daher scheint letztlich die Vorstellung weit hergeholt, es könnte zu regelrechten Zirkeln maschineller Kommunikation kommen, wenn Lernende Texte von einem KI-System erzeugen lassen und diese von einem anderen KI-System (teil-)automatisiert korrigiert bzw. bewertet werden (de Waard 2023).

Differenzierung und Inklusion

Eine für die Erstellung von Lehrmaterialien und Aufgaben besonders interessante Funktion der sprachverarbeitenden KI-Modelle ist die Möglichkeit, Texte in unterschiedlichen sprachlichen Stilen und auch unterschiedlichen Kompetenzniveaus zu erzeugen (Kap. 3.1.1). Auf diese Weise kann Lernstoff ohne den üblicherweise damit verbundenen Aufwand differenziert nach unterschiedlichen Lernniveaus angeboten werden (Blume 2023; Mohr 2023, S. 8). Denkbar sind auch Übersetzungen in einfache Sprache zur Unterstützung eines inklusiven Unterrichts (Kap. 4.4.2). Der Deutsche Ethikrat (2023, S. 179) betont allerdings, dass KI-Systeme »nicht die generelle Lösung für Fragen von Inklusion« sind, sondern nur unterstützend wirken können (ähnlich: Kreer 2023).

60 www.edu.sot.tum.de/en/hctl/forschung/peer/ (19.4.2023); www.uni-hildesheim.de/fb3/institute/iwist/forschung/forschungsprojekte/aktuelle-forschungsprojekte/ki-unterstuetztes-textfeedback-in-englischsprachigen-lehrveranstaltungen (19.4.2023); Brandstätter (2023).

Medienkompetenz

Zu den Herausforderungen durch ChatGPT zählt auch, dass den Lernenden *neue Kompetenzen* vermittelt werden müssen. Bereits durch Suchmaschinen wie Google haben sich Recherchekompetenzen grundlegend verändert. Auch für die (kritische) Bewertung der Funktionsweise und der Ergebnisse von ChatGPT (Dolderer 2022) bzw. des (kritischen) Hinterfragens der KI-Technologie und -Industrie (Sheldon 2022), werden neue Kompetenzen benötigt (Floridi/Chiriatti 2020, S. 692f.). Nicht zuletzt wurde auf die Bedeutung der kompetenten Wahl von Prompts für eine effiziente Nutzung des Systems hingewiesen (van Deyzen 2023; Kap. 3.1.2).

Diese Kompetenzen sollten im Bildungsprozess vermittelt werden, aber auch Gegenstand von Fort- und Weiterbildungen für Lehrende sein (Salden et al. 2023, S. 19) – nicht nur für solche, die diese Kompetenzen selbst vermitteln, sondern für alle, die mit KI-Modellen zur Sprachverarbeitung konfrontiert sind. Außerdem könnte sich die Bedeutung bisheriger Kompetenzen verschieben – neben dem Faktenwissen, das zur Beurteilung der Ergebnisse der KI-Systeme wichtig bleibt, gewinnt die Fähigkeit zur Anwendung von Wissen weiter an Bedeutung (Blume 2023).

Für die *Teilhabe an der digitalen Gesellschaft* wird es zukünftig eine Rolle spielen, Systeme wie ChatGPT kompetent nutzen zu können, etwa als Anforderung in vielen Berufen. Erfahrungen mit bisherigen digitalen Werkzeugen haben gezeigt, dass neue Medien häufig bestehende Bildungsungleichheiten verstärken, weil gute Schüler/innen stärker von ihnen profitieren als schwächere. Gleichzeitig steigt mit der Verbreitung digitaler Anwendungen das Anspruchsniveau an menschliche Tätigkeiten, sodass ein »Schereneffekt entsteht« (Honegger 2023a, b). KI-Systeme wie ChatGPT könnten allerdings auch einen ausgleichenden Effekt in Bezug auf Bildungsungleichheit haben, indem sie denjenigen Schüler/innen Hilfe anbieten, die in ihrer Familie keine oder wenig Unterstützung bekommen (Blume 2023). Bisherige Beobachtungen der Nutzung von ChatGPT in der Praxis deuten allerdings auf eine Verstärkung von Ungleichheiten hin (Brandstätter 2023; Wedig 2023).

Auch eventuelle Kosten des Angebots könnten soziale Ungleichheiten in Bezug auf den Bildungserfolg verstärken. Daran knüpfen Forderungen an, allen Menschen unabhängig von ihrer finanziellen Situation den Zugang zu Systemen wie ChatGPT zu gewährleisten (Dang 2023).

Prüfungen

Texterzeugnisse spielen in den allermeisten Fachgebieten eine fundamentale Rolle für die Bewertung von Lernfortschritten. Der sprachliche Ausdruck wird behelfsweise herangezogen, um die Entwicklung von Wissen und (z. B. gedankliche) Kompetenzen beurteilen zu können, die nicht direkt beobachtbar

sind (Mahowald/Ivanova 2022). ChatGPT stellt diesen angenommenen Zusammenhang infrage und damit die bisherige Prüfungspraxis an Schulen und insbesondere an Hochschulen (wo die Distanz zwischen Lehrenden und Lernenden größer ist) vor große Herausforderungen.

Im Gegensatz zu Plagiaten ist es im Fall von Texten, die mit sprachverarbeitenden KI-Systemen erzeugt wurden, bislang nicht möglich, sie automatisiert zuverlässig zu erkennen. Für die Überprüfung wurde bereits eine ganze Reihe von – ebenfalls auf KI-Technologien beruhenden – Anwendungen entwickelt (Polomski 2023), die sich jedoch bisher als nicht effektiv erwiesen haben (Gao et al. 2022). Zur *Erkennung von Texten, die von KI-Modellen erzeugt wurden*, wird daher auch eine Kennzeichnung durch eine Art Wasserzeichen diskutiert (Heikkilä 2023b). Forschende gehen von einem Wettlauf zwischen KI-Systemen zur Erzeugung und solchen zur Entdeckung künstlicher Texte aus (Heidt 2023). Von Überprüfungen studentischer Texte durch KI-Systeme rät das Bayerische Kompetenzzentrum für Fernprüfungen ab, weil die Funktion der Tools nicht nachprüfbar und ihr Ergebnis daher nicht rechtssicher sei, zudem bestünden datenschutzrechtliche Einwände (Besner et al. 2023; Hoeren 2023; Sperl 2023).

Ein *Verbot von ChatGPT als unerlaubtes Hilfsmittel* dürfte sich in der Praxis kaum durchsetzen lassen, insbesondere, weil sich eine Nutzung nur unter Umständen nachweisen lässt. Auch gegenüber Mitteln wie der Durchführung von handschriftlichen Prüfungen bestehen Bedenken, weil diese manche Lernende benachteiligen (Heidt 2023). Teilweise könnten nur mit großem Aufwand abgesicherte Prüfungsumgebungen geschaffen werden (für Onlineprüfungen auch mit Videobeaufsichtigung), die eine Kontrolle ermöglichen (Besner et al. 2023). Dagegen werden didaktische oder organisatorische Lösungen wie die explizite Einbindung von ChatGPT als Hilfsmittel, die Protokollierung oder Erläuterung des Arbeitsprozesses als Teil der Aufgabe oder ergänzende mündliche Prüfungsleistungen als Alternativen diskutiert (van Dis et al. 2023; Gimpel et al. 2023, S. 31ff.; Haverkamp 2022).

Nicht zuletzt erfordern die faktischen Fehler, die in Texten von ChatGPT vorkommen, ebenso wie die Beschränktheit der Trainingsdaten von den Lernenden eine kompetente Beurteilung der Texte, wenn sie nicht Gefahr laufen wollen, eine wenig überzeugende Prüfungsleistung und damit potenziell eine schlechte Bewertung zu erhalten (Blume 2023). Ein bloßes Delegieren von Prüfungsleistungen an das System dürfte daher ohnehin keine sinnvolle Strategie darstellen. Allerdings sind Bildungsinstitutionen gefordert, ihre Prüfungsregeln zu überprüfen, um die Nutzung bzw. Nichtnutzung von Hilfsmitteln wie ChatGPT rechtssicher zu gestalten (Hoeren 2023).

5.1.3 Institutionenperspektive

Institutionen des Bildungssystems sind durch Systeme wie ChatGPT zum einen *herausgefordert*, ihre Prüfungs- wie auch ihre Lehrpraxis zu überdenken und ggf. neu auszurichten. Eine Vielzahl an Veranstaltungen, internen Aktivitäten sowie Empfehlungen der Schulen und Hochschulen wie auch der Bildungsverwaltung zeigt, dass diesbezüglich ein ausgeprägtes Bewusstsein herrscht (Gimpel et al. 2023; Gross/dpa 2023; Schulministerium NRW 2023). Zum anderen können die Systeme zur *Erfüllung von Verwaltungsaufgaben* genutzt werden. Analog zum Einsatz von KI-basierten Chatbots in der öffentlichen Verwaltung lassen sich etwa an Hochschulen Angebote wie die Rekrutierung neuer Studierender, die Studienberatung oder Informationen über das Campusleben durch den Einsatz von Chatbots unterstützen. KI-Modelle zur Sprachverarbeitung lassen sich zu Erstellung von Veranstaltungshinweisen, von Kursbeschreibungen sowie von Anerkennungsverfahren einsetzen.[61]

Auf der Ebene des Bildungssystems insgesamt ergeben sich Notwendigkeiten und damit *Chancen* für die Weiterentwicklung von Bildungsformen, wenn etwa Lehrende gemeinsam mit den Lernenden erarbeiten, wie neue Vermittlungs- und Aneignungsformen unter Einbeziehung von KI-Technologien aussehen könnten. Solche Erprobungen finden derzeit etwa am »Virtuellen Kompetenzzentrum – Schreiben lehren und lernen mit Künstlicher Intelligenz«[62] an der Fachhochschule Kiel statt. Außerdem wird erwartet, dass die Auseinandersetzung mit ChatGPT Anlass dazu gibt, »über Begriffe wie ›Information‹ und ›Lernen‹ gründlich nachzudenken« (Krischke 2023, S. 4) und dabei bestehende Bildungspraktiken zu hinterfragen. Daraus können sich Ideen für neue Kursformate und Prüfungsformen etwa an Hochschulen ergeben (Thorp 2023, S. 313), aber auch ein kritischer Blick auf die Bewertung von wissenschaftlichen Leistungen allein anhand quantitativer Maße (Else 2023, S. 423) oder – im Bereich der Forschung – auf die Leistungsfähigkeit des Peer-review-Systems (Seife 2023).

Die Nutzung von ChatGPT in und durch Bildungsinstitutionen birgt aber auch eine ganze Reihe von *Risiken*. Neben der Herausforderung, einen Umgang mit den Grenzen und unerwünschten Auswirkungen des KI-Systems (Kap. 3.2) zu finden, bestehen sie zum einen darin, dass der Schutz personenbezogener Daten nicht ohne weiteres gewährleistet werden kann, zum anderen in der

61 Christian Spannagel, mündliche Kommunikation, Webinar des Multimediakontors Hamburg am 14.3.2023. Eine unreflektierte Automatisierung der Kommunikation birgt allerdings die Gefahr des Verlusts zugewandter menschlicher Umgangsformen. An einer US-amerikanischen Hochschule verschickte ein College nach einer Amoktat ein Schreiben, das unter Nutzung von ChatGPT verfasst worden war (https://vanderbilthustler.com/2023/02/17/peabody-edi-office-responds-to-msu-shooting-with-email-written-using-chatgpt/, 19.4.2023). Die Collegeverwaltung entschuldigte sich anschließend.

62 www.vkkiwa.de (19.4.2023).

Notwendigkeit, *Benachteiligung oder Diskriminierung* im Zuge der Nutzung auszuschließen (Schwarz 2023). Während letzteres aufgrund der Intransparenz des Systems und der mangelhaften Sicherheitsvorkehrungen kaum möglich ist, wird auch eine *datenschutzkonforme Anwendung* im Rahmen von Unterricht oder Lehre als nicht möglich angesehen (Honegger 2023a; Thiede 2023). So ist der Zugriff auf ChatGPT Personen unter 18 Jahren nur mit Einverständnis der Sorgeberechtigten erlaubt, aufgrund der Erhebung und – weitgehend unklaren Verwendung – von Daten der Nutzenden (Kap. 6.1) ist auch bei Erwachsenen im Bildungskontext eine Nutzung, die die Eingabe persönlicher Daten erfordert, nicht gestattet. In Italien wurde Ende März 2023 eine vorübergehende vollständige Sperrung von ChatGPT veranlasst, unter anderem mit Verweis auf den Jugendschutz, weil OpenAI das Alter der Nutzenden nicht ausreichend kontrolliert (Hahn 2023c).

Schulen und Hochschulen in Deutschland nutzen bislang zwei Wege, um trotz der Datenschutzprobleme einen Einsatz zu ermöglichen. Zum einen wird das System mit dem Zugang einer Lehrperson (sofern diese dazu bereit ist) oder mit anonymen, z. B. fiktiven Nutzungsdaten verwendet. In diesem Fall ist zu beachten, dass auch durch die Nutzung selbst keine personenbezogenen Daten in das System eingegeben werden. Zum anderen können über die Bildungsinstitution bzw. über Drittanbieter Zugangskonten eingerichtet werden, deren Daten gesetzeskonform verarbeitet und nicht an OpenAI weitergegeben werden (Wedig 2023); andere Anbietende von Bildungsplattformen arbeiten mit pseudonymisierten Daten (Beer 2023). Eine lokale Installation eines KI-Modells zur Sprachverarbeitung, sofern dies technisch bzw. finanziell realisierbar ist (Kap. 3.3.3), könnte eine datenschutzkonforme Anwendung erleichtern. Eine rechtliche Klärung der Praxis steht noch aus.[63]

5.2 Chancen und Risiken im Bereich der Forschung

Im Bereich der Forschung haben erste Veröffentlichungen, die ChatGPT als Autor/in nennen, für Aufsehen gesorgt (Stokel-Walker 2023). Die Redaktionen der Zeitschriften Nature und Science machten daraufhin klar, dass eine Autorschaft eines KI-Systems nicht akzeptiert werde, weil nur Menschen die Verantwortung für die veröffentlichten Forschungsergebnsise übernehmen können (Nature Editorial 2023; Thorp 2023). Auch das deutsche Urheberrecht sieht nur natürliche Personen als Urheber vor (Hoeren 2023, S. 26). Eine Nutzung von KI-Systemen im Forschungsprozess wie auch beim Verfassen einer Publikation ist dagegen zulässig, sofern diese – entsprechend der Regeln guter wissenschaftlicher Praxis – transparent gemacht wird.

63 Mündliche Auskunft von Teilnehmenden einer Onlinekonferenz zu künstlicher Intelligenz in der Bildung am 2.2.2023.

5.2.1 Anwendungsmöglichkeiten von KI-Modellen zur Sprachverarbeitung

Auch unabhängig von ChatGPT gibt es eine ganze Reihe von KI-basierten Anwendungen, die das wissenschaftliche Arbeiten unterstützen (de Waard 2023). So lassen sich beispielsweise Daten aus einer Veröffentlichung extrahieren oder Studienergebnisse zusammenfassen – für letztere Aufgabe kann ein System auf Basis von GPT-3 genutzt werden (Schlender 2022). Insbesondere beim *Schreiben von wissenschaftlichen Texten* wie Zeitschriftenbeiträgen oder Anträgen, in denen vergleichsweise standardisierte Abschnitte wie ein Überblick über die verwendeten Methoden oder den Forschungsstand enthalten sind, kann ein KI-Modell zur Sprachverarbeitung Unterstützung leisten. Zukünftig könnte auch eine Hilfestellung bei Reviewprozessen möglich sein.

Als positive Aspekte einer solchen Nutzung werden genannt (Berdejo-Espinola/Amano 2023; van Dis et al. 2023):

- die Beschleunigung von Forschung, Entwicklung und damit letztlich Innovationen durch Verkürzung der Zeit bis zur Publikation von Forschungsergebnissen,
- die Entlastung der Forschenden von Routineaufgaben zugunsten der Konzentration auf neue Entdeckungen sowie
- die Verringerung der Benachteiligung von Forschenden, die nicht kompetent in der Nutzung der englischen Sprache sind, bei Publikationen in internationalen Zeitschriften.

Außer dem Schreiben können Chatbots zukünftig auch Ideen für die eigentliche *Forschungsarbeit* beisteuern, etwa für das Design wissenschaftlicher Experimente oder bei der Interpretation von Daten (van Dis et al. 2023). Für die Auswertung von Text- oder Sprachdaten z. B. in den Sozialwissenschaften könnten sie ebenso hilfreich sein wie für Programmieraufgaben. Auch die Linguistik, auf deren Erkenntnissen die Modelle zu einem guten Teil beruhen, könnte umgekehrt vom Einsatz bzw. der Erforschung der sprachverarbeitenden KI-Modelle profitieren. Anwendungen in der Forschung zu Molekülstrukturen zeigen, dass die Modelle auch auf biologischen Code anwendbar sind (Dalla-Torre et al. 2023; Lin et al. 2023; Kap. 3.1).

5.2.2 Probleme und Risiken der Anwendung in der Forschung

Auch im Bereich der Forschung ist eine Anwendung von sprachverarbeitenden KI-Modellen nur unter strikter – und aufwendiger – Kontrolle möglich. Neben

dem Fabulieren können auch die Ungenauigkeiten und der mögliche Bias der Modelle ein Problem darstellen (van Dis et al. 2023).[64]

Für Anwendungen in der wissenschaftlichen Arbeit ebenfalls sehr problematisch ist die *fehlende Zuordnung von Informationen zu einer Quelle* (Rogers 2023). Sie ist zum einen als Ausweis von Urheberschaft relevant, zum anderen aber auch zur Kontextualisierung von Ideen und als Nachweis der Verbundenheit mit bestimmten Denkweisen bzw. -schulen. Im Zuge des Trainings und der Parametrisierung des KI-Modells geht diese Verbindung der Konzepte zu ihrem Ursprung verloren, was die Anwendung im wissenschaftlichen Kontext problematisch macht.

Ein Risiko, potenziell auch für das Verhältnis von Wissenschaft und Öffentlichkeit, stellt die *fehlende Faktenorentierung* der KI-Modelle dar. »Indem Sprachmodelle Texte produzieren, die wissenschaftlich klingen, es aber nicht sind, könnten sie das Problem der Desinformation verschärfen und das Vertrauen in die Wissenschaft erodieren. Für die Forschung bedeutet das, dass gesichertes Wissen bald in noch größerer Konkurrenz mit plausibler Propaganda stünde. Das Risiko ist umso größer, wenn Wissenschaftler/innen selbst solche Tools nutzen, um einfach nur schnell zu publizieren. Denn klar ist, dass auch Forschende anfällig für plausiblen Unsinn sind, oder, anders gesagt, der Publikationsdruck sie oftmals dazu zwingt, Plausibilität genügen zu lassen.« (Fecher/Schulz 2023)

Das schnelle Publizieren könnte durch Systeme wie ChatGPT noch beschleunigt werden, die Rate an wissenschaftlichen Publikationen damit weiterwachsen. Bereits jetzt lässt sich die *Menge an Publikationen* kaum noch einer sorgfältigen Prüfung unterziehen, auch der Überblick über den Forschungsstand wird dadurch erschwert. Dabei besteht zum einen die Hoffnung, dass durch eine verbreitete Verwendung von ChatGPT diese problematischen Auswüchse der akademischen Forschung sichtbar werden und ins Bewusstsein der wissenschaftlichen Gemeinschaft und der Öffentlichkeit gelangen (Else 2023; Seife 2023). Zum anderen dürften sich durch die Verwendung KI-gestützter Hilfsmittel auch die Arbeitslast und der Konkurrenzdruck weiter erhöhen (van Dis et al. 2023), so dass Forschende gar nicht anders können, als Hilfsmittel wie ChatGPT zu nutzen, wodurch der Prozess sich selbst verstärkt.

64 Interessanterweise wird vorgeschlagen, dass die Archive wissenschaftlicher Zeitschriften für das Training der KI-Modelle genutzt werden sollten (van Dis et al. 2023) – der Fokus liegt allerdings auf Open-Source-Modellen von gemeinnützigen Organisationen.

6 Rechtliche Aspekte und Fragen der Nachhaltigkeit

Die bisherige Betrachtung hat eine Reihe von rechtlichen Aspekten der Nutzung von ChatGPT und vergleichbaren Systemen aufgeworfen, insbesondere den Schutz persönlicher Daten (Kap. 6.1) und Fragen des Urheberrechts (Kap. 6.2) betreffend. Außerdem geraten Aspekte der Nachhaltigkeit in den Blick (Kap. 6.3).

6.1 Datenschutz

Fragen des Datenschutzes bei ChatGPT betreffen sowohl die *im Training verwendeten* als auch die *im Zuge der Nutzung entstehenden Daten*. In Bezug auf erstere ist nicht auszuschließen, dass personenbezogene Informationen als Teil der Trainingsdaten Eingang in das Modell gefunden haben. Ein solcher Fall könnte nur durch effektive Filter ausgeschlossen werden, was angesichts der sehr großen Datenmenge eine Herausforderung darstellen dürfte. Bei einem KI-Modell zur Sprachverarbeitung lässt sich nach Abschluss des Trainings weder ohne weiteres überprüfen, welche Daten im Modell enthalten sind, noch lassen sich diese korrigieren bzw. löschen (Gal 2023).

Allerdings lassen sich unter Umständen durch rückwärtsgerichtetes Suchen Daten aus dem Trainingsmaterial rekonstruieren, was im Fall von personenbezogenen Daten hoch problematisch ist (Carlini et al. 2021; Wayner 2023). Dabei ist nicht ganz klar, inwiefern das Auslesen von Informationen im Netz zur Sammlung der Trainingsdaten rechtlich erlaubt ist. Ein Unternehmen, das Bilddaten von Privatpersonen sammelte, um damit Algorithmen zur Gesichtserkennung zu trainieren, wurde zu hohen Geldstrafen verurteilt – allerdings ist dieser Fall nicht mit KI-Modellen zur Spracherzeugung vergleichbar (Poireault 2023; Zakir 2023).

Bei der Nutzung von KI-Modellen zur Sprachverarbeitung könnten sensible und/oder personenbezogene Daten, die von den Nutzenden eingegeben werden, von den Betreiber/innen missbräuchlich verwendet werden. Die Datenschutzbestimmungen von OpenAI sind zwar sehr detailliert, bleiben in diesem Punkt allerdings vage (Thiede 2023). Einige Unternehmen haben ihren Angestellten daher die Nutzung von ChatGPT untersagt (Lukpat 2023). Im Kontext von schulischer Bildung und Lehre ist eine Nutzung nur möglich, wenn dabei sichergestellt werden kann, dass keine personenbezogenen Daten an das System übermittelt werden (Kap. 5.1.3). Eine datenschutzrechtliche Klärung steht diesbezüglich noch aus.

6.2 Urheberrecht

Auch in Bezug auf urherrechtliche Aspekte ergeben sich einige offene Fragen (zum Folgenden: Hoeren 2023):

Wie ist die Nutzung von möglicherweise urheberrechtlich geschützen Daten für das Training zu bewerten?

Im Urheberrecht (§ 44b UrhG) existiert zwar EU-weit eine Schrankenregelung, die eine automatisierte Analyse von Werken zulässt. Aufgrund der Neuheit der KI-Modelle zur Sprachverarbeitung lässt sich allerdings nicht klar beurteilen, ob deren Trainingsverfahren den Bedingungen dieses Gesetzes entspricht (Hoeren 2023, S. 28), Zeitungsverleger bestreiten dies (Voß 2023). Im Zusammenhang mit Bildgeneratoren bzw. der Plattform Github sind derzeit mehrere Verfahren anhängig, die die Rechte an den für das Training der Modelle verwendeten Daten betreffen (Brittain 2023; Vincent 2023; Wittenhorst 2022), die Fälle sind allerdings nicht unbedingt mit sprachverarbeitenden KI-Modellen vergleichbar.

Wer besitzt die Rechte an den Ergebnissen eines KI-Modells zur Sprachverarbeitung?

In diesem Fall ist im deutschen Recht klar geregelt, dass nur natürliche Personen Anspruch auf urheberrechtlichen Schutz erheben können (Kreutzer 2021). Inwiefern die Erzeugnisse des KI-Modells frei von Urheberrechten sind, hängt zum einen davon ab, ob das Ausgangsmaterial in den Erzeugnissen noch weitgehend unverändert enthalten ist. In diesem Fall bestehen auch für die KI-Erzeugnisse Urheberrechte der Rechteinhaber des Ausgangsmaterials. Allerdings ist dies aufgrund der fehlenden Quellenverweise für die Nutzenden kaum zu erkennen (Hoeren 2023, S. 28; Kasneci et al. 2023, S. 6). Zum anderen kommt es für die Frage des Schutzes der Ergebnisse darauf an, ob darin eine Schöpfung im Sinn des Urheberrechts zu erkennen ist und ob sie einer natürlichen Person zugeordnet werden kann. Dies könnte der Fall sein, wenn ein Prompt so detailliert formuliert wird, dass darin wesentliche Gestaltungsentscheidungen über das Ergebnis enthalten sind – in diesem Fall könnte der/die Promptautor/in möglicherweise Urheberrechte beanspruchen (Hoeren 2023, S. 26).

Müssen Texte, die von einem KI-Modell zur Spracherzeugung stammen, als solche gekennzeichnet werden?

Für die Frage einer Kennzeichnungspflicht sind die Lizenz- und Nutzungsbedingungen des Betreibers bzw. der Betreiberin sowie die Regeln des jeweiligen Verwendungskontextes bestimmend. Grundsätzlich besteht keine Kennzeich-

nungspflicht, im akademischen Kontext sowie bei schulischen Prüfungen sind entsprechende Pflichten allerdings meist durch die Prüfungsordnungen bzw. die Regeln guter wissenschaftlicher Praxis festgelegt (Hoeren 2023, S. 29).

Dürfen fremde Texte als Prompts in das System eingegeben werden?

Da sich OpenAI die Nutzung der als Prompt eingegebenen Daten laut der Nutzungsbedingungen vorbehält, sind eventuelle Rechte an den Prompts zu beachten. Prüfungsleistungen beispielsweise sind urheberrechtlich geschützt, es wäre also unzulässig, sie – beispielsweise für eine Bewertung durch ChatGPT – in das System einzugeben (Hoeren 2023, S. 37).

6.3 Nachhaltigkeitsaspekte

ChatGPT und andere sehr große KI-Modelle zur Sprachverarbeitung werfen Probleme sowohl in Hinblick auf ökologische als auch soziale Aspekte der Nachhaltigkeit auf (zu ökonomischen Aspekten der Nachhaltigkeit s. Kap. 2.5). Mit Blick auf *Umweltauswirkungen* wird auf den sehr hohen Energieverbrauch der Modelle in der Trainingsphase hingewiesen (Bender et al. 2021; Kap. 2.4). Dieser fällt zwar nur einmal an, allerdings deutet die rasche Abfolge neuer Modellvarianten darauf hin, dass die Lebensdauer einer Modellvariante nicht allzu hoch ist. Außerdem wird damit gerechnet, dass mit weiterhin exponentiell steigender Größe der Modelle bald eine Grenze dessen erreicht ist, was an Rechenleistung und Energie dafür aufgewendet werden kann (Patel 2023).

Mit Blick auf *soziale Aspekte der Nachhaltigkeit* ist darauf zu verweisen, dass die Entwicklung von KI-Systemen immer wieder mit der Inanspruchnahme großer Mengen menschlicher Arbeitskraft verbunden ist. Damit ist nicht in erster Linie die Arbeit der Entwickler/innen gemeint, sondern diejenigen, die für meist geringe Entlohnung und ohne Aufstiegs- und Weiterbildungsmöglichkeiten einfache Tätigkeiten wie das Kodieren von Daten übernehmen, um die Systeme zu trainieren (Rohde et al. 2019, S. 55). Bereits die Entwicklung der ImageNet-Datenbank, auf der wesentliche Fortschritte der Bilderkennung beruhen, wurde nach Aussagen der Begründerin nur möglich dank der Clickworker auf der Mechanical-Turk-Plattform von Amazon.[65] Im Fall von OpenAI wurden Arbeiter/innen in Kenia gegen geringen Lohn einer psychisch z. T. sehr belastenden Tätigkeit ausgesetzt (Perrigo 2023). Nicht zuletzt wird auch hinterfragt, inwiefern die kostenlose Nutzung von kreativen Werken durch private, nicht gemeinnützige Unternehmen wie OpenAI fair und rechtmäßig ist (Wayner 2023).

65 https://learning.acm.org/techtalks/ImageNet (19.4.2023, ab Min. 25:40).

7 Weiterführende Fragen

Die breite öffentliche Debatte, die nach Veröffentlichung von ChatGPT eingesetzt hat, mag *einerseits* als Hype erscheinen, wie er typischerweise die Einführung vieler neuer Technologien begleitet. Insofern hat das vorliegende *Hintergrundpapier* des TAB versucht, diesem Textformat gerecht zu werden und hinter die auf der Oberfläche sichtbaren Entwicklungen zu schauen. Dargestellt wurden dabei – vor dem Hintergrund einer weiterhin sehr dynamisch verlaufenden technologischen Entwicklung und öffentlichen Debatte –

- die technologische Entwicklung der KI-Modelle zur Sprachverarbeitung ausgehend von der wegweisenden Transformerarchitektur bis zum jüngst angekündigten GPT-4,
- Grundzüge der wirtschaftlichen und rechtlichen Rahmenbedingungen dieser Entwicklung,
- Möglichkeiten und Grenzen, die diesen Systemen der Künstlichen-Intelligenz-Forschung inhärent sind (einschließlich der Versuche, sie zu überwinden),
- Anwendungspotenziale in unterschiedlichen gesellschaftlichen Bereichen, sowie
- erste Überlegungen zu möglichen Auswirkungen solcher Anwendungen

Diese Debatte kann aber *andererseits* auch als Ausdruck eines starken öffentlichen Interesses an innovativen Technologien angesehen werden. Sie zeigt die Bereitschaft, sich intensiv mit sprachverarbeitenden KI-Modellen und ihren Auswirkungen zu beschäftigen. Als Beispiel sei auf die Aktivitäten an Schulen und Hochschulen verwiesen, in denen sich akute Betroffenheit (in Bezug auf den Umgang mit Prüfungsleistungen) widerspiegelt, aber auch Interesse an einer gemeinsamen, reflexiven Gestaltung der Nutzung dieser neuen Technologie.

Zum Abschluss dieses Papiers sollen weiterführende Fragen entlang der Dimensionen der Regulierung, technologischer Weiterentwicklungen, des Austauschs von Wissen und der Forschungsförderung angerissen werden, die möglicherweise Anregungen für die weitere Debatte geben können.

7.1 Regulierung

Sprachverarbeitende KI-Modelle greifen in die Grundlagen der menschlichen Wissensverarbeitung und des menschlichen Zusammenlebens ein. Daher erscheint eine staatliche Festlegung der Rahmenbedingungen, unter denen sie entwickelt und angewendet werden, angemessen und ggf. angebracht. Die bisherige Entwicklung und vor allem die öffentliche Verfügbarkeit von ChatGPT und

seinen Nachfolgemodellen hat einige *Gefahren einer unregulierten Entwicklung* deutlich gemacht. Eine größere Transparenz von Daten und Modellverhalten, die Möglichkeit der Kennzeichnung bzw. der Erkennbarkeit der Erzeugnisse der Systeme und ein wirksamer Schutz der Daten der Nutzenden erscheinen dabei besonders dringlich.

Insbesondere die Institutionen der Europäischen Union erarbeiten derzeit mit dem *Gesetz über Künstliche Intelligenz* (engl. AI Act) ein Regelwerk auch für sprachverarbeitende KI-Modelle. Bis die Verhandlungen darüber abgeschlossen sein werden und das Gesetz in Kraft getreten ist, stellt sich die Frage, ob und an welchen Stellen mit Blick auf die skizzierten Anwendungsmöglichkeiten und Implikationen akute Regelungslücken bestehen und mit welchen Maßnahmen im Rahmen der bestehenden Gesetze diese ggf. geschlossen werden können.[66] Gleichzeitig könnte es nötig werden, für die jeweiligen Geschäftsbereiche der öffentlichen Verwaltung Leitfäden für den Umgang mit den neuen technologischen Möglichkeiten zu entwickeln.

Zudem stellt sich die Frage, welche Ansätze einer nicht zwingenden bzw. *nichtgesetzlichen Regulierung* (engl. soft law) möglicherweise kurzfristig Wirkung entfalten können. So könnten Entwickler/innen von KI-Systemen zu stärkeren interdisziplinären und transdisziplinären, also in die Gesellschaft hineinwirkenden Diskussionen aufgefordert werden (Kap. 7.3). Ebenfalls könnte eine bezüglich der Grundlagen und Auswirkungen der Systeme transparentere Entwicklung angestoßen werden, etwa durch gezielte Förderung von Forschung zu einem möglichen, in den Modellen inhärenten Bias und zu aussagekräftigen Benchmarktests für die Systeme (Ananthaswamy 2023, S. 204; van Miltenburg et al. 2023).

7.2 Technologische Weiterentwicklungen

Verschiedene Weiterentwicklungen sollen Probleme, die von den KI-Systemen aufgeworfen werden, abmildern oder beheben. Es werden Technologien zur Identifizierung der Erzeugnisse von KI-Systemen mittels Wasserzeichen (Kirchenbauer et al. 2023) sowie zur Bekämpfung von Desinformation entwickelt (Marcus 2023). Auch die Prüfsteine (Benchmarks und Audits) in der Zulassung von KI-Systemen (Liang et al. 2022) können zwar nicht allein technisch durchgeführt werden (Papakyriakopoulos et al. 2021), benötigen aber technische Unterstützung.

66 In den USA, wo es bislang keine spezifischen gesetzlichen Regelungen für KI gibt, ist die Handels- und Verbraucherschutzbehörde FTC öffentlich aktiv geworden und weist Anbieter/innen von KI-Systemen regelmäßig auf die Regeln hin, nach denen sie gegen unlautere Praktiken vorgehen kann (Atleson 2023). Die Behörde wurde Ende März von einer gemeinnützigen Forschungsorganisation aufgefordert, Ermittlungen gegen OpenAI einzuleiten (Hahn 2023c).

Eine weitere Frage in Zusammenhang mit der technologischen Entwicklung ist die nach möglichen Verkleinerungen bzw. Verschlankungen der Systeme, auch um den ökologischen Fußabdruck von Entwicklung und Betrieb zu verringern. Der Technology Policy Council der US-amerikanischen Informatikgesellschaft Association for Computing Machinery empfiehlt darüber hinaus »koordinierte, klare und durchsetzbare staatliche Maßnahmen und Gesetze«, um den Beitrag des IT-Sektors zur Erreichung des Ziels der Verringerung klimaschädlicher Emissionen sicherzustellen (ACM 2021).

7.3 Austausch von Wissen und Stakeholderbeteiligung

Gleich mehrere Gruppen von Expert/innen und Wissenschaftler/innen haben Anfang des Jahres 2023 zu einer intensiveren Reflexion der Entwicklung von sprachverarbeitenden KI-Modellen aufgerufen, mit offenbar ganz unterschiedlichen Intentionen.[67] An Hochschulen und Bildungseinrichtungen hat bereits eine intensive Debatte begonnen, dabei stellt sich allerdings die Frage, wie alle relevanten Stakeholder einbezogen werden können. In parlamentarischen Anfragen und Debatten der Bundestagsabgeordneten wurde ChatGPT bereits thematisiert (z. B. Bundesregierung 2023b; Deutscher Bundestag 2023a und 2023b, S. 10305ff.), in verschiedenen Bundesbehörden wurden Tests damit durchgeführt (Bundesregierung 2023a). Eine Anwendung sprachverarbeitender KI-Modelle für genuin politische Zwecke, etwa innerhalb von Parteien oder der parlamentarischen Arbeit, wurde in Deutschland bisher noch kaum thematisiert.[68]

Weitere Fragen betreffen den *Wissensaustausch zwischen gesellschaftlichen Gruppen*: Welche Kompetenzen sind für einen (kritisch-reflexiven) Umgang mit KI-Modellen zur Sprachverarbeitung nötig, und wie lassen sie sich am besten vermitteln? Wie lassen sich diejenigen, die praktische Erfahrungen mit der Anwendung von KI-Systemen beispielsweise im Bildungsbereich sammeln, untereinander vernetzen und ihre Erfahrungen systematisch auswerten? Welches interdisziplinäre Wissen fehlt, um die Blackboxen der KI-Modelle zu öffnen und sie theoretisch besser zu durchdringen? Und wie kann öffentlich über

67 https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (19.4.2023) – kritisch dazu Becker (2023) sowie Spielkamp (2023); www.openpetition.eu/petition/online/securing-our-digital-future-a-cern-for-open-source-large-scale-ai-research-and-its-safety (19.4.2023); www.law.kuleuven.be/ai-summer-school/open-brief/open-letter-manipulative-ai (19.4.2023); van Dis et al. (2023).

68 Erste Überlegungen zu »Einsatzmöglichkeiten von KI-gestützten Technologien zur Verbesserung der parlamentarischen Funktionserfüllung« wurden allerdings von Pilniok (2021, hier S. 181) entwickelt.

die Technologie und ihre Möglichkeiten, aber auch Grenzen gesprochen werden, ohne dadurch unerwünschten Entwicklungen Vorschub zu leisten?

7.4 Forschungspolitik und -bedarf

Mit Blick auf die wissenschaftliche Forschung lassen sich mehrere Gebiete identifizieren, in denen Wissenslücken zu bestehen scheinen. Dies betrifft zum einen die bereits erwähnten Fragen der Überprüfung und des Benchmarkings von KI-Systemen sowie deren theoretische Beschreibung, zum anderen wird immer wieder auch eine verstärkt interdisziplinäre, die sozial- und computerwissenschaftliche Perspektive übergreifende Forschung gefordert (Crawford/ Calo 2016, S. 311). Auch ein kontinuierliches Monitoring der Entwicklung, wie es bereits von unterschiedlichen Stellen angestrebt wird, kann hilfreich sein, möglicherweise negative Entwicklungen frühzeitig zu identifizieren.

Einer solchen Auseinandersetzung mit den vielfältigen Aspekten der Nutzung von KI-Modellen zur Sprachverarbeitung steht eine offenbar in starkem Maße von *Konkurrenzkampf* getriebene Herangehensweise der Entwickler/innen gegenüber (Briegleb 2023; Roose 2023b). Während für den Schutz von Persönlichkeitsrechten und Daten bei den IT-Konzernen eigene Abteilungen eingerichtet wurden, steht die Prüfung ethischer Fragen der Anwendung von KI-Systemen demgegenüber zurück (Tiku et al. 2023; Wolfangel 2021). Doch alternative Ansätze von verantwortungsvoller Forschung und Innovation wurden bereits entwickelt (TAB 2016b). Auch eine enge Verknüpfung der Forschungsaktivitäten und -ergebnisse mit der regulatorischen Arbeit könnte ein Gegengewicht bilden.

8 Literatur

Abid, A.; Farooqi, M.; Zou, J. (2021): Persistent Anti-Muslim Bias in Large Language Models. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21). Association for Computing Machinery, New York, NY, S. 298–306

Abraham, T.M. (2023): Tweet »How does GPT-4 do in the medical domain? (...)« vom 24.3.2023, https://twitter.com/iScienceLuvr/status/1639154976097460226 (19.4.2023)

ACM (Association for Computing Machinery) (2021): ACM TechBrief: Computing and Climate Change. ACM Technology Policy Council, New York, NY (DOI: 10.1145/3483410)

Ahmed, N.; Wahed, M.; Thompson, N.C. (2023): The growing influence of industry in AI research. In: Science 379 (6635), S. 884–886

AKI (Akademie für Künstliche Intelligenz AKI GmbH) (2023): Große KI-Modelle für Deutschland. Machbarkeitsstudie. Berlin

Alba, D. (2022): OpenAI Chatbot Spits Out Biased Musings, Despite Guardrails. 8.12.2022, www.bloomberg.com/news/newsletters/2022-12-08/chatgpt-open-ai-s-chatbot-is-spitting-out-biased-sexist-results (19.4.2023)

Albergotti, R.; Matsakis, L. (2023): OpenAI has hired an army of contractors to make basic coding obsolete. 27.1.2023, www.semafor.com/article/01/27/2023/openai-has-hired-an-army-of-contractors-to-make-basic-coding-obsolete (19.4.2023)

Albert, A. (2023): Tweet vom 16.3.2023, https://twitter.com/alexalbert__/status/1636500543337299969 (19.4.2023)

Allyn, B. (2023): A robot was scheduled to argue in court, then came the jail threats. 25.1.2023, www.npr.org/2023/01/25/1151435033/a-robot-was-scheduled-to-argue-in-court-then-came-the-jail-threats (19.4.2023)

Alouani, N. (2023): Why You're the Biggest Loser in the AI Wars. 19.2.2023, https://nabilalouani.substack.com/p/why-youre-the-biggest-loser-in-the (19.4.2023)

Alston, E. (2023): New! Try Zapier's ChatGPT plugin. 23.3.2023, https://zapier.com/blog/announcing-zapier-chatgpt-plugin/ (19.4.2023)

Altman, S. (2022): Tweet vom 11.12.2022, https://twitter.com/sama/status/1601731295792414720 (19.4.2023)

Ananthaswamy, A. (2023): In AI, is bigger better? In: Nature 615, 9.3.2023, S. 202–205

Anthony, L.F.W.; Kanding, B.; Selvan, R. (2020): Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. Paper presented at ICML Workshop on »Challenges in Deploying and monitoring Machine Learning Systems«, 17.7.2020, arXiv:2007.03051

Atleson, M. (2023): Chatbots, deepfakes, and voice clones: AI deception for sale. 20.3.2023, www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale (19.4.2023)

Autor, D.; Mindell, D.; Reynolds, E. (2020): The Work of the Future: Building Better Jobs in an Age of Intelligent Machines. Cambridge, MA

Azaria. A. (2022): ChatGPT Usage and Limitations. 27.12.2022, https://hal.science/hal-03913837 (19.4.2023)

Bain & Company (2023): Bain & Company announces services alliance with OpenAI to help enterprise clients identify and realize the full potential and maximum value of AI. Pressemitteilung vom 21.2.2023, www.bain.com/vector-digital/partnerships-alliance-ecosystem/openai-alliance/ (19.4.2023)
Baron, N.S. (2023): How ChatGPT robs students of motivation to write and think for themselves. 19.1.2023, https://theconversation.com/how-chatgpt-robs-students-of-motivation-to-write-and-think-for-themselves-197875 (19.4.2023)
Bastian, M. (2023): CNET investigation shows lots of flaws in AI-written articles. 25.1.2023, https://the-decoder.com/cnet-investigation-shows-lots-of-flaws-in-ai-written-articles/ (19.4.2023)
Baumann, A.-L.; Egenberger, V.; Supik, L. (2018): Erhebung von Antidiskriminierungsdaten in repräsentativen Wiederholungsbefragungen. Bestandsaufnahme und Entwicklungsmöglichkeiten. Antidiskriminierungsstelle des Bundes, Berlin
Bayer, L. (2023): Introducing Q-Chat, the world's first AI tutor built with OpenAI's ChatGPT. 28.2.2023, https://quizlet.com/blog/meet-q-chat (19.4.2023)
Becker, K. (2023): Prof. Urs Gasser zur Forderung eines Moratoriums für die KI-Entwicklung: »Eine Pause beim Training von Künstlicher Intelligenz hilft nicht«. Pressimitteilung der Technischen Universität München vom 3.4.2023, www.tum.de/aktuelles/alle-meldungen/pressemitteilungen/details/eine-pause-beim-training-von-kuenstlicher-intelligenz-hilft-nicht (19.4.2023)
Beer, K. (2023): Sofatutor-Gründer zu ChatGPT: Es ist zu früh, um über Schwächen der KI zu reden. 11.2.2023, www.heise.de/hintergrund/Sofatutor-Gruender-zu-ChatGPT-Es-ist-zu-frueh-um-ueber-Schwaechen-der-KI-zu-reden-7491429.html (19.4.2023)
Beetz, C. (Produzent); Block, H,; Riesewieck, M. (Regie) (2018): The Cleaners. Film, Deutschland
Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. (2021): On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 2021, S. 610–623
Berdejo-Espinola, V.; Amano, T. (2023): AI tools can improve equity in science. In: Science 379(6636), S. 991
Bergstrom, C. (2023): Your chatbot is not »hallucinating«. 17.2.2023, https://post.news/article/2Lr2DCy9lQz0pbzrVwrtgBD6I81 (19.4.2023)
Berins, L. (2023): Judith Simon über Chatbots: »ChatGPT versteht nicht, es simuliert nur Sprache«. 31.1.2023, www.fr.de/kultur/gesellschaft/judith-simon-ueber-chatbots-chatgpt-versteht-nicht-es-simuliert-nur-sprache-92060094.html (19.4.2023)
Besner, A.; Gerstner, M.; Strasser, A. (2023): Erste Einschätzungen zum Umgang mit ChatGPT in Fernprüfungen an bayerischen Univeristäten. Bayerisches Kompetenzzentrum für Fernprüfungen, Technische Universität München, München
Biddle, S. (2022): The Internet's New Favorite AI Proposes Torturing Iranians and Surveilling Mosques. 8.12.2022, https://theintercept.com/2022/12/08/openai-chatgpt-ai-bias-ethics/ (19.4.2023)
BigScience Workshop (Le Scao, T.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A.S.; Yvon, F.; Gallé, M.; Tow, J. et al.) (2023): BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. arXiv:2211.05100

Bjork, C. (2023): 9 ChatGPT threatens language diversity. More needs to be done to protect our differences in the age of AI. 9.2.2023, https://theconversation.com/chatgpt-threatens-language-diversity-more-needs-to-be-done-to-protect-our-differences-in-the-age-of-ai-198878 (19.4.2023)
Blume, B. (2023): Das Ende vom Lernen wie wir es kennen. 20.1.2023, https://deutsches-schulportal.de/kolumnen/chatgpt-das-ende-vom-lernen-wie-wir-es-kennen/ (19.4.2023)
Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; Kalai,A. (2016): Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16). Red Hook, NY, USA, S. 4356–4364
Bommasani, R., Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E. et al. (2021): On the opportunities and risks of foundation models. arXiv:2108.07258
Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; van Den Driessche, G.B.; Lespiau, J.; Damoc, B.; Clark, A.; De Las Casas, D. et al. (2022): Improving Language Models by Retrieving from Trillions of Tokens. In: Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research 162, S. 2206–2240
Borji, A. (2023): A Categorical Archive of ChatGPT Failures. arXiv:2302.03494
Bornstein, M.; Appenzeller, G.; Casado, M. (2023): Who owns the generative AI platform? 19.1.2023, https://a16z.com/2023/01/19/who-owns-the-generative-ai-platform/ (19.4.2023)
Brandstätter, P. (2023): Hausaufgaben aus der Maschine. 18.3.2023, https://taz.de/ChatGPT-loest-Bildungskrise-aus/!5920652/ (19.4.2023)
Branwen, G. (2023): GPT-3 Creative Fiction. 11.3.2023, https://gwern.net/gpt-3 (19.4.2023)
Briegleb, V. (2023): ChatGPT-Hype: Google will sich im KI-Wettrennen nicht geschlagen geben. 20.1.2023, www.heise.de/news/ChatGPT-Hype-Google-will-sich-im-KI-Wettrennen-nicht-geschlagen-geben-7466242.html (19.4.2023)
Brittain, B. (2023): Lawsuits accuse AI content creators of misusing copyrighted work. 17.1.2023, www.reuters.com/legal/transactional/lawsuits-accuse-ai-content-creators-misusing-copyrighted-work-2023-01-17/ (19.4.2023)
Brockman, G.; Sutskever, I. (2015): Introducing OpenAI. 11.12.2015, https://openai.com/blog/introducing-openai (24.3.2023)
Brown University (2023): Brown scholars put their heads together to decode the neuro science behind ChatGPT. Pressemitteilung vom 9.2.2023, www.brown.edu/news/2023-02-09/neuroscience-chatbot (19.4.2023)
Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplany, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A. et al. (2020): Language models are few-shot learners. arXiv:2005.14165v4
Brühl, J. (2023): »Chat-GPT wird benutzt, um Bullshit zu automatisieren«. In: Süddeutsche Zeitung vom 21.3.2023, S. 15
Bruell, A. (2023): Sports Illustrated Publisher Taps AI to Generate Articles, Story Ideas. 3.2.2023, www.wsj.com/articles/sports-illustrated-publisher-taps-ai-to-generate-articles-story-ideas-11675428443 (19.4.2023)
Brynjolfsson, E. (2022): The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence. In: Daedalus 151(2), S. 272–287

Buchanan, B.; Lohn, A.; Musser, M.; Sedova, K. (2021): Truth, Lies, and Automation. How language models could change disinformation. Center for Security and Emerging Technology, Washington, DC

Bünte, O.; dpa (2023): ChatGPT: Rekord-Wachstum und Abo-Modell in den USA. 2.2.2023, www.heise.de/news/ChatGPT-startet-Abo-Modell-in-den-USA-7480052.html (19.4.2023)

Bundesregierung (2018): Strategie Künstliche Intelligenz der Bundesregierung. Berlin

Bundesregierung (2021): Datenstrategie der Bundesregierung. Kabinettsfassung, 27. Januar 2021. Berlin

Bundesregierung (2022): Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten Anke Domscheit-Berg, Dr. Petra Sitte, Nicole Gohlke, weiterer Abgeordneter und der Fraktion DIE LINKE – Drucksache 20/317 – Künstliche Intelligenz im Geschäftsbereich der Bundesregierung. Deutscher Bundestag, Drucksache 20/430, Berlin

Bundesregierung (2023a): Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten Barbara Lenk, Eugen Schmidt, Edgar Naujok, weiterer Abgeordneter und der Fraktion der AfD – Drucksache 20/5465 – Zum Textgenerator ChatGPT des Unternehmens Open AI. Deutscher Bundestag, Drucksache 20/6044, Berlin

Bundesregierung (2023b): Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten René Springer, Jürgen Pohl, Gerrit Huy, weiterer Abgeordneter und der Fraktion der AfD – Drucksache 20/5656 – ChatGPT und die Zukunft der Arbeit. Deutscher Bundestag, Drucksache 20/6062, Berlin

Burrell, J. (2023): It's time to challenge the narrative about ChatGPT and the future of journalism. 9.2.2023, www.poynter.org/ethics-trust/2023/opinion-chatgpt-will-not-replace-humans/ (19.4.2023)

CAIS (Center for Advanced Internet Studies) (2023): Pressemitteilung: ChatGPT, wie viele Menschen kennen Dich bereits? 1.2.2023, www.cais-research.de/news/chatgpt-wie-viele-menschen-kennen-dich-bereits/ (19.4.2023)

Campolo, A.; Crawford, K. (2020): Enchanted Determinism: Power without Responsibility in Artificial Intelligence. In: Engaging Science, Technology, and Society 6, S. 1–19

Caprarella, J. (2023): Künstliche Intelligenz in der Politik: Rumäniens Ministerpräsident stellt KI als Berater ein. 2.3.2023, https://de.nachrichten.yahoo.com/kunstliche-intelligenz-in-der-politik-rumaniens-ministerprasident-stellt-ki-als-berater-ein-122954805.html (19.4.2023)

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; Oprea, A.; Raffel, C. (2021): Extracting Training Data from Large Language Models. arXiv:2012.07805

Ceney, A.; Tolond, S.; Glowinski, A.; Marks, B.; Swift, S.; Palser, T. (2021): Accuracy of online symptom checkers and the potential impact on service utilisation. In: PLoS ONE 16(7), e0254088 (DOI: 10.1371/journal.pone.0254088)

Che, C.; Liu, J. (2023): China's Answer to ChatGPT Gets an Artificial Debut and Disappoints. 16.3.2023, www.nytimes.com/2023/03/16/world/asia/china-baidu-chatgpt-ernie.html (19.4.2023)

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; de Oliveira Pinto, H.P.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; Ray, A. et al. (2021): Evaluating large language models trained on code. arXiv:2107.03374

Choi, J.H.; Hickman, K.E.; Monahan, A.; Schwarcz, D.B. (2023): ChatGPT Goes to Law School. Minnesota Legal Studies Research Paper No. 23-03 (DOI: 10.2139/ssrn.4335905)
Chomsky, N. Roberts, I.; Watumull, J. (2023): Noam Chomsky: The False Promise of ChatGPT. 8.3.2023, www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html (19.4.2023)
Christian, J. (2023a): CNET's AI Journalist Appears to Have Committed Extensive Plagiarism. 23.1.2023, https://futurism.com/cnet-ai-plagiarism (19.4.2023)
Christian, J. (2023b): Magazine Publishes Serious Errors in First AI-Generated Health Article. 9.2.2023, https://futurism.com/neoscope/magazine-mens-journal-errors-ai-health-article (19.4.2023)
Ciesielski, R.; Barthel, J. (2023): Bachelorarbeit in drei Tagen mit ChatGPT? 31.3.2023, www.br.de/nachrichten/wissen/bachelorarbeit-in-drei-tagen-mit-chatgpt-kuenstliche-intelligenz,TZo8lwF (19.4.2023)
Clarke, N. (2023): A Concerning Trend. 15.2.2023, http://neil-clarke.com/a-concerning-trend/ (19.4.2023)
Cowls, J.; Tsamados, A.; Taddeo, M.; Floridi, L. (2023): The AI gambit: leveraging artificial intelligence to combat climate change – opportunities, challenges, and recommendations. In: AI & Society 38(1), S. 283–307
Crawford, K.; Calo, R. (2016): There is a blind spot in AI research. In: Nature 538, 20.10.2016, S. 311–313
Dalla-Torre, H.; Gonzalez, L.; Mendoza Revilla, J.; Lopez Carranza, N.; Grzywaczewski, A.H.; Oteri, F.; Dallago, C.; Trop, E.; Sirelkhatim, H.; Richard, G.; Skwark, M. et al. (2023): The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. bioRxiv 2023.01.11.523679
Dang, A. (2023): KI-Forscher: »Ich sehe die Euphorie auch skeptisch«. 10.2.2023, www.derstandard.at/story/2000143432882/ki-forscher-ich-sehe-die-euphorie-auch-skeptisch (19.4.2023)
Darcy, A. (2023): Why Generative AI Is Not Yet Ready for Mental Healthcare. 1.3.2023, https://woebothealth.com/why-generative-ai-is-not-yet-ready-for-mental-healthcare/ (19.4.2023)
Dastin, J.; Nellis, S. (2023): For tech giants, AI like Bing and Bard poses billion-dollar search problem. 23.2.2023, www.reuters.com/technology/tech-giants-ai-like-bing-bard-poses-billion-dollar-search-problem-2023-02-22/ (19.4.2023)
Davis, P. (2023): Did ChatGPT Just Lie To Me? 13.1.2023, https://scholarlykitchen.sspnet.org/2023/01/13/did-chatgpt-just-lie-to-me/ (19.4.2023)
Daws, R. (2020): Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves. 28.10.2020, www.artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/ (19.4.2023)
Dax, P. (2023): Warum ChatGPT so schnell kein Job-Killer ist. 9.3.2023, https://futurezone.at/digital-life/chatgpt-arbeitsmarkt-jobabbau-experten-stefan-strauss-clemens-heitzinger-josef-trappel/402355518 (19.4.2023)
Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. (2009): ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, S. 248–255
Dennis, A.; Kim, A.; Rahimi, M.; Ayabakan, S. (2020): User reactions to COVID-19 screening chatbots from reputable providers. In: Journal of the American Medical Informatics Association 27(11), S. 1727–1731

Deutscher Bundestag (2023a): Schriftliche Fragen mit den in der Woche vom 13. Februar 2023 eingegangenen Antworten der Bundesregierung. Deutscher Bundestag, Drucksache 20/5694, Berlin
Deutscher Bundestag (2023b): Stenografischer Bericht, 86. Sitzung vom 10. Februar 2023. Deutscher Bundestag, Plenarprotokoll 20/86, Berlin
Deutscher Ethikrat (2023): Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz. Stellungnahme. Berlin
Dev, S.; Sheng, E.; Zhao, J.; Amstutz, A.; Sun, J.; Hou, Y.; Sanseverino, M.; Kim, J.; Nishi, A.; Peng, N.; Chang, K.-W. (2022): On Measures of Biases and Harms in NLP. In: Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, S. 246–267
van Deyzen, B. (2023): Education experts discuss ChatGPT: »An extra classmate has joined the class«. 19.1.2023, https://communities.surf.nl/en/ai-in-education/article/education-experts-discuss-chatgpt-an-extra-classmate-has-joined-the (19.4.2023)
van Dis, E.A.M.; Bollen, J.; van Rooij, R.; Zuidema, W.; Bockting, C.L. (2023): ChatGPT: five priorities for research. In: Nature Bd. 614, 9.2.2023, S. 225
Doctorow, C. (2020): Our Neophobic, Conservative AI Overlords Want Everything to Stay the Same. 1.1.2020, https://blog.lareviewofbooks.org/provocations/neophobic-conservative-ai-overlords-want-everything-stay/ (19.4.2023)
Dolderer, M. (2022): AI transforming higher education? How a chatbot might bring the change we've all been waiting for. 19.12.2022, https://hochschulforumdigitalisierung.de/de/blog/AI-higher-education-gbt-3 (19.4.2023)
Driess, D.; Xia, F.; Sajjadi, M.S.M.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; Huang, W. et al. (2023): PaLM-E: An Embodied Multimodal Language Model. arXiv:2303.03378
Du, N.; Huang, Y.; Dai, A.M.; Tong, S.; Lepikhin, D.; Xu, Y.; Krikun, M.; Zhou, Y.; Wei Yu, A.; Firat, O.; Zoph, B. et al. (2022): GLaM: Efficient Scaling of Language Models with Mixture-of-Experts. In: Proceedings of the 39th International Conference on Machine Learning, PMLR 162, S. 5547–5569
Dunhill, J. (2023): GPT-4 Hires And Manipulates Human Into Passing CAPTCHA Test. 16.3.2023, www.iflscience.com/gpt-4-hires-and-manipulates-human-into-passing-captcha-test-68016 (19.4.2023)
Eloundou, T.; Manning, S.; Mishkin, P.; Rock, D. (2023): GPTs are GPTs– An Early Look at the Labor Market Impact Potential of Large Language Models. arXiv:2303.10130
Else, H. (2023): Abstracts written by ChatGPT fool scientists. In: Nature Bd. 613, 19.1.2023, S. 423
Emmerich, N. (2023): ChatGPT in der Bildung: »Hausaufgaben sind tot«. 25.1.2023, www.gew.de/aktuelles/detailseite/hausaufgaben-sind-tot (19.4.2023)
Engelmann, J.; Puntschuh, M. (2020): KI im Behördeneinsatz – Erfahrungen und Empfehlungen. Berlin
Europäische Kommission (2021): Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union. Brüssel, 21.4.2021, COM(2021) 206 final, https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0019.02/DOC_1&format=PDF (24.3.2023)
Europäische Kommission (2022): Ethische Leitlinien für Lehrkräfte über die Nutzung von KI und Daten für Lehr- und Lernzwecke. Luxemburg

Fecher, B.; Schulz, W. (2023): Künstliche Intelligenz in der Forschung: Wie ChatGPT das wissenschaftliche Arbeiten verändern wird. 15.2.2023, www.tagesspiegel.de/wissen/chatgpt-sind-maschinen-die-besseren-forscher-9344372.html (19.4.2023)

Felton, J. (2023): Google And Bing's AI Chatbots Appear To Be Citing Each Other's Lies. 23.3.2023, www.iflscience.com/google-and-bings-ai-chatbots-appear-to-be-citing-each-others-lies-68116 (19.4.2023)

Ferrer, X.; van Nuenen, T.; Such, J.M.; Criado, N. (2021): Discovering and Categorising Language Biases in Reddit. In: Proceedings of the International AAAI Conference on Web and Social Media 15(1), S. 140–151

Finnie-Ansley, J.; Denny, P.; Becker, B.A.; Luxton-Reilly, A.; Prather, J. (2022): The Robots Are Coming: Exploring the Implications of OpenAI Codex on Introductory Programming. In: ACE '22: Proceedings of the 24th Australasian Computing Education ConferenceFebruary 2022, S. 10–19 (DOI: 10.1145/3511861.3511863)

Fischer, G. (2023): »Künstliche Intelligenz« à la GPT3: Die große Remix-Maschine. 18.2.2023, https://irights.info/artikel/kuenstliche-intelligenz-a-la-gpt3-die-grosse-remix-maschine/31752 (19.4.2023)

Fitzpatrick K.K.; Darcy, A.; Vierhile, M. (2017): Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. In: JMIR Mental Health 4(2), e19 (DOI: 10.2196/mental.7785)

Floridi, L.; Chiriatti, M. (2020): GPT-3: Its Nature, Scope, Limits, and Consequences. In: Minds & Machines 30, S. 681–694

Förtsch, M. (2023): Alpaca: Die Universität Stanford hat einen ChatGPT-Konkurrenten entwickelt, der auf Billig-Computern läuft. 20.3.2023, https://1e9.community/t/alpaca-die-universitaet-stanford-hat-einen-chatgpt-konkurrenten-entwickelt-der-auf-billig-computern-laeuft/18930 (19.4.2023)

Frieder, S.; Pinchetti, L.; Griffiths, R.-R.; Salvatori, T.; Lukasiewicz, T.; Petersen, P.C.; Chevalier, A.; Berner, J. (2023): Mathematical Capabilities of ChatGPT. arXiv:2301.13867

Fulterer, R. (2023): Interview: »Es ist absurd, bei Chat-GPT von künstlicher Intelligenz zu sprechen«. 25.2.2023, www.nzz.ch/technologie/es-ist-absurd-bei-chat-gpt-von-kuenstlicher-intelligenz-zu-sprechen-ld.1726924 (19.4.2023)

Funk, J. (2022): AI and Economic Productivity: Expect Evolution, Not Revolution. 5.12.2022, https://spectrum.ieee.org/ai-and-economic-productivity-expect-evolution-not-revolution (19.4.2023)

Gal, U. (2023): ChatGPT is a data privacy nightmare. If you've ever posted online, you ought to be concerned. 8.2.2023, https://theconversation.com/chatgpt-is-a-data-privacy-nightmare-if-youve-ever-posted-online-you-ought-to-be-concerned-199283 (19.4.2023)

Ganguli, D.; Hernandez, D.; Lovitt, L.; DasSarma, N.; Henighan, T.; Jones, A.; Joseph, N.; Kernion, J.; Mann, B.; Askell, A.; Bai, Y. et al. (2022): Predictability and Surprise in Large Generative Models. In: FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, June 2022, S. 1747–1764

Gao, C.A.; Howard, F.M.; Markov, N.S.; Dyer, E.C.; Ramesh, S.; Luo, Y.; Pearson, A.T. (2022): Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. bioRxiv (DOI: 10.1101/2022.12.23.521610)

Gawlik, P.; Schiffer, C. (2023): ChatGPT - Schafft die KI das bayerische Abitur? 12.2.2023, www.br.de/nachrichten/netzwelt/chatgpt-schafft-die-ki-das-bayerische-abitur,TVBjrXE (19.4.2023)
Gershgorn, D. (2017): The data that transformed AI research – and possibly the world. 26.7.2017, https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world (19.4.2023)
Geuter, J. (2023): Bullshit, der (e)skaliert. 16.3.2023, www.golem.de/news/chat gpt-bard-und-co-bullshit-der-e-skaliert-2303-172677-3.html (19.4.2023)
Gibney, E. (2022): Open-source language AI challenges big tech's models. In: Nature 606, 30.6.2022, S. 850–851
Gimpel, H.; Hall, K.; Decker, S.; Eymann, T.; Lämmermann, L.; Mädche, A.; Röglinger, M.; Ruiner, C.; Schoch, M.; Schoop, M.; Urbach, N.; Vandirk, S. (2023): Unlocking the Power of Generative AI Models and Systems such as GPT-4 and ChatGPT for Higher Education: A Guide for Students and Lecturers. Hohenheim
Glaese, A.; McAleese, N.; Trebacz, M.; Aslanides, J.; Firoiu, V.; Ewalds, T.; Rauh, M.; Weidinger, L.; Chadwick, M.; Thacker, P.; Campbell-Gillingham, L. et al. (2022): Improving alignment of dialogue agents via targeted human judgements. arXiv:2209.14375
Goldstein, J.A.; Sastry, G.; Musser, M.; DiResta, R.; Gentzel, M.; Sedova, K. (2023): Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. 11.1.2023, https://cdn.openai.com/papers/forecasting-misuse.pdf (19.4.2023)
Golumbia, D. (2022): ChatGPT should not exit. 14.12.2022, https://davidgolumbia.medium.com/chatgpt-should-not-exist-aab0867abace (19.4.2023)
Google Search Central (2023): Leitfaden der Google Suche zu KI-generierten Inhalten. 8.2.2023, https://developers.google.com/search/blog/2023/02/google-search-and-ai-content (19.4.2023)
Gozalo-Brizuela, R.; Garrido-Merchan, E.C. (2023): ChatGPT is not all you need. A State of the Art Review of large Generative AI models. arXiv:2301.04655v1
Grävemeyer, A. (2022): Wandlungsfähige Schreib-KI. In: c't Heft 9/2022, S. 60-63
Greis, F. (2023): Google startet Experimentierphase von Chatbot Bard. 21.3.2023, www.golem.de/news/konkurrenz-zu-chatgpt-google-startet-experimentier phase-von-chatbot-bard-2303-172808.html (19.4.2023)
Greshake, K.; Abdelnabi, S.; Mishra, S.; Endres, C.; Holz, T.; Fritz, M. (2023): More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models. arXiv:2302.12173
Grimm, K. (2023): ChatGPT & Co. – Nutzung von künstlicher Intelligenz auf dem freien Projektmarkt. 2.3.2023, www.freelancermap.de/blog/kuenstliche-intelli genz-umfrage/ (19.4.2023)
Gross, G.; dpa (2023): Texte aus dem Generator. Wie gehen die Unis mit ChatGPT um? In: Der Tagesspiegel vom 1.2.2023, S. B27
Gutiérrez, J.D. (2023): ChatGPT in Colombian Courts. 23.2.2023, https://verfassungsblog.de/colombian-chatgpt (19.4.2023)
Hahn, S. (2023a): GPT-4: »In einer Welt rasender KI-Entwicklung« – Fakten auf einen Blick. 16.3.2023, www.heise.de/news/GPT-4-In-einer-Welt-rasender-KI-Entwicklung-Fakten-auf-einen-Blick-7546721.html (19.4.2023)

Hahn, S. (2023b): Luminous schließt Europas KI-Lücke: Aleph Alpha auf Augenhöhe mit US-Anbietern. 21.2.2023, www.heise.de/news/Luminous-schliesst-Europas-KI-Luecke-Aleph-Alpha-auf-Augenhoehe-mit-US-Anbietern-7521254.html (19.4.2023)
Hahn, S. (2023c): OpenAI: US-Verbraucherschutzbehörde soll ermitteln, Italien sperrt ChatGPT. 31.3.2023, www.heise.de/news/OpenAI-US-Verbraucherschutzbehoerde-soll-ermitteln-Italien-sperrt-ChatGPT-8328351.html (19.4.2023)
Hahn, S. (2023d): OpenAI stellt GPT-4 vor: Sprachmodell versteht jetzt auch Bilder. 14.3.2023, www.heise.de/news/OpenAI-stellt-GPT-4-vor-Sprachmodell-versteht-jetzt-auch-Bilder-7545722.html (19.4.2023)
Hanfeld, M. (2023): Streit um Leistungsschutzrecht – Es sollen nicht nur Promille sein. 23.7.2022, www.faz.net/aktuell/feuilleton/medien/google-und-corint-media-rufen-im-streit-um-presseleistungsschutzrecht-schiedsstelle-an-18192954.html (19.4.2023)
Hardman, P. (2023): Introducing: ChatGPT Edu-Mega-Prompts. 26.1.2023, https://drphilippahardman.substack.com/p/introducing-chatgpt-edu-mega-prompts (19.4.2023)
Hartmann, J.; Schwenzow, J.; Witte, M. (2023): The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orien tation. arXiv:2301.01768
Hartnett, K. (2022): Interview: Die Zukunft der KI. In: Spektrum Highlights 1/22: Künstliche Intelligenz – Wie man Robotern das Denken beibringt. Heidelberg, S. 45–47
Hashimoto, T.; Srivastava, M.; Namkoong, H.; Liang, P. (2018): Fairness Without Demographics in Repeated Loss Minimization. In: Proceedings of the 35th International Conference on Machine Learning, PMLR 80, S. 1929–1938
Haven, J. (2022): ChatGPT and the future of trust. 19.12.2022, www.niemanlab.org/2022/12/chatgpt-and-the-future-of-trust/ (19.4.2023)
Haverkamp, H. (2022): Ein Lehrer lässt KI bei Klassenarbeiten zu – das hat er dabei gelernt. 30.10.2022, https://the-decoder.de/ein-lehrer-laesst-ki-bei-klassenarbeiten-zu-das-hat-er-dabei-gelernt/ (19.4.2023)
Heaven, D. (2019): Deep trouble for deep learning. In: Nature 574, 10.10.2019, S. 163–166
Heaven, W.D. (2022): Why Meta's latest large language model survived only three days online. 18.11.2022, www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/ (19.4.2023)
Heaven, W.D. (2023a): Google just launched Bard, its answer to ChatGPT – and it wants you to make it better, 21.3.2023, www.technologyreview.com/2023/03/21/1070111/google-bard-chatgpt-openai-microsoft-bing-search/ (19.4.2023)
Heaven, W.D. (2023b): The inside story of how ChatGPT was built from the people who made it. 3.3.2023, www.technologyreview.com/2023/03/03/1069311/inside-story-oral-history-how-chatgpt-built-openai/ (19.4.2023)
Heesen, J.; Bieber, D.; Lauber-Rönsberg, A.; Neuberger, C.; Elmer, C.; Hühnert, T. (2023): Künstliche Intelligenz im Journalismus. Whitepaper aus der Plattform Lernende Systeme, München (DOI: 10.48669/pls_2023-1)
Heidt, A. (2023): »Arms race with automation«: professors fret about AI-generated coursework. Nature Technology Feature, 24.1.2023, www.nature.com/articles/d41586-023-00204-z (19.4.2023)

Heikkilä, M. (2023a): Could ChatGPT do my job? 31.1.2023, www.technology review.com/2023/01/31/1067436/could-chatgpt-do-my-job/ (19.4.2023)
Heikkilä, M. (2023b): Wie Wasserzeichen Texte von KI-Chatbots sichtbar machen könnten. 2.2.2023, www.heise.de/hintergrund/Wie-Wasserzeichen-Texte-von-KI-Chatbots-sichtbar-machen-koennten-7477158.html (19.4.2023)
Hendler, J. (2023): Understanding the limits of AI coding. In: Science 379(6632), S. 548
Hern, A. (2017): Facebook translates »good morning« into »attack them«, leading to arrest. 24.10.2017, www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest (19.4.2023)
Hern, A. (2023): Sci-fi publisher Clarkesworld halts pitches amid deluge of AI-generated stories. 21.2.2023, www.theguardian.com/technology/2023/feb/21/sci-fi-publisher-clarkesworld-halts-pitches-amid-deluge-of-ai-generated-stories (19.4.2023)
Hoeren, T. (2023): Rechtsgutachten zum Umgang mit KI-Software im Hochschulkontext. In: Salden, P.; Leschke, J. (Hg.): Didaktische und rechtliche Perspektiven auf KI-gestütztes Schreiben in der Hochschulbildung. Bochum, S. 22–40
Honegger, B.D. (2023a): ChatGPT & Schule. 4.2.2023, https://mia.phsz.ch/LLM (19.4.2023)
Honegger, B.D. (2023b): Digitaler Schereneffekt. 2.1.2023, https://beat.doebe.li/bibliothek/w03389.html (19.4.2023)
Hooker, S. (2020): The Hardware Lottery. In: Communications of the ACM 64(12), S. 58–65
Horizont Online; dpa (2023): Kantar-Umfrage: Ein Viertel der Deutschen hat bereits ein KI-Tool wie ChatGPT genutzt. 4.3.2023, www.horizont.net/tech/nachrichten/kantar-umfrage-ein-viertel-der-deutschen-hat-bereits-ein-ki-tool-wie-chatgpt-genutzt-210290 (19.4.2023)
Hsu, T.; Thompson, S.A. (2023): Disinformation Researchers Raise Alarms About A.I. Chatbots. 8.2.2023, www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html (19.4.2023)
Hughes, A. (2023): ChatGPT: Everything you need to know about OpenAI's GPT-4 tool. 16.3.2023, www.sciencefocus.com/future-technology/gpt-3/ (19.4.2023)
Hutson, M. (2021): The language machines. In: Nature 591, 4.3.2021, S. 22–25
Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. (2022): Survey of Hallucination in Natural Language Generation. In: ACM Computing Surveys 55(12), 248 (DOI: 10.1145/3571730)
Kahn, J. (2023): The inside story of ChatGPT: How OpenAI founder Sam Altman built the world's hottest technology with billions from Microsoft. 25.1.2023, https://fortune.com/longform/chatgpt-openai-sam-altman-microsoft/ (19.4.2023)
Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T.B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D. (2021): Scaling Laws for Neural Language Models. arXiv:2001.08361
Karpf, D. (2022): Money Will Kill ChatGPT's Magic. 21.12.2022, www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-chatbots-openai-cost-regulations/672539/ (19.4.2023)
Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; Krusche, S. et al. (2023): ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. Position Paper. München

Kastl, B. (2023): Was war noch mal das Problem? 12.2.2023, https://netzpolitik.org/2023/degitalisierung-was-war-nochmal-das-problem/ (19.4.2023)
Khullar, D. (2023): Can A.I. Treat Mental Illness? 27.2.2023, www.newyorker.com/magazine/2023/03/06/can-ai-treat-mental-illness (19.4.2023)
Kim, T.J.; Stenert, K. (2023): ChatGPT-Recht - rechtliche Betrachtung von ChatGPT. 4.3.2023, www.sbs-legal.de/blog/chatgpt-recht-was-chatgpt-rechtlich-bedeuten-kann (19.4.2023)
Kirchenbauer, J.; Geiping, J.; Wen, Y.; Katz, J.; Miers, I.; Goldstein, T. (2023): A Watermark for Large Language Models. arXiv:2301.10226
Klinge, J.-M. (2022): Ein Schreib-Workshop durch eine Künstliche Intelligenz. 29.12.2022, https://halbtagsblog.de/2022/12/28/ein-schreib-workshop-durch-eine-kuenstliche-intelligenz/ (19.4.2023)
Köver, C. (2023): Wofür braucht OpenAI so viel Geld? 25.1.2023, https://netzpolitik.org/2023/10-milliarden-fuer-start-up-wofuer-braucht-openai-so-viel-geld/ (19.4.2023)
Kohne, A.; Kleinmanns, P.; Rolf, C.; Beck, M. (2020): Chatbots. Aufbau und Anwendungsmöglichkeiten von autonomen Sprachassistenten. Wiesbaden
Kompetenzplattform KI.NRW (2021): Moderne Sprachtechnologien. Konzepte, Anwendungen, Chancen. Sankt Augustin
Kreer, C. (2023): Das nächste große Ding für digitale Zugänglichkeit? 16.3.2023, https://netzpolitik.org/2023/gpt-4-das-naechste-grosse-ding-fuer-digitale-zugaenglichkeit/ (19.4.2023)
Kreutzer, T. (2021): Welche Regeln gelten für die Erzeugnisse Künstlicher Intelligenz? 25.2.2021, https://irights.info/artikel/welche-regeln-gelten-fuer-die-erzeugnisse-kuenstlicher-intelligenz/30724 (19.4.2023)
Kreye, A. (2023): Im Tal der Ahnungsvollen. 1.3.2023, www.sueddeutsche.de/kultur/ki-elon-musk-mark-rolston-argodesign-1.5760548 (19.4.2023)
Krischke, W. (2023): Trainingsrunden für Sprachroboter. In: Frankfurter Allgemeine Zeitung vom 25.1.2023, S. 4
Krizhevsky, A.; Sutskever, I.; Hinton, G.E. (2017): ImageNet classification with deep convolutional neural networks. In: Communications of the ACM 60(6), S. 84–90
Kroker, M. (2022): »Das Tool deckt auf, was im Wissenschaftsbetrieb falsch läuft«. 15.12.2022, www.wiwo.de/technologie/digitale-welt/chatgpt-das-tool-deckt-auf-was-im-wissenschaftsbetrieb-falsch-laeuft/28866804.html (19.4.2023)
Kühl, E. (2023): Ein Llama auf Abwegen. 6.3.2023, www.zeit.de/digital/internet/2023-03/llama-ki-chatbot-meta-leak/komplettansicht (19.4.2023)
Kullmann, S. (2023): Das Ende der Trefferlisten: Informationsinfrastrukturen und KI. 21.2.2023, https://dgi-info.de/das-ende-der-trefferlisten-informationsinfrastrukturen-und-ki/ (19.4.2023)
Kung, T.H. et al. (2023): Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. medRxiv 2022.12.19.22283643
Lacker, K. (2020): Giving GPT-3 a Turing Test. 6.7.2020, https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html (19.4.2023)
Landymore, F. (2023): CNET Is Quietly Publishing Entire Articles Generated By AI. 11.1.2023, https://futurism.com/the-byte/cnet-publishing-articles-by-ai (19.4.2023)
Langston, J. (2020): Microsoft announces new supercomputer, lays out vision for future AI work. 19.5.2020, https://news.microsoft.com/source/features/ai/openai-azure-supercomputer/ (19.4.2023)

Lehner, M. (2021): Sind Künstliche Intelligenzen die besseren Journalist:innen? 21.12.2021, https://medium.com/br-next/sind-k%C3%BCnstliche-intelligenzen-die-besseren-journalist-innen-53e6a4ede70d (19.4.2023)
Lennon, R. (2023): Thread »Anatomy of a ChatGPT Mega-Prompt« vom 16.1.2023, https://twitter.com/thatroblennon/status/1615104249192488980 (19.4.2023)
van Lente, H.; Splitters, C.; Peine, A. (2013): Comparing technological hype cycles: Towards a theory. In: Technological Forecasting & Social Change 80(8), S. 1615–1628
Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; Newman, B. et al. (2022): Holistic Evaluation of Language Models. arXiv:2211.09110
Lin, S.; Hilton, J.; Evans, O. (2022): TruthfulQA: Measuring How Models Mimic Human Falsehoods arXiv:2109.07958
Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A. et al. (2023): Evolutionary-scale prediction of atomic-level protein structure with a language model. In: Science 379(6637), S. 1123–1130
Lordick, N. (2023): KI-Tools werden den akademischen Betrieb nicht zum Einsturz bringen. 23.2.2023, https://news.rub.de/wissenschaft/2023-02-23-wissenschaftsdidaktik-ki-tools-werden-den-akademischen-betrieb-nicht-zum-einsturz-bringen (19.4.2023)
Luccioni, A.S.; Viguier, S.; Ligozat, A.-L. (2022): Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. arXiv:2211.02001
Lukpat, A. (2023): JPMorgan Restricts Employees From Using ChatGPT. 22.2.2023, www.wsj.com/articles/jpmorgan-restricts-employees-from-using-chatgpt-2da5dc34 (19.4.2023)
Mahowald, K.; Ivanova, A.A. (2022): Google's powerful AI spotlights a human cognitive glitch: Mistaking fluent speech for fluent thought. 24.6.2022, https://theconversation.com/googles-powerful-ai-spotlights-a-human-cognitive-glitch-mistaking-fluent-speech-for-fluent-thought-185099 (19.4.2023)
Manning, S. (2023): Einfach schreiben mit künstlicher Intelligenz. 18.1.2023, https://multisprech.org/2023/01/18/einfach-schreiben-mit-kuenstlicher-intelligenz/ (19.4.2023)
Marche, S. (2022):The College Essay Is Dead. 6.12.2022, www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-writing-college-student-essays/672371/ (19.4.2023)
Marcus, G. (2022): How come GPT can seem so brilliant one minute and so breathtakingly dumb the next? 1.12.2022, https://garymarcus.substack.com/p/how-come-gpt-can-seem-so-brilliant (19.4.2023)
Marcus, G. (2023): Why Are We Letting the AI Crisis Just Happen? 13.3.2023, www.theatlantic.com/technology/archive/2023/03/ai-chatbots-large-language-model-misinformation/673376/ (19.4.2023)
Marcus, G.; Davis, E. (2023a): How not to test GPT-3. 18.2.2023, https://garymarcus.substack.com/p/how-not-to-test-gpt-3 (19.4.2023)
Marcus, G.; Davis, E. (2023b): Large Language Models like ChatGPT say The Darnedest Things. 10.1.2023, https://garymarcus.substack.com/p/large-language-models-like-chatgpt (19.4.2023)

Markov, T.; Zhang, C.; Agarwal, S.; Eloundou, T.; Lee, T.; Adler, S.; Jiang, A.; Weng, L. (2022): A Holistic Approach to Undesired Content Detection in the Real World. arXiv:2208.03274v1

Markovski, Y. (o.D.): How your data is used to improve model performance. https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance (19.4.2023)

Marx, J.P.S. (2023): ChatGPT im Studium: Die Top 10 Befehle für effektives Lernen. 17.1.2023, https://shribe.de/chatgpt-studium/ (19.4.2023)

McCormick, P. (2021): The Model of Everything. 14.10.2021, www.notboring.co/p/the-model-of-everything (19.4.2023)

McCormick, P. (2023): Attention is All You Need. 27.3.2023, www.notboring.co/p/attention-is-all-you-need (19.4.2023)

McQuillan, D. (2018): Data Science as Machinic Neoplatonism. In: Philosophy & Technology 31, S. 253–272

Mehdi, Y. (2023): Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web. 7.2.2023, https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/ (19.4.2023)

Meineck, S. (2023): Sind Bild-Generatoren böse? 12.1.2023, https://netzpolitik.org/2023/aufschrei-unter-kuenstlerinnen-bild-generatoren-stable-diffusion-dall-e-2/ (19.4.2023)

Metz, C. (2023): Why Do A.I. Chatbots Tell Lies and Act Weird? Look in the Mirror. 26.2.2023, www.nytimes.com/2023/02/26/technology/ai-chatbot-information-truth.html (19.4.2023)

Metzler, D.; Tay, Y.; Bahri, D.; Najork, M. (2021): Rethinking search: making domain experts out of dilettantes. In: ACM SIGIR Forum 55(1), 13 (DOI: 10.1145/3476415.3476428)

Meyer, E.; Weßels, D. (2023): Natural Language Processing im akademischen Schreibprozess – mehr Motivation durch Inspiration? In: Schmohl, T.; Watanabe, A.; Schelling, K. (Hg.): Künstliche Intelligenz in der Hochschulbildung: Chancen und Grenzen des KI-gestützten Lernens und Lehrens. Bielefeld, S. 227–251

Mickle, T.; Metz, C.; Grant, N. (2023): The Chatbots Are Here, and the Internet Industry Is in a Tizzy. 8.3.2023, www.nytimes.com/2023/03/08/technology/chatbots-disrupt-internet-industry.html (19.4.2023)

Microsoft (2019): OpenAI forms exclusive computing partnership with Microsoft to build new Azure AI supercomputing technologies. Pressemitteilung, 22.7.2019, https://news.microsoft.com/2019/07/22/openai-forms-exclusive-computing-partnership-with-microsoft-to-build-new-azure-ai-supercomputing-technologies/ (19.4.2023)

Microsoft (2023): Microsoft and OpenAI extend partnership. Pressemitteilung, 23.1.2023, https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/ (19.4.2023)

Miller, S.; Gilbert, S.; Virani, V.; Wicks, P. (2020): Patients' Utilization and Perception of an Artificial Intelligence-Based Symptom Assessment and Advice Technology in a British Primary Care Waiting Room: Exploratory Pilot Study. In: JMIR Human Factors 7(3), e19713 (DOI: 10.2196/19713)

van Miltenburg, E.; Clinciu, M.; Dušek, O.; Gkatzia, D.; Inglis, S.; Leppänen, L.; Mahamood, S.; Schoch, S.; Thomson, C.; Wen, L. (2023): Barriers and enabling factors for error analysis in NLG research. In: Northern European Journal of Language Technology 9(1), 4529 (DOI: 10.3384/nejlt.2000-1533.2023.4529)
Mitchell, M.; Krakauer, D.C. (2023): The debate over understanding in AI's large language models. In: Proceedings of the National Academy of Sciences 120(13), e2215907120 (DOI: 10.1073/pnas.2215907120)
Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. (2019): Model Cards for Model Reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19), New York, NY, S. 220–229
Mohr, G.; Reinmann, G.; Blüthmann, N.; Lübcke, E.; Kreinsen, M. (2023): Übersicht zu ChatGPT im Kontext Hochschullehre. Hamburger Zentrum für Universitäres Lehren und Lernen (HUL), Hamburg
Mok, A.; Zinkula, J. (2023): ChatGPT may be coming for our jobs. Here are the 10 roles that AI is most likely to replace. 2.2.2023, www.businessinsider.com/chatgpt-jobs-at-risk-replacement-artificial-intelligence-ai-labor-trends-2023-02 (19.4.2023)
Mollick, E. (2023): A quick and sobering guide to cloning yourself. 10.2.2023, https://oneusefulthing.substack.com/p/a-quick-and-sobering-guide-to-cloning (19.4.2023)
Mollick, E.; Mollick, L. (2023): Using AI to Implement Effective Teaching Strategies in Classrooms: Five Strategies, Including Prompts (DOI: 10.2139/ssrn.4391243)
Monroe, D. (2023): ChatGPT: Super Rentier. 20.1.2023, https://monroelab.net/chatgpt-super-rentier (19.4.2023)
Morris, G.E.; Kennedy, C. (2017): Personal finance questions elicit slightly different answers in phone surveys than online. 4.8.2017, www.pewresearch.org/fact-tank/2017/08/04/personal-finance-questions-elicit-slightly-different-answers-in-phone-surveys-than-online/ (19.4.2023)
Narayanan, A.; Kapoor, S. (2022): ChatGPT is a bullshit generator. But it can still be amazingly useful. 6.12.2022, https://aisnakeoil.substack.com/p/chatgpt-is-a-bull shit-generator-but (19.4.2023)
Narayanan, A.; Kapoor, S. (2023): GPT-4 and professional benchmarks: the wrong answer to the wrong question. 20.3.2023, https://aisnakeoil.substack.com/p/gpt-4-and-professional-benchmarks (19.4.2023)
Nature Editorial (2023): Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. In: Nature 613, 26.1.2023, S. 612
Nayak, P. (2019): Understanding searches better than ever before. 25.10.2019, https://blog.google/products/search/search-language-understanding-bert/ (19.4.2023)
Neff, G.; Nagy, P. (2016): Talking to Bots: Symbiotic Agency and the Case of Tay. In: International Journal of Communication 10, S. 4915–4931
Neuralmagic (2023): SparseGPT: Remove 100 Billion Parameters for Free. Pressemitteilung vom 21.3.2023, https://neuralmagic.com/blog/sparsegpt-remove-100-bil lion-parameters-for-free/ (19.4.2023)
OpenAI (2020): OpenAI licenses GPT-3 technology to Microsoft. 22.9.2020, https://openai.com/blog/openai-licenses-gpt-3-technology-to-microsoft (19.4.2023)

OpenAI (2022a): Introducing ChatGPT. 30.11.2022, https://openai.com/blog/chatgpt (19.4.2023)
OpenAI (2022b): Snapshot of ChatGPT model behavior guidelines. https://cdn.openai.com/snapshot-of-chatgpt-model-behavior-guidelines.pdf (19.4.2023)
OpenAI (2023a): ChatGPT plugins. 23.3.2023, https://openai.com/blog/chatgpt-plugins (24.3.2023)
OpenAI (2023b): How should AI systems behave, and who should decide? 16.2.2023, https://openai.com/blog/how-should-ai-systems-behave (19.4.2023)
OpenAI (2023c): Introducing ChatGPT Plus. 1.2.2023, https://openai.com/blog/chatgpt-plus (19.4.2023)
OpenAI (2023d): March 20 ChatGPT outage: Here's what happened. 24.3.2023, https://openai.com/blog/march-20-chatgpt-outage (19.4.2023)
OpenAI (2023e): GPT-4 System Card. 23.2.2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf (19.4.2023)
OpenAI (2023f): GPT-4 Technical Report. arXiv:2303.08774
Ouyang, L.; Wu, J.; Jiang, J.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J. et al. (2022): Training language models to follow instructions with human feedback. arXiv:2203.02155
Ovadya, A. (2021): When we change the efficiency of knowledge operations, we change the shape of society. 2.4.2021, https://aviv.medium.com/when-we-change-the-efficiency-of-knowledge-operations-we-change-the-shape-of-society-d48ca870ff5b (19.4.2023)
Papakyriakopoulos, O.; Watkins, E.A.; Winecoff, A.; Jaźwińska, K.; Chattopadhyay, T. (2021): Qualitative Analysis for Human Centered AI. arXiv:2112.03784
Patel, D. (2023): The AI Brick Wall – A Practical Limit For Scaling Dense Transformer Models, and How GPT 4 Will Break Past It. 24.1.2023, www.semianalysis.com/p/the-ai-brick-wall-a-practical-limit (19.4.2023)
Patel, D.; Ahmad, A. (2023): The Inference Cost Of Search Disruption – Large Language Model Cost Analysis. 9.2.2023, www.semianalysis.com/p/the-inference-cost-of-search-disruption (19.4.2023)
Patterson, D.; Gonzalez, J.; Le, Q.; Liang, C.; Munguia, L.-M.; Rothchild, D.; So, D.; Texier, M.; Dean, J. (2021): Carbon Emissions and Large Neural Network Training. arXiv:2104.10350
Perez, E.; Ringer, S.; Lukošiūtė, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S.; Jones, A. et al. (2022): Discovering Language Model Behaviors with Model-Written Evaluations. arXiv:2212.09251
Perrigo, B. (2023): Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. 18.1.2023, https://time.com/6247678/openai-chatgpt-kenya-workers/ (19.4.2023)
Pilniok, A. (2021): Das Zeitalter der Künstlichen Intelligenz als Herausforderung für die Parlamente. In: Zeitschrift für Parlamentsfragen 52(1), S. 159–181
Poireault, F. (2023): #DataPrivacyWeek: ChatGPT's Data-Scraping Model Under Scrutiny From Privacy Experts. 27.1.2023, www.infosecurity-magazine.com/news-features/chatgpts-datascraping-scrutiny/ (19.4.2023)
Polomski, J. (2023): KI Texte erkennen – diese 12 Tools helfen dabei. 5.2.2023, https://jens.marketing/ki-texte-erkennen-tools/ (19.4.2023)
Porath, G. (2023): Einsatzmöglichkeiten von ChatGPT in HR. 30.1.2023, www.haufe.de/personal/hr-management/einsatzmoeglichkeiten-von-chatgpt-in-hr_80_586532.html (19.4.2023)

Puppis, M.; Schenk, M.; Hofstetter, B. (Hg.) (2017): Medien und Meinungsmacht. TA Swiss 65/2017, Zürich
Qin, C.; Zhang, A.; Zhang, Z.; Chen, J.; Yasunaga, M.; Yang, D. (2023): Is ChatGPT a General-Purpose Natural Language Processing Task Solver? arXiv:2302.06476
Radford, A.; Narasimhan, K.; Sallmans, T.; Sutskever, I. (2018): Improving Language Understanding by Generative Pre-Training. OpenAI Technical report, San Francisco, CA, https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (19.4.2023)
Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. (2019): Language models are unsupervised multitask learners. OpenAI Technical report, San Francisco, CA, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (19.4.2032)
Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. (2021): Zero-Shot Text-to-Image Generation. In: Proceedings of the 38th International Conference on Machine Learning, PMLR 139, S. 8821–8831
Rahwan, I.; Cebrian, M.; Obradovich, N.; Bongard, J.; Bonnefon, J.-F.; Breazeal, C.; Crandall, J.W.; Christakis, N.A.; Couzin, I.D.; Jackson, M.O.; Jennings, N.R. et al. (2019): Machine behaviour. In: Nature 568, 25.4.2019, S. 477–486
Rathenau Instituut (2020): Look who's talking. Tools for the responsible use of speech technology (Hamer, J., Doesborgh, S.; Kool, L.). Den Haag
Renieris, E.M. (2023): Claims That AI Productivity Will Save Us Are Neither New, nor True. 8.3.2023, www.cigionline.org/articles/claims-that-ai-productivity-will-save-us-are-neither-new-nor-true/ (19.4.2023)
Rettberg, J.W. (2022): ChatGPT is multilingual but monocultural, and it's learning your values. 6.12.2022, https://jilltxt.net/right-now-chatgpt-is-multilingual-but-monocultural-but-its-learning-your-values/ (19.4.2023)
Riera, K.; Rousseau, A.-L.; Baudelaire, C. (2020): Doctor GPT-3: hype or reality? 27.10.2020, www.nabla.com/blog/gpt-3/ (19.4.2023)
Rogers, A. (2023): The attribution problem with generative AI. 1.11.2022, https://hackingsemantics.xyz/2022/attribution/ (19.4.2023)
Rohde, F.; Wagner, J.; Reinhard, P.; Petschow, U.; Meyer, A.; Voß, M.; Mollen, A. (2019): Nachhaltigkeitskriterien für künstliche Intelligenz. Schriftenreihe des IÖW 220/21, Berlin
Roose, K. (2022): The Brilliance and Weirdness of ChatGPT. 5.12.2022, www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html (19.4.2023)
Roose, K. (2023a): Bing's A.I. Chat: »I Want to Be Alive«. 16.2.2023, www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html (19.4.2023)
Roose, K. (2023b): How ChatGPT Kicked Off an A.I. Arms Race. 3.2.2023, www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html (19.4.2023)
Roth, M. (2023): ChatGPT: Wie ein KI-Werkzeug Schule in Sachsen-Anhalt verändert. 3.2.2023, www.mdr.de/nachrichten/sachsen-anhalt/podcast-digital-leben-ki-chatgpt-schule-bildung-lernen-100.html (19.4.2023)
Rozado, D. (2023): The Political Bias of ChatGPT. In: Social Sciences 12(3): 148 (DOI: 10.3390/socsci12030148)
Rudin, C. (2019): Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. In: Nature Machine Intelligence 1, S. 206–215

Rumbelow, J.; Watkins, M. (2023): SolidGoldMagikarp (plus, prompt generation). 5.2.2023, www.alignmentforum.org/posts/aPeJE8bSo6rAFoLqg/solidgoldmagikarp-plus-prompt-generation (19.4.2023)
Salden, P.; Lordick, N.; Wiethoff, M. (2023): KI-basierte Schreibwerkzeuge in der Hochschule: Eine Einführung. In: Salden, P.; Leschke, J. (Hg.): Didaktische und rechtliche Perspektiven auf KI-gestütztes Schreiben in der Hochschulbildung. Bochum, S. 4–21
Sanderson, K. (2023): GPT-4 is here: what scientists think. In: Nature 615, 30.3.2023, S. 773
Scheuer, S. (2022): Multisearch: Google will die Internetsuche neu erfinden. 12.5.2022, www.handelsblatt.com/technik/it-internet/internet-multisearch-google-will-die-internetsuche-neu-erfinden/28335010.html (19.4.2023)
Schick, T.; Dwivedi-Yu, J.; Dessìy, R.; Raileanu, R.; Lomeli, M.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. (2023): Toolformer: Language Models Can Teach Themselves to Use Tools. arXiv:2302.04761
Schiffer, C. (2022): Warum Alexa für Amazon zum Milliarden-Flop wird. 3.12.2022, www.br.de/nachrichten/netzwelt/warum-alexa-fuer-amazon-zum-milliarden-flop-wird,TOTErHS (19.4.2023)
Schinkels, P. (2023): Ach, wie trollig. 2.3.2023, www.zeit.de/digital/internet/2023-03/chatgpt-chatbots-desinformation-soziale-medien (19.4.2023)
Schlender, H. (2022): Vom plappernden Papageien zum nützlichen Werkzeug? 19.4.2022, www.wissenschaftskommunikation.de/kuenstliche-intelligenz-papagei-nuetzliches-werkzeug-56903 (19.4.2023)
Schulministerium NRW (Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen) (2023): Umgang mit textgenerierenden KI-Systemen. Ein Handlungsleitfaden. Düsseldorf
Schwarz, S. (2023): ChatGPT und Hausaufgaben, Prüfungen. 23.1.2023, https://community.beck.de/2023/01/23/chatgpt-und-hausaufgaben-pruefungen (19.4.2023)
Seghier, M.L. (2023): ChatGPT: not all languages are equal. In: Nature 615, 9.3.2023, S. 216
Seife, C. (2023): A.I. Like ChatGPT Is Revealing the Insidious Disease at the Heart of Our Scientific Process. 31.1.2023, https://slate.com/technology/2023/01/ai-chatgpt-scientific-literature-peer-review.html (19.4.2023)
Seymour, T.; Frantsvog, D.; Kuma, S. (2011): History Of Search Engines. In: International Journal of Management & Information Systems 15(4), S. 47–58
Shah, C.; Bender, E. (2021): Situating Search. In: Proceedings of the 2022 Conference on Human Information Interaction and Retrieval (CHIIR '22), Regensburg, March 14–18. Association for Computing Machinery, New York, NY, S. 221–232
Shanahan, M. (2023): Talking About Large Language Models. arXiv:2212.03551
Sharma, A.; Lin, I.W.; Miner, A.S.; Atkins, D.C.; Althoff, T. (2022): Human-AI Collaboration Enables More Empathic Conversations in Text-based Peer-to-Peer Mental Health Support. arXiv:2203.15144v1
Sheldon, J. (2022): o.T. (Linked-In-Beitrag), www.linkedin.com/posts/joshsheldon_marketing-technology-chatgpt-activity-7007083252779778049-LNiZ/ (19.4.2023)
Sheng, E.; Chang, K.-W.; Natarajan, P.; Peng, N. (2021): Societal Biases in Language Generation: Progress and Challenges. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing 1, S. 4275–4293

Shevlin, H.; Halina, M. (2019): Apply rich psychological terms in AI with care. In: Nature Machine Intelligence 1, S. 165–167
Simon, J. (2021): Large Language Models: A New Moore's Law? 26.10.2021, https://huggingface.co/blog/large-language-models (19.4.2023)
Smith, S.; Patwary, M.; Norick, B.; LeGresley, P.; Rajbhandari, S.; Casper, J.; Liu, Z.; Prabhumoye, S.; Zerveas, G.; Korthikanti, V.; Zhang, E. et al. (2022): Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. arXiv:2201.11990
Snoswell, A.J.; Burgess, J. (2022): The Galactica AI model was trained on scientific knowledge – but it spat out alarmingly plausible nonsense. 30.11.2022, https://theconversation.com/the-galactica-ai-model-was-trained-on-scientific-knowledge-but-it-spat-out-alarmingly-plausible-nonsense-195445 (19.4.2023)
Spannagel, C. (2023): ChatGPT und die Zukunft des Lernens: Evolution statt Revolution. 24.1.2023, https://hochschulforumdigitalisierung.de/de/blog/chatgpt-und-die-zukunft-des-lernens-evolution-statt-revolution (19.4.2023)
Spataro, J. (2023): Introducing Microsoft 365 Copilot – your copilot for work. 16.3.2023, https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/ (19.4.2023)
Sperl, A. (2023): ChatGPT. 28.2.2023, www.fernuni-hagen.de/zli/blog/chatgpt/ (19.4.2023)
Spiegel Online (2023): Microsoft legt Bing-Chatbot an die Leine. 19.2.2023, www.spiegel.de/netzwelt/wegen-uebergriffiger-antworten-microsoft-legt-bing-chatbot-an-die-leine-a-a70246a7-0a89-475e-b353-1cf380055ac0 (19.4.2023)
Spielkamp, M. (2023): AlgorithmWatch fordert Regulierung von »General Purpose AI« in der KI-Verordnung der EU. Pressemitteilung vom 13.4.2023, https://algorithmwatch.org/de/regulierung-general-purpose-ai-ki-verordnung/ (19.4.2023)
Stock, L. (2023): ChatGPT an Universitäten – wie KI Studierenden helfen kann. 20.1.2023, www.dw.com/de/chatgpt-an-universit%C3%A4ten-wie-ki-studierenden-helfen-kann/a-64418962 (19.4.2023)
Stokel-Walker, C. (2023): ChatGPT listed as author on research papers. In: Nature 613, 9.2.2023, S. 620–621
Strauß, S. (2021): Deep Automation Bias: How to Tackle aWicked Problem of AI? In: Big Data and Cognitive Computing 5(2), 18 (DOI: 10.3390/bdcc5020018)
Strobelt, H.; Webson, A.; Sanh, V.; Hoover, B.; Beyer, J.; Pfister, H.; Rush, A.M. (2022): Interactive and Visual Prompt Engineering for Ad-hoc Task Adaptation with Large Language Models. In: IEEE Transactions on Visualization and Computer Graphics 29(1), S. 1146–1156
Szöke, D. (2023): Mittels Sprachmodell Robotik steuern: Google und TU Berlin stellen PaLM-E vor. 15.3.2023, www.heise.de/news/Mittels-Sprachmodell-Robotik-steuern-Google-und-TU-Berlin-stellen-PaLM-E-vor-7543506.html (19.4.2023)
TAB (Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag) (2011): Schwerpunkt: Hope-, Hype- und Fear-Technologien. TAB-Brief Nr. 39, Berlin
TAB (2016a): Digitale Medien in der Bildung (Albrecht, S.; Revermann, C.). TAB-Arbeitsbericht Nr. 171, Berlin
TAB (2016b): »Responsible Research and Innovation« als Ansatz für die Forschungs-, Technologie- und Innovationspolitik – Hintergründe und Entwicklungen (Lindner, R.; Goos, K.; Güth, S.; Som, O.; Schröder, T.). TAB-Hintergrundpapier Nr. 22, Berlin

TAB (2017a): Chancen und Risiken mobiler und digitaler Kommunikation in der Arbeitswelt. (Börner, F.; Kehl, C.; Nierling, L.). TAB-Arbeitsbericht Nr. 174, Berlin
TAB (2017b): Online-Bürgerbeteiligung an der Parlamentsarbeit (Oertel, B.; Kahlisch, C.; Albrecht, S.). TAB-Arbeitsbericht Nr. 173, Berlin
TAB (2017c): Social Bots. TA-Vorstudie (Kind, S.; Jetzke, T.; Weide, S.; Ehrenberg-Silies, S.; Bovenschulte, M.). Horizon Scanning Nr. 3, Berlin
TAB (2018): Robotik und assistive Neurotechnologien in der Pflege – gesellschaftliche Herausforderungen (Kehl, C.). TAB-Arbeitsbericht Nr. 177, Berlin
TAB (2019a): Deepfakes - Manipulation von Filmsequenzen (Bovenschulte, M.). Themenkurzprofil Nr. 25, Berlin
TAB (2019b): Legal Tech – Potenziale und Wirkungen (Kind, S.; Ferdinand, J.-P.; Priesack, K.). TAB-Arbeitsbericht Nr. 185, Berlin
TAB (2022a): Algorithmen in digitalen Medien und ihr Einfluss auf die Meinungsbildung (Oertel, B.; Dametta, D.; Kluge, J.; Todt, J.). TAB-Arbeitsbericht Nr. 204, Berlin
TAB (2022b): Data-Mining – gesellschaftspolitische und rechtliche Herausforderungen (Gerlinger, K.). TAB-Arbeitsbericht Nr. 203, Berlin
TAB (2022c): Künstliche Intelligenz und Distributed-Ledger-Technologie in der öffentlichen Verwaltung (Evers-Wölk, M.; Kluge, J.; Steiger, S.). TAB-Arbeitsbericht Nr. 201, Berlin
TAB (2022d): Sprich mit mir! Perspektiven für den Einsatz KI-basierter Dialogsysteme (Peters, R.). Themenkurzprofil Nr. 52, Berlin
Tamkin, A.; Brundage, M.; Clark, J.; Ganguli, D. (2021): Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. arXiv:2102.02503v1
Tamkin, A.; Ganguli, D. (2021): How Large Language Models Will Transform Science, Society, and AI. 5.2.2021, https://hai.stanford.edu/news/how-large-language-models-will-transform-science-society-and-ai (19.4.2023)
The Economist (2022): The world that Bert built. Briefing vom 11.6.2022, S. 17–20
Thiede, D. (2023): ChatGPT – AI Chat. 8.2.2023, https://datenschutz-schule.info/datenschutz-check/chatgpt-ai-chat/ (19.4.2023)
Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; Li, Y. et al. (2022): LaMDA: Language Models for Dialog Applications. arXiv:2201.08239
Thorp. H.H. (2023): ChatGPT is fun, but not an author. In: Science 379(6630), S. 313
Tiku, N.; De Vynck, G.; Oremus, W. (2023): Big Tech was moving cautiously on AI. Then came ChatGPT. 3.2.2023, www.washingtonpost.com/technology/2023/01/27/chatgpt-google-meta/ (19.4.2023)
Toews, R. (2022a): A Wave Of Billion-Dollar Language AI Startups Is Coming. 27.3.2022, www.forbes.com/sites/robtoews/2022/03/27/a-wave-of-billion-dollar-language-ai-startups-is-coming/ (19.4.2023)
Toews, R. (2022b): Language is the next great frontier in AI. 13.2.2022, www.forbes.com/sites/robtoews/2022/02/13/language-is-the-next-great-frontier-in-ai/ (19.4.2023)
Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A. et al. (2023): LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971

Towey, H. (2023): Chat GPT plant einen Familienurlaub in Costa Rica – und beweist mit dem Ergebnis, dass es Reise-Experten nicht ersetzen kann. 20.3.2023, www.businessinsider.de/wirtschaft/international-business/chatgpt-plant-familienurlaub-in-costa-rica-das-ging-dort-schief/ (19.4.2023)
Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. (2017): Attention Is All You Need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, S. 6000–6010
Véliz, C. (2023): Chatbots shouldn't use emojis. In: Nature 615, 16.3.2023, S. 375
Venkit, P.N.; Srinath, M.; Wilson, S. (2022): A Study of Implicit Language Model Bias Against PeopleWith Disabilities. In: Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, S. 1324–1332
Vincent, J. (2023): Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content. 17.1.2023, www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit (19.4.2023)
Vinsel, L. (2021): You're Doing It Wrong: Notes on Criticism and Technology Hype. 1.2.2021, https://sts-news.medium.com/youre-doing-it-wrong-notes-on-criticism-and-technology-hype-18b08b4307e5 (19.4.2023)
Vogelgesang, J.; Bleher, J.; Krupitzer, C.; Stein, A.; Jung, R. (2023): Nutzung von ChatGPT in Lehre und Forschung – eine Einschätzung der AIDAHO-Projektgruppe. Positionspapier, Universität Hohenheim, https://aidaho.uni-hohenheim.de/fileadmin/einrichtungen/aidaho/Dokumente/AIDAHO_ChatGPT_Positionspapier_23-02-09.pdf (19.4.2023)
Volkert, J. (2023): Experiment ohne Einwilligung: Menschen suchen Hilfe – und landen bei einer KI. 11.1.2023, www.heise.de/news/Experiment-ohne-Einwilligung-Menschen-suchen-Hilfe-und-landen-bei-einer-KI-7455475.html (19.4.2023)
Volland, M. (2023): Large-Language-Modelle und mögliche Anwendungsbereiche im Recht. 2.1.2023, https://lrz.legal/de/lrz/large-language-modelle-und-moegliche-anwendungsbereiche-im-recht (19.4.2023)
Voß, O. (2023): Chatbot-Suche: Verleger fordern Lizenzgebühren von Microsoft und Google. 13.2.2023, www.tagesspiegel.de/wirtschaft/chatbot-suche-verleger-fordern-lizenzgebuhren-von-microsoft-und-google-9336164.html (19.4.2023)
de Waard, A. (2023): Guest Post – AI and Scholarly Publishing: A View from Three Experts. 18.1.2023, https://scholarlykitchen.sspnet.org/2023/01/18/guest-post-ai-and-scholarly-publishing-a-view-from-three-experts/ (19.4.2023)
Wang, S.H. (2023): OpenAI — explain why some countries are excluded from ChatGPT. In: Nature 615, 2.3.2023, S. 34
Waters, R. (2023): Man beats machine at Go in human victory over AI. 19.2.2023, https://arstechnica.com/information-technology/2023/02/man-beats-machine-at-go-in-human-victory-over-ai/ (19.4.2023)
Wayner, P. (2023): 10 reasons to worry about generative AI. 13.2.2023, www.infoworld.com/article/3687211/10-reasons-to-worry-about-generative-ai.html (19.4.2023)
WD (Wissenschaftliche Dienste) (2022): Wartezeiten auf eine Psychotherapie. Umfragen und Studien. Deutscher Bundestag, Ausarbeitung Nr. WD 9 – 3000 – 059/22, Berlin

Wedig, M. (2023): ChatGPT in der Schule: »KI ersetzt nicht den gemeinsamen Unterricht«. 12.3.2023, www.spiegel.de/deinspiegel/chatgpt-in-der-schule-ki-ersetzt-nicht-den-gemeinsamen-unterricht-a-416fae9d-c377-41cb-94b5-d642c69da133 (19.4.2023)
Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E.H. et al. (2022): Emergent Abilities of Large Language Models. In: Transactions on Machine Learning Research (8/2022), https://openreview.net/pdf?id=yzkSU5zdwD (19.4.2023)
Weiß, E.-M. (2023a): Baidu veröffentlicht ErnieBot – ChatGPT-Konkurrenz für seine Suchmaschine. 7.2.2023, www.heise.de/news/Baidu-plant-ChatGPT-Konkurrenz-fuer-seine-Suchmaschine-7478392.html (19.4.2023)
Weiß, E.-M. (2023b): Clyde wird intelligent: Discord setzt auf KI von OpenAI. 9.3.2023, www.heise.de/news/Clyde-wird-intelligent-Discord-setzt-auf-KI-von-OpenAI-7540550.html (19.4.2023)
Weßels, D. (2022): ChatGPT – ein Meilenstein der KI-Entwicklung. 20.12.2022, www.forschung-und-lehre.de/lehre/chatgpt-ein-meilenstein-der-ki-entwicklung-5271 (19.4.2023)
White House; Europäische Kommission (2022): The Impact of Artificial Intelligence on the Future of Workforces in the European Union and the United States of America. 5.12.2022, www.whitehouse.gov/wp-content/uploads/2022/12/TTC-EC-CEA-AI-Report-12052022-1.pdf (19.4.2023)
Wiggers, K. (2022): You.com launches an AI-powered writing tool powered by OpenAI. 15.3.2022, https://venturebeat.com/ai/you-com-partners-with-openai-to-launch-an-ai-powered-writing-tool/ (19.4.2023)
Wilkens, U.; Herrmann, T. (2016): Gibt es eine Arbeitswissenschaft der Digitalisierung? Ein Diskursbeitrag. In: Schlick, C. (Hg.): Megatrend Digitalisierung - Potenziale der Arbeits- und Betriebsorganisation. Berlin, S. 215–230
Wittenbrink, N.; Demirci, S.; Wischmann, S. (2023): Resiliente und robuste KI-Systeme im praktischen Einsatz. In: Wittpahl, V. (Hg.): Resilienz. Berlin, Heidelberg, S. 199–211
Wittenhorst, T. (2022): Microsoft, GitHub und OpenAI verklagt: KI-Programmierhilfe Copilot kopiert Code. 6.11.2022, www.heise.de/news/Microsoft-GitHub-und-OpenAI-verklagt-KI-Programmierhilfe-Copilot-kopiert-Code-7331566.html (19.4.2023)
Wolf, M.J.; Miller, K.W.; Grodzinsky, F.S. (2017): Why We Should Have Seen That Coming. In: The ORBIT Journal 1 (2), S. 1–12
Wolfangel, E. (2021): Wie viel Ethik verträgt Google? 2.2.2021, www.zeit.de/digital/2021-02/google-ethik-timnit-gebru-technologie-forschung (19.4.2023)
Wolfangel, E. (2023): Wie man Chatbots die Wahrheit beibringt. 26.2.2023, www.zeit.de/digital/2023-02/ki-sprachmodelle-chatbots-kuenstliche-intelligenz-wahrheit (19.4.2023)
Wolfram, S. (2023a): ChatGPT Gets Its »Wolfram Superpowers«! 23.3.2023, https://writings.stephenwolfram.com/2023/03/chatgpt-gets-its-wolfram-superpowers/ (19.4.2023)
Wolfram, S. (2023b): What Is ChatGPT Doing … and Why Does It Work? 14.2.2023, https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/ (19.4.2023)
Wright, J.W. (2023): A new era for AI and Google Workspace. 14.3.2023, https://workspace.google.com/blog/product-announcements/generative-ai (19.4.2023)

Yin, B.; Corradi, F.; Bohté, S.M. (2021): Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks. In: Nature machine intelligence 3, S. 905–913

Zakir, T. (2023): ChatGPT: The Bot That Needs a Chat About Regulation. 7.3.2023, www.nyujlb.org/single-post/chatgpt-the-bot-that-needs-a-chat-about-regulation (19.4.2023)

Zhdanov, F. (2023): What happens to a large language model (LLM) after it's trained. 21.1.2023, https://venturebeat.com/ai/what-happens-to-an-llm-after-its-trained (19.4.2023)

BÜRO FÜR TECHNIKFOLGEN-ABSCHÄTZUNG BEIM DEUTSCHEN BUNDESTAG

Karlsruher Institut für Technologie

Neue Schönhauser Straße 10
10178 Berlin

Telefon: +49 30 28491-0
E-Mail: buero@tab-beim-bundestag.de
Web: www.tab-beim-bundestag.de
Twitter: @TABundestag