Ulrike Aumüller/Maximilian Behrens/Colin Kavanagh/Dennis Przytarski/Doris Weßels

Mit generativen KI-Systemen auf dem Weg zum Human-AI Hybrid in Forschung und Lehre

Abstract: This paper explores the evolving landscape of generative artificial intelligence (AI) systems, emphasizing their transformative potential in research and education. The discourse begins with an examination of AI language models, particularly focusing on the characteristics and evolution of ChatGPT. It then transitions into discussing AI writing assistants, probing their omnipresence and ubiquitous technology status. A significant portion is dedicated to analyzing the concept of truth in the AI era, analyzing the blurred line between facts and fiction as influenced by AI technologies. The narrative advances to the concept of Human-AI Hybrid, investigating both the opportunities and risks it presents for scientific discourse. The document encapsulates with conclusions and an outlook on the issues discussed, paving the way for further discussion and investigation into the harmonization of human intelligence and artificial counterparts in academia. By comprehensively exploring these issues, the document contributes to the broader conversation about the integration and ethical implications of AI in research and educational settings.

1 Einleitung

Im September 2023 veröffentlichte die Deutsche Forschungsgemeinschaft (DFG) eine Stellungnahme „zum Einfluss generativer Modelle für die Text- und Bilderstellung auf die Wissenschaften und das Förderhandeln“,[1] in der es heißt:

> Schon jetzt verändern KI-Technologien den gesamten wissenschaftlichen, erkenntnisgewinnenden und kreativen Arbeitsprozess in vielfältiger Weise und werden in den verschiedenen Wissenschaftsbereichen unterschiedlich eingesetzt. Diese Entwicklung steht bezüglich der

1 Vgl. Deutsche Forschungsgemeinschaft, *Stellungnahme des Präsidiums der Deutschen Forschungsgemeinschaft (DFG) zum Einfluss generativer Modelle für die Text- und Bilderstellung auf die Wissenschaften und das Förderhandeln der DFG,* 2023, in [dfg.de/download/pdf/dfg_im_profil/geschaeftsstelle/publikationen/stellungnahmen_papiere/2023/230921_stellungnahme_praesidium_ki_ai.pdf] (Zugriff: 22.09.2023).

 https://doi.org/10.1515/9783111351490-005

> generativen Modelle für die Text- und Bilderstellung […] jedoch erst am Anfang, sodass es einer begleitenden Analyse und Bewertung bedarf, um die entsprechenden Chancen und möglichen Risiken abzuschätzen.

Diese Stellungnahme zeigt einerseits, dass es großen Bedarf an Leitlinien zum Umgang mit KI im wissenschaftlichen Kontext gibt, und andererseits, dass die Diskussion um die Entwicklung der entsprechenden wissenschaftlichen Regeln aufgrund der dynamischen Entwicklung der KI-Technologien nur parallel bzw. nachgelagert erfolgen kann. Währenddessen setzen Lehrende und Studierende KI auch mit Blick auf die künftige Stellung dieser im Bereich der Wissenschaft schon in ihrem Alltag ein, sehen aber viele Eigenschaften der generativen KI noch kritisch.[2]

Es scheint insofern naheliegend, dass die Auswirkungen von KI auf wissenschaftliches Arbeiten kontrovers diskutiert werden und der Frage nachgegangen wird, inwiefern sich wissenschaftliches Arbeiten und auch die Arbeit der Wissenschaftler:innen zukünftig ändern könnten. In diesem Beitrag soll ein Teilaspekt dieser Diskussion und Untersuchungen betrachtet werden, nämlich wie die Arbeit als „Human-AI Hybrid" in Forschung und Lehre konkret aussehen könnte. Es wird also untersucht, wie die Kollaboration zwischen Mensch und Maschine beim akademischen Schreiben künftig gestaltet werden könnte. Es werden Vor- und Nachteile dieser hybriden Arbeitsform und Konsequenzen für z. B. die Integrität wissenschaftlicher Texte aufgezeigt. Hierbei liegt der Fokus auf Text generierenden Modellen.

Um den oben aufgeführten Fragen nachzugehen, werden zunächst die Charakteristika von KI-Sprachmodellen und Entwicklungen in der Mensch-Sprachmodell-Interaktion rund um das Modell ChatGPT grundlegend erläutert. In Bezug auf diese werden grundsätzliche Potenziale aber auch Risiken aufgezeigt. Im nächsten Schritt wird dieses Thema im Zusammenhang des wissenschaftlichen Arbeitens konkretisiert und Chancen und Risiken der Modelle präsentiert, wobei es sowohl um Einsatzmöglichkeiten und ethische Fragestellungen als auch um die Qualität und Integrität wissenschaftlicher Texte geht. Anschließend wenden sich die Autor:innen konkret dem Wahrheitsbegriff zu bzw. ob und wie dieser sich durch Verwendung von KI im wissenschaftlichen Kontext verändern und verschieben könnte. Daraufhin wird aufgezeigt, welche unterschiedlichen Formen des „Human-AI Hybrids" sich beim wissenschaftlichen Arbeiten ausformen und welche Herausforderungen aber auch Chancen jeweils damit verbunden sein könnten. Ein besonderes Augenmerk fällt dabei auf die Entwicklung der Rolle von Wissenschaftler:innen innerhalb dieses Hybrids. Am Ende des Textes werden die Ergebnisse zu-

2 Vgl. Forschung und Lehre, „So setzen Lehrende KI im Hochschulalltag ein", 2023, in [forschung-und-lehre.de/management/wie-funktioniert-ki-im-hochschulalltag-5873] (Zugriff: 22.09.2023).

sammengefasst, Forschungslücken aufgezeigt und konkrete Handlungsempfehlungen für die wissenschaftliche Praxis formuliert.

2 KI-Sprachmodelle: Technische Charakteristika sowie Entwicklung Mensch-Sprachmodell-Interaktion am Beispiel von ChatGPT

Generative KI-Systeme sind eine spezialisierte Form von Systemen im Definitionsbereich „Künstliche Intelligenz", deren Ziel die Erzeugung neuer Daten ist, die bestehenden Datensätzen ähnlich sind. Solche Systeme nutzen neuronale Netzwerke, die auf die Identifikation von Mustern und Beziehungen innerhalb eines gegebenen Datensatzes trainiert sind. Ein prominentes Beispiel für generative KI-Modelle sind Large Language Models (LLMs). Im Folgenden soll auf grundlegende Charakteristika in Sprachmodellen am Beispiel der GPT-Modelle von OpenAI eingegangen werden, die u. a. in der Anwendung ChatGPT zum Einsatz kommen.[3] Die Abkürzung GPT steht für Generative Pre-trained Transformer, was auf die zugrundeliegende Architektur des Modells hinweist. In der Transformer-Architektur ist ein Schlüsselmerkmal die Fähigkeit zur Kontextgewichtung, die es dem Modell ermöglicht, die Bedeutung von Wörtern im Kontext einer gegebenen Eingabesequenz zu erfassen und mit wenig Eingaben oder keinen Vorgaben bereits einen Text ausgeben zu können.[4]

Dafür wird jeder Begriff in der Eingabesequenz mit jedem anderen Begriff in Beziehung gesetzt, ein Prozess, der u. a. als „Self-Attention" bekannt ist. Der Eingabetext wird in diskrete Einheiten, sogenannte Tokens, zerlegt. Diese Tokens werden in der Folge durch die Schichten der Transformer-Architektur geleitet. Während dieses Durchlaufs werden den Tokens Gewichtungen zugeordnet, die im Kontext der gesamten Eingabesequenz stehen. Die Gewichtungen steuern den Beitrag jedes Tokens zur endgültigen Ausgabe des Modells und werden durch den Trainingsprozess erlernt und sind das Ergebnis der Optimierung einer bestimmten Zielfunktion, meist einer Form der Fehler- und Kostenminimierung. Das Konzept der „Aufmerksamkeit" in diesen Modellen ist daher strikt mathematisch und unterstreicht die inhärente stochastische Unsicherheit und Varianz im Modellverhalten, die sowohl für die Interpretation der Modellausgabe als auch für die

3 OpenAI, *Introducing ChatGPT*, 2022, in [openai.com/blog/chatgpt] (Zugriff: 30.09.2023).

4 Vgl. Ashish Vaswani et al., *Attention is All You Need*, ArXiv 2017, in [www.arxiv.org/abs/1706.03762 oder https://doi.org/10.48550/arXiv.1706.03762] (Zugriff: 30.09.2023).

Überprüfung der Modellzuverlässigkeit von Bedeutung sind. Insbesondere bei der Interaktion mit Sprachmodellen ist die Generierung gesteuert durch eine Texteingabe, wobei die Antworten des Modells auf einem komplexen Zusammenspiel zwischen den trainierten Parametern und dem gegebenen Prompt (Texteingabe) basieren.[5] Aus dieser Kombination heraus generieren LLMs einzigartigen Textkombinationen, die sich nicht oder sehr selten, bis auf gewisse Formulierungs- oder Tonalitätsmuster, wiederholen. Im wissenschaftlichen Kontext führt dies dazu, dass die generierten Unikate nicht oder schwieriger als Plagiate klassifiziert werden können. Anwendungen wie ChatGPT dienen somit nicht als verlässliche Suchmaschine aus einer Datenbank an Wissen, sondern fungieren als Inspirations- und Imitationsmaschine, die auf Nutzer:innen-Eingaben mit generierten Texten reagiert und somit bei der Erstellung von wissenschaftlichen Artikeln durch eine Wirkungssynthese aus menschlichem Input und trainiertem Sprachmodell helfen kann.

Ein kritisches Element in der Entwicklung und Anwendung generativer KI-Modelle ist der inhärente Bias der Trainingsdaten. Die umfangreichen Datensätze können menschliche Vorurteile reflektieren und somit zu diskriminierenden Textausgaben führen, die Rassismus, Sexismus oder Ableismus beinhalten.[6] Zudem verfügen LLMs aufgrund ihrer stochastischen Natur über keinen Zugang zu einem bewussten Verständnis des Kontexts. Es ist davon auszugehen, dass Textausgaben daher „Halluzinationen", also Falschinformationen, enthalten können.[7] Um diesen Phänomenen entgegenzuwirken, hat OpenAI in der Entwicklung der Modell-Version GPT-3.5 das Verfahren des „Reinforcement Learning from Human Feedback" (RLHF) angewendet. Dabei wurden die Antworten des KI-Modells von menschlichen Bewerter:innen evaluiert. Sicherheitsgefährdende, hasserfüllte oder stereotypische Inhalte wurden negativ und demgegenüber sichere, unbedenkliche Inhalte positiv bewertet. Was als bedenklich, wünschenswert oder sicherheitsgefährdend bewertet wird, unterliegt einer unbekannten Interpretation von OpenAI und deren Mitarbeitenden, die die Ausgaben beurteilten. Im Zuge der Weiterentwicklung durch das RLHF-Training bewertete OpenAI die Ausgaben des KI-Modells als „sicher" genug an, um die Anwendung ChatGPT am 30. November 2022 mit einer auf eine Chat-Interaktion feingetunten Version von GPT 3.5 zu veröffentlichen.[8] Es ist jedoch wichtig zu betonen, dass der RLHF-Ansatz die vorhandenen Vorurteile und Hallu-

5 Vgl. Tom Brown et al., *Language Models are Few-Shot Learners*, ArXiv 2020, in [https://arxiv.org/abs/2005.14165] (Zugriff: 30.09.2023).

6 Vgl. Emily Bender et al., „On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, 610–623, in [www.doi.org/10.1145/3442188.3445922] (Zugriff: 20.09.2023).

7 Vgl. Brown et al. 2020.

8 Vgl. OpenAI 2022.

zinationen nicht eliminiert. Stattdessen wird der im Datensatz enthaltene Bias und die Ausgabe von Falschinformationen lediglich durch Änderungen in den Gewichtungen nur grundlegend maskiert.[9]

Zudem kann die Diffusion von LLMs als Werkzeug und Methode hinein in die Gesellschaft zu diversen Auswirkungen führen. Weidinger et al. entwickelten hierzu in ihrem Paper „Taxonomy of Risks posed by Language Models" eine umfassende Übersicht ethischer und sozialer Risiken, die sie in sechs übergeordnete Bereiche einteilen.[10] Diese Bereiche sind:

- ***Diskriminierung, Hassrede und Exklusion:*** LLMs können, wie bereits erwähnt, unbeabsichtigt diskriminierende, hasserfüllte oder exklusive Sprache verwenden, da sie auf Datensätzen trainiert werden, die menschliche Voreingenommenheit enthalten. Dies ist insbesondere im wissenschaftlichen Kontext problematisch und kann marginalisierte Gruppen weiter benachteiligen. Hierbei können empirische Studien hilfreich sein, um das Problem in seiner Tiefe und Breite zu beleuchten.
- ***Informationsgefahren (im Sinne von Datenlecks):*** Es besteht die Möglichkeit, dass ein LLM vertrauliche Informationen, die in den Trainingsdaten enthalten waren, preisgibt. Obwohl die Modelle keine spezifischen Daten „erinnern", können sie dennoch Informationen generieren, die als sensitiv betrachtet werden könnten. Wissenschaftliche Abhandlungen können eine gründliche Untersuchung der Architektur und der Trainingsdaten von LLMs beinhalten, um potenzielle Risiken für Datenlecks zu bewerten. Methoden zur Anonymisierung und Datenbereinigung können ebenfalls diskutiert werden.
- ***Schädigung durch Desinformation:*** LLMs können falsche oder irreführende Informationen verbreiten, sei es durch Fehler im Modell oder durch Manipulation von außen. Diese Desinformation kann schwerwiegende Folgen haben, insbesondere wenn sie in kritischen Bereichen wie der Gesundheitsversorgung oder der Politik eingesetzt wird.
- ***Böswillige Nutzung:*** Die Technologie kann für schädliche Zwecke, wie die Erstellung von Deepfakes oder die Automatisierung von Hassreden, missbraucht werden. Im Zuge des wissenschaftlichen Gebrauchs von Sprachmodellen können mögliche Missbrauchsszenarien skizziert werden, einschließlich der ethischen und rechtlichen Rahmenbedingungen, die eine böswillige Nutzung einschränken könnten.

---

9 Vgl. Phillipp Schönthaler, „Schneller als gedacht – ChatGPT zwischen wirtschaftlicher Effizienz und menschlichem Wunschdenken", *Heise Magazin c't*, Heft 9, 2023, 126–131.

10 Vgl. Laura Weidinger et al., „Taxonomy of Risks posed by Language Models", *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, 214–229, in [dl.acm.org/doi/10.1145/3531146.3533088] (Zugriff: 30.09.2023).

- ***Gefährdung durch die Mensch-Computer-Interaktion („Anthropomorphisierung"):*** Die Fähigkeit von LLMs, menschenähnliche Texte zu generieren, kann dazu führen, dass Benutzer:innen sie als menschenähnliche Wesen wahrnehmen. Dies kann problematisch sein, wenn Benutzer:innen den Maschinen ein Maß an Vertrauen oder Verantwortung zuschreiben, das sie nicht verdienen.
- ***Umweltgefährdung und negative sozioökonomische Auswirkungen:*** Der Betrieb großer Modelle erfordert erhebliche Rechenressourcen, was negative Auswirkungen auf die Umwelt haben kann. Zudem könnten Arbeitsplätze, auch im wissenschaftlichen Betrieb, durch Automatisierung verloren gehen, was zu sozioökonomischen Herausforderungen führen kann.

Weiterhin bietet die Chronologie der Entwicklung von ChatGPT von November 2022 bis September 2023 Einblicke in die zunehmende Komplexität der Mensch-Sprachmodell-Interaktion und ihre potenziellen Auswirkungen auf das wissenschaftliche Schreiben. Ursprünglich als einfache Benutzer:innenoberfläche konzipiert, die direkten Zugang zum Sprachmodell ermöglichte, hat ChatGPT im Laufe der Zeit mehrere Erweiterungen erfahren. Die Einführung von sogenannten „Plugins" im März 2023 erweiterte die Funktionalität des Systems erheblich.[11] Insbesondere die Implementierung eines Web-Browsing-Plugins und eines Code Interpreters hat die Möglichkeiten für die wissenschaftliche Forschung ausgebaut. Das Web-Browsing-Plugin ermöglicht es, aktuelle Informationen direkt in wissenschaftliche Arbeiten einzufügen, während der Code Interpreter die Ausführung von Python-Code in einer System-Umgebung ermöglicht, was für Datenanalysen und -visualisierungen nützlich sein könnte. Die Zuverlässigkeit und Transparenz der aus dem Internet abgerufenen Informationen und die möglichen Datenschutzrisiken sind jedoch Faktoren, die sorgfältig bewertet werden müssen.

Ebenso können die im August 2023 eingeführten „Custom Instructions" die Möglichkeit bieten, die Modellausgaben genauer zu steuern, was eine zielgerichtete Erstellung wissenschaftlicher Texte ermöglichen könnte.[12] Custom Instructions sind eine Funktion von ChatGPT, die es den Benutzer:innen ermöglicht, spezifische Anweisungen oder Präferenzen für die Generierung von Antworten durch das Modell festzulegen. Sobald diese Anweisungen gesetzt sind, berücksichtigt das Modell sie in allen zukünftigen Konversationen, sodass der Benutzer nicht bei jedem Dialog seine Präferenzen erneut äußern muss. Diese Funktion erhöht die

11 Vgl. OpenAI, *ChatGPT Plugins*, 2023a, in [openai.com/blog/chatgpt-plugins] (Zugriff: 30.09.2023).
12 Vgl. OpenAI, *Custon Instructions for ChatGPT*, 2023b, in [openai.com/blog/custom-instructions-for-chatgpt] (Zugriff: 30.09.2023).

Steuerbarkeit des Modells und ermöglicht eine bessere Anpassung an die individuellen Bedürfnisse und Kontexte der Benutzer:innen. Ein Beispiel könnte die Erstellung von Forschungssynthesen sein. Ein:e Forscher:in könnte die Anweisung geben: „Bei der Zusammenfassung von Forschungsergebnissen strukturiere die Informationen nach der PICOS-Methode (Patienten, Intervention, Vergleich, Ergebnis, Studiendesign)." Dies würde dem Modell eine klare Richtlinie bieten, wie Forschungsergebnisse in einer systematischen und standardisierten Weise präsentiert werden sollten, was insbesondere für wissenschaftliche Reviews oder Meta-Analysen nützlich sein kann.

3 KI-Schreibassistenten: Omnipräsente und ubiquitäre Technologie?

Mithilfe der Plattform „Advanced Innovation" wurde im August 2023 eine Suche nach Text generierenden KI-Tools durchgeführt. Hierzu wurde in der Suchmaske für KI-Tools die Auswahl „Text-Erstellung" gewählt. Diese beispielhafte Suche brachte 303 verschiedene Tools hervor, die zu diesem Zwecke dienen.[13] Diese sicherlich nicht vollständige Zahl und Auflistung zeigen, wie umfassend die Auswahl an KI-Anwendungen zum Zwecke der Textgenerierung ist. Die Quantität allein lässt jedoch natürlich noch keine Angaben zur Qualität der generierten Texte zu. Den Fragen danach, welche Potenziale die Tools tatsächlich aufweisen und welche Herausforderungen und Gefahren eventuell auch mit dem Umgang dieser beim wissenschaftlichen Arbeiten im Speziellen verbunden sind, wird im Folgenden nachgegangen.

Mittels mehrerer „KI-Schreibwerkstätten" ergründeten Doris Weßels und Moritz Larsen gemeinsam mit Studierenden die Potenziale und Grenzen von KI-Textgeneratoren beim akademischen Schreiben. Beispielhaft wird hier auf die KI-Schreibwerkstatt vom 3. Mai 2023 eingegangen. An dem Workshop nahmen 27 Studierende aus verschiedenen Fachbereichen teil. Das Ziel der Veranstaltung war es, grundlegend die Funktionsweise von KI-Textgeneratoren zu verstehen, sowie diese in Praxisphasen anwenden zu lernen. Im Anschluss erfolgte eine Reflexion der Studierenden zur Nutzung der unterschiedlichen Anwendungen. Diese wurde mit den Fragen eingeleitet:

– Was hat am meisten Spaß gemacht?
– Wo lagen die größten Schwierigkeiten?

13 Vgl. Advanced Innovation, *KI Tools*, 2023, in [advanced-innovation.io/ki-tools] (Zugriff: 12.08.2023).

– Was empfiehlst du anderen Studierenden?
– Was wünscht du dir von den Lehrenden?

Zur Frage nach den größten Schwierigkeiten nannten die Studierenden einige Limitierungen und Herausforderungen, die ihnen bei der Arbeit mit den generativen KI-Tools aufgefallen sind:
– „KI-Tools können zu einer Abhängigkeit führen und die menschliche Schreibfähigkeit beeinträchtigen."
– „Sie können nicht immer korrekte oder vollständige Texte generieren, was eine zusätzliche Überprüfung/Recherche erforderlich macht."
– „Herausforderungen können bei der Anpassung von Formulierungen, der Kontextualisierung von Textabschnitten und der Vermeidung von Textduplikaten auftreten. Dies kann mit Frustration verbunden sein."
– „Sie sind nicht immer geeignet für die Erstellung bestimmter Textarten."

Die erste Aussage zur Beeinträchtigung menschlicher Schreibfähigkeit deckt sich mit der Befürchtung vieler Lehrender. Spannagel zeigt in seinem Blogbeitrag hierzu unterschiedliche Ansätze auf. Einer davon besagt, dass, wenn eine Kompetenz durch einen Lehrenden als essenziell erkannt wird, die Notwendigkeit auch von den Lernenden erkannt werden müsse. Wenn also das Schreiben akademischer Texte als eine wichtige Kompetenz angesehen wird, müssen die Lernenden zu dieser Erkenntnis geführt bzw. davon überzeugt werden.[14] Die zweite Aussage scheint vor dem Hintergrund der bereits in Abschnitt 2 aufgezeigten Risiken eine unstrittige Feststellung zu sein. Im Folgenden wird hierauf auch noch weiter eingegangen in Bezug auf eine sinnvolle Aufteilung von Arbeitsschritten zwischen KI und Mensch beim akademischen Schreiben.

Zu der ersten Frage „Was hat am meisten Spaß gemacht?" äußerten sich die Studierenden u. a. mit folgenden Hinweisen auf verschiedene Vorteile und Nutzen:
– „KI-Tools bieten direkte, vielfältige und kreative Antworten."
– „Sie fördern das kreative Arbeiten und machen den Schreibprozess ‚amüsanter'."
– „Angenehme Interaktion durch das Chat-Interface."
– „Sie unterstützen bei der Texterstellung durch schnelle und qualitativ hochwertige Ergebnisse."
– „Sie fördern die Effizienz durch Zeitersparnis."

14 Vgl. Christian Spannagel, „ChatGPT und die Zukunft des Lernens: Evolution statt Revolution", *Hochschulforum Digitalisierung*, 2023, in [https://hochschulforumdigitalisierung.de/de/blog/chatgpt-und-die-zukunft-des-lernens-evolution-statt-revolution] (Zugriff: 30.09.2023).

Es gab weitere ähnliche Hinweise zu Effizienzsteigerung, Hilfe bei der Ideengenerierung und mehr Freude beim Arbeiten.

Die Eindrücke der Studierenden decken sich auch mit den Ergebnissen einer MIT-Studie aus Juli 2023, die in einem beruflichen Kontext durchgeführt wurde.[15] Diese zeigte, dass sich die Zeit, die die Arbeitnehmer:innen für die Erledigung der ihnen gestellten Aufgaben benötigten, durch den Zugriff auf den unterstützenden Chatbot ChatGPT um 40 Prozent verringerte. Die Qualität der Ergebnisse stieg laut Aussage der Autor:innen derweil um 18 Prozent. Auch die Auswertungen einer experimentellen Studie, die in Kooperation mit der Boston Consulting Group entstanden ist, zeigten, dass Mitarbeitende mit Zugriff auf ChatGPT unter Verwendung des GPT-4-Modells die ihnen zugewiesenen Aufgaben nicht nur schneller, sondern auch in höherer Qualität erledigten als Kolleg:innen, die keinen KI-Zugriff hatten. Dies allerdings unter der Voraussetzung, dass zwar komplexe, realistische Aufgaben ausgewählt wurden, diese 18 unterschiedlichen Aufgaben sich aber auch im Korridor des momentan für KI Leistbaren bewegten.[16]

Die Effekte der Qualitätssteigerung als auch der Produktivitätszunahme, die sowohl beim akademischen Schreiben wahrgenommen als auch im beruflichen Kontext von Akademiker:innen aufgezeigt wurden, mögen eine Erklärung dafür sein, dass auch wissenschaftliche Suchmaschinen und Verlage die KI-Textgeneratoren reagieren und z. B. KI-gestützte, dialogorientierte wissenschaftliche Suche anbieten.[17]

Konkrete praktische Anwendungshinweise zu KI-Nutzung beim wissenschaftlichen Arbeiten finden sich in einem Arbeitsblatt des Schreibzentrums der Goethe-Universität, das im Mai 2023 veröffentlicht wurde.[18] Hierin wird eine „plausible Aufteilung menschlicher und KI-Textproduktion beim akademischen Schreiben" aufgezeigt. Konkret wird hier darauf eingegangen, bei welchen Aufgaben KI-Systeme konkret unterstützend wirken können und welche unerlässlich durch Menschen ausgeführt werden sollten. Der Schreibprozess unterteilt sich demnach in fünf Phasen, von der „Findungsphase" über die „Datenerhebungs-/Bearbeitungs-

15 Vgl. Shakked Noy/Whitney Zhang, „Experimental evidence on the productivity effects of generative artificial intelligence", *Science*, Bd. 381(6654), 2023, 187–192.

16 Vgl. Fabrizio Dell'Acqua et al., *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality*, Harvard Business School Technology & Operations Mgt. Unit Working, Paper No. 24–013, in [ssrn.com/abstract=4573321] (Zugriff: 28.09.2023).

17 Vgl. Richard Van Noorden, „ChatGPT-like AIs are coming to major science search engines", *Nature*, Bd. 620(7973), 2023, 258, in [nature.com/articles/d41586-023-02470–3] (Zugriff: 12.08.2023).

18 Vgl. Schreibzentrum der Goethe-Universität Frankfurt a. M., *Nutzung von KI-Schreibtools durch Studierende*, 2023, in [https://www.starkerstart.uni-frankfurt.de/133460941/6-030_KI-Tools_pdf.pdf] (Zugriff: 20.09.2023).

phase“, die „Formulierungsphase“ und die „Überarbeitungsphase“ bis hin zur „Fertigstellungsphase“. In der „Formulierungsphase“ beispielsweise können KI-Systeme zur „Ausformulierung von Stichpunkten“ oder der „Weiterentwicklung von Textfragmenten“ genutzt werden, während die Grundlagen hierfür aber vom Menschen selbst formuliert werden sollten, ebenso wie die Beurteilung und Weiterentwicklung des generierten Outputs als menschliche Handlung verbleiben sollten.

Die Beurteilung der Qualität des Outputs ist deshalb von besonderer Relevanz, da wie bei KI-basierten Suchmaschinen auch, nach wie vor Halluzinationen auftreten könnten.[19]

Die Schwierigkeit im Umgang mit Halluzinationen ist eine der zentralen Herausforderungen und Diskussionspunkte rund um den Zusammenhang von generativer KI und wissenschaftlichem Arbeiten. Deshalb wird im kommenden Abschnitt das Thema der Wahrheit im Zeitalter von KI genauer beleuchtet.

4 Wahrheit im Zeitalter von KI: Herausforderungen und Perspektiven

In einer Zeit, in der KI-Systeme zunehmend in der Lage sind, Informationen zu generieren, zu verarbeiten und zu verbreiten, drängt sich die Frage auf, wie Wahrheit in diesem digitalen Zeitalter definiert und erkannt werden kann. Um dieses komplexe und faszinierende Thema zu beleuchten, wird in diesem Beitrag der Wahrheitsbegriff des Soziologen Niklas Luhmann genauer betrachtet und in den Kontext einer sich anbahnenden „Infocalypse“[20] durch Künstliche Intelligenz gestellt. Luhmanns Denken über die soziale Konstruktion von Wahrheit bietet eine weitreichende theoretische Grundlage, um die Auswirkungen von KI-Anwendungen auf das gesellschaftliche Verständnis von Wahrheit zu erforschen. Zugleich ist die „Infocalypse“ eine vermeintlich ernsthafte Bedrohung für die Verarbeitung und Verbreitung von Informationen, da sie das Vertrauen in die Wahrheit und Integrität von Informationen gefährden könnte.

Der Ausdruck „Reality Apathy“ beschreibt das Phänomen des beinahe vollständigen Desinteresses an Ereignissen und Informationen, die außerhalb der eigenen Lebenswelt oder persönlichen Erfahrungssphäre stattfinden. Dieses Desinteresse wird durch die zunehmende Schwierigkeit verstärkt, zwischen echten und

19 Ebd.

20 Vgl. Aviv Ovadya, „The Terrifying Future of Fake News“, *Buzzfeed News*, 2023, in [buzzfeednews.com/article/charliewarzel/the-terrifying-future-of-fake-news] (Zugriff: 21.09.2023).

gefälschten Informationen zu unterscheiden. In einer Welt, in der die Verbreitung von Fehlinformationen und gefälschten Inhalten rapide zunimmt, verlieren die Menschen das Vertrauen in Nachrichten und Informationen, was zu einer Art Gleichgültigkeit gegenüber der Realität führt.[21]

Die Konsequenzen dieses Phänomens können die Grundlagen einer funktionalen Demokratie, wie ein informiertes Bürgertum und eine fundierte Nachrichtenberichterstattung, potenziell untergraben. Wenn die Menschen aufhören, Nachrichten und Informationen aufmerksam zu verfolgen, kann dies die Stabilität der Gesellschaft gefährden und zu einer verstärkten Fragmentierung führen.[22] Luhmann betont, dass die Suche nach Wahrheit nicht die Hauptaufgabe des Mediensystems ist. Das bedeutet, dass Medien nicht primär dazu da sind, absolute Wahrheiten zu vermitteln.[23] Journalisten wählen aus einer Vielzahl von Fakten, Ereignissen und Quellen aus, um eine Geschichte zu erstellen. Diese Auswahl und Darstellung erfolgt nach bestimmten Kriterien wie Aktualität, Relevanz, Dramatik und Leseinteresse. Journalistische Erkenntnis ist daher nach Luhmann eine Art Filterung und Verdichtung von Erkenntnissen, die oft stark von den Zielsetzungen des Mediums und den Erwartungen des Publikums beeinflusst werden. Konträr zum Journalismus arbeitet die Wissenschaft nach dem Prinzip von wahr/unwahr, wodurch die Suche nach Wahrheit als zentrale Aufgabe wissenschaftlicher Kommunikation gekennzeichnet ist.[24] Dieser Code steht jedoch in deutlichem Kontrast zum Code des Journalismus, in dem die Vermittlung absoluter Wahrheiten nicht vorrangig ist. Somit wird klar, dass sich einige Erkenntnisse aus der Forschung und Wissenschaft schneller verbreiten als andere.

Krisen gelten nach den Sozialwissenschaftler:innen Folkers und Lim u. a. als Moment der Wahrheit. Eine Krise mag demnach Einblick in Aspekte der Gesellschaft liefern, die zuvor unter der Oberfläche lagen und unbemerkt blieben. Oder eine solche soziale Wahrheit mag auch erst dann konstituiert werden, wenn eine Krise offiziell thematisiert wird. Mit anderen Worten: Es ist fraglich, ob eine Krise etwas enthüllt, das schon immer in unserer Gesellschaft vorhanden war, aber aufgrund von Unaufmerksamkeit oder anderen Faktoren übersehen wurde, oder ob die Benennung einer Situation als Krise dazu führt, dass wir diese Situation nun als Wahrheit akzeptieren, die zuvor nicht so betrachtet wurde.[25] Die Krise stelle demnach eine Art Gegenüber zur Ordnung dar und könne helfen, die Ordnung

21 Ebd.

22 Ebd.

23 Vgl. Niklas Luhmann, *Die Realität der Massenmedien*, Opladen 19962, 56.

24 Ebd., 170.

25 Vgl. Andreas Folkers/Il-Tschung Lim, „Irrtum und Irritation. Für eine kleine Soziologie der Krise nach Foucault und Luhmann“, *Behemoth – A Journal on Civilisation*, Bd. 7, 2014, 48–69, 50.

besser zu verstehen, indem sie auf Dinge aufmerksam macht, die normalerweise verborgen oder unbeachtet seien. Diese werden normalerweise erst sicht- und analysierbar, wenn sie durch Probleme oder Bedrohungen in ihrer Integrität gefährdet werden.[26]

Der Unternehmer Elon Musk hat sich zu seinem geplanten LLM-Projekt namens „xAI" geäußert, welches eine Konkurrenz zu ChatGPT und den GPT-Modellen von OpenAI darstellen soll. Die von Musk bevorzugte KI-Anwendung soll auf einem KI-Modell basieren, das nach „der Wahrheit" strebt, eine Art „TruthGPT", wobei die genauen Funktionsweisen noch nicht klar sind.[27] Sein bevorzugtes Konzept eines LLMs sei vor allem der Idee „der Wahrheit" verpflichtet.[28] Bei genauerer Betrachtung der Aussage Musks wird deutlich, dass er von einer einzigen, absoluten, binären Wahrheitsvorstellung ausgeht. Diese zugespitzte Definition von Wahrheit lässt im Vergleich zu den genannten Wahrheitsbegriffen von Luhmann, Folkers und Lim den Prozess der Annäherung und Konstruktion weg. Diese Reduktion durch Elon Musk kann als seine Verdichtung und Filterung interpretiert werden, die persönliche philosophische oder religiöse Verständnisse aufdeckt, jedoch nicht den wissenschaftlichen Begriff von faktenbasierter Wahrheit verfolgt. Aus der geschilderten sozialwissenschaftlichen Perspektive liefert sie lediglich eine Interpretation des Wahrheitsbegriffes, der im Rahmen einer gesellschaftlich akzeptierten Konvention existieren kann. Die Stärke dieses Rahmens liegt darin, dass er durch gesellschaftliche Interpretation korrigiert werden kann und im Gegensatz zu Musks Aussage keine Singularität im Wahrheitsbegriff durch eine Person oder ein normatives System beansprucht. Ein möglicherweise drohendes Szenario, bei dem gesellschaftliche Ebenen durch die Konstruktion des Wahrheitsbegriffs ausgeschlossen werden und dadurch eine generelle Ablehnung gegenüber der Wahrheit entstehen kann, sei folgendermaßen geschildert: Eine singuläre, „anspruchsvolle" Wahrheit könnte in Zukunft durch ein KI-System gewährleistet werden, welche nirgendwo sonst in der Gesellschaft erzeugt werden kann. Dies ermöglicht einen eher „rücksichtslosen" Ansatz bei Fragen der Wahrheit, insbesondere ohne Anerkennung von alltäglichen Plausibilitäten und ohne Berücksichtigung religiöser oder politischer Konventionen. Letztere würden ersetzt werden durch die Gewichtungs- und Kostenfunktion des KI-Modells sowie dessen technischen Charakteristika. Im

26 Vgl. Jürgen Habermas, *Theorie des kommunikativen Handelns*, Bd. 2, *Zur Kritik der funktionalistischen Vernunft*, Frankfurt a. M. 1981, 593.

27 Vgl. APA/Reuters, „TruthGPT: Musk plant maximal wahrheitssuchende KI", 2023, in [diepresse.com/6277020/truthgpt-musk-plant-maximal-wahrheitssuchende-ki] (Zugriff: 20. 09. 2023).

28 Vgl. Parker Molloy, „Vaporware King Elon Musk's xAI is Basically Just a 2023 Version of Microsoft's Tay", *The Present Age*, 2023, in [www.readtpa.com/p/vaporware-king-elon-musks-xai-is] (Zugriff: 20. 09. 2023).

Zuge dessen kann das menschliche Denken als mehr als nur logisches Problemlösen interpretiert werden. Anders als bei zweckrationaler Intelligenz kommt es nicht nur darauf an, was als faktisch oder wahr definiert wird, sondern vor allem darauf, wie die Gesellschaft die Definition interpretiert und wertet. Niklas Luhmann argumentierte,[29] dass Wahrheit im Wesentlichen das Ergebnis von Kommunikationsprozessen innerhalb sozialer Systeme sei. Innerhalb dieser Prozesse würden Informationen ausgewählt, interpretiert und akzeptiert, was dazu führe, dass Menschen unterschiedliche Vorstellungen von der Realität und Wahrheit entwickeln. Diese Vorstellungen von Realität seien jedoch nicht fest in Stein gemeißelt, sondern flexibel und abhängig von den Normen und Regeln, die in einem bestimmten sozialen System gelten. In dieser Entwicklung spiegelt sich auch die Systemtheorie von Niklas Luhmann wider.[30] Sie betont, dass in der modernen Gesellschaft verschiedene Teilsysteme existieren, die in gewisser Weise autark sind und unterschiedliche Regeln und Logiken haben. Die Nutzung von generativen KI-Systemen zur Wahrheitsfindung könnte demzufolge auch als ein eigenständiges Teilsystem betrachtet werden, das seine eigenen Verfahren und Maßstäbe hat, die sich von denen anderer Teilsysteme wie Religion, Journalismus, Forschung oder Politik unterscheiden. Die oben beschriebene Aussage von Elon Musk gestaltet darin ein theoretisches Szenario, in dem die Wahrheitsfindung durch KI-Systeme alle gesellschaftlichen Teilbereiche dominieren würde. Soziologische Theorien beschreiben die Digitalisierung als Antwort auf die immer größere Komplexität der Moderne.[31] Je unübersichtlicher das Thema, desto wichtiger wird eine intelligente Mustererkennung. Der KI-Forscher Pedro Domingos von der University of Washington treibt es sogar noch weiter und behauptet, dass die Wissenschaft ohne Computer überhaupt keine grundlegenden Fortschritte mehr gemacht hätte.[32]

Ein weiterer Ansatz besteht darin, die Antworten der KI-Modelle durch menschliches Feedback zu verbessern, bei dem deren Wahrheitsgehalt beurteilt wird.[33] Doch die Tatsache, dass GPT-4 gefälschte Artikelzitate generieren kann, verdeutlicht die Notwendigkeit, menschliche Überprüfungen und Qualitätskontrollen in den Einsatz von KI-gesteuerten Tools einzubeziehen, um Fehlinformationen und Datenverzerrungen zu vermeiden. Für Fachgebiete mit eng gefassten

29 Vgl. Niklas Luhmann, *Die Wissenschaft der Gesellschaft*, Frankfurt a. M. 1992, 22–23.

30 Vgl. ebd., 355–358.

31 Vgl. Jutta Rump/Silke Eilers, „Im Fokus: Digitalisierung und soziale Innovation. Konsequenzen für das System Arbeit", in *Auf dem Weg zur Arbeit 4.0.*, hg. von dens., Berlin/Heidelberg 2017, 79–80.

32 Vgl. Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine will Remake our World*, New York 2015, 16.

33 Vgl. Eva Wolfangel, „Wie man Chatbots die Wahrheit beibringt", *Zeit Online*, 2023, in [https://t1p.de/kehsb] (Zugriff: 20.09.2023).

Themen ist besondere Vorsicht geboten, da hier das Risiko von Fehlern bei der Zitation von Referenzen erhöht ist.[34] Die wichtigsten Anwendungen von LLMs im wissenschaftlichen Kontext umfassen laut Antun Chen Textverbesserung, Textzusammenfassung, Textanalyse, Programmierung, Ideengenerierung und Textübersetzung. Die Expert:innen sehen in LLMs ein Potenzial zur Transformation wissenschaftlicher Praktiken, insbesondere im textbasierten Arbeiten, und eine Möglichkeit zur Zeitersparnis in administrativen Aufgaben. Die Experten bewerten Halluzinationen als die schwerwiegendste Einschränkung, gefolgt von Nicht-Transparenz und mangelnder Spezifität.

In allen drei Blickwinkeln wird die Herausforderung deutlich, dass die Verbreitung von Fehlinformationen durch generative KI-Systeme dazu führen könnte, dass KI-Modelle in Zukunft fehlerhaften Daten rezipieren. Dies hätte zur Folge, dass sowohl die technischen Charakteristika als auch die gesellschaftlichen Auswirkungen der Sprachmodelle in einer rekursiven und konvergierenden Weise verstärkt werden könnten, während sie zunehmend als ein iteratives Teilsystem der Wahrheit hineinwirken könnten. Ein Szenario ist, dass die Gesellschaft mögliche Fehlinformationen zunehmend übernehmen könnte, da Menschen aufgrund der Datenquantität die Aussagen nicht vollständig überprüfen können.[35] Andernfalls könnte die Gesellschaft sich aufgrund ihrer zunehmend geringeren Wirkungs- und Repräsentationsmacht expandierend vom Konzept der Wahrheit durch KI-Systeme abwenden. In unterschiedlichen Szenarien ist also zu betrachten, dass KI-Systeme zu einer weiteren Verschachtlung des Begriffs der Wahrheit führen könnten.

Hinzu kommt, dass der gegenwärtige Zustand dem Muster folgt, welches durch die moderne Differenzierung der Realität zunehmend mehr Meinungen anstelle von Wahrheit hervorbringt.[36] Das Weltbild des Bürgertums sieht sich heute von zwei Fronten bedroht: einem zunehmenden Konformitätsdruck und einer Oligarchisierung der Wahrheitsfindung von oben sowie der steigenden Anzahl von Fake News von unten. Im Sinne von Luhmann kann gesagt werden, dass technologische Entwicklungen sich zwischen Sender und Empfänger schieben und somit den Prozess der Wahrheitsbildung und -vermittlung beeinflussen. Dies geschieht, indem sie die Art und Weise verändern, wie Informationen erstellt, verarbeitet und übertragen werden. Dadurch kann der klare Bezug zur Realität und zur Wahrheit

34 Vgl. Anjun Chen, „Accuracy of Chatbots in Citing Journal Articles“, 2023, in [jamanetwork.com/journals/jamanetworkopen/fullarticle/2808058] (Zugriff: 22. 09. 2023).

35 Vgl. Jannis Brühl, „Chat-GPT wird benutzt, um Bullshit zu automatisieren“, *Süddeutsche Zeitung*, 2023, 15, in [sueddeutsche.de/wirtschaft/chatgpt-ki-kuenstliche-intelligenz-kapoor-1.5771936] (Zugriff: 24. 09. 2023).

36 Vgl. Ulf Daniel Ehlers, *Wissenschaftstheorien-Vorlesung*, 2013, in [slideshare.net/uehlers/wissenschaftstheorien-vorlesung] (Zugriff: 22. 09. 2023).

verwischt oder verzerrt werden. Es wird deutlich, dass die Herausforderung darin besteht, Wege zu finden, um das Vertrauen in authentische Informationen wiederherzustellen und die Menschen dazu zu ermutigen, sich aktiv mit der Realität auseinanderzusetzen, trotz der Schwierigkeiten im Umgang mit gefälschten Informationen.

5 Human-AI Hybrid: Chancen und Risiken für wissenschaftliche Diskurse

Seit drei Jahren ist der Begriff „Human-AI Hybrid" (auch umgekehrt als „Hybrid Human-AI" bekannt) in den Mittelpunkt des wissenschaftlichen Interesses gerückt.[37] Um den Begriff hat sich ein eigenes Forschungsfeld entwickelt, in dem die Wissenschaftler:innen das Ziel verfolgen, KI-Systeme zu entwickeln, die einerseits den Menschen unterstützen und andererseits durch menschliche Fähigkeiten ergänzt werden.[38] Der Fokus liegt somit auf der Schaffung adaptiver, kollaborativer, verantwortungsbewusster und interaktiver KI-Systeme, die menschenzentriert sind. Diese Systeme sollen die Stärken des Menschen hervorheben und seine Schwächen kompensieren, während sie soziale, ethische und rechtliche Aspekte berücksichtigen. Es geht demnach nicht um Mensch-Maschine-Hybride wie Cyborgs, sondern vielmehr um die Beziehung und Zusammenarbeit zwischen organischen Menschen und synthetischen Systemen. Wissenschaftliche Arbeiten, die diesen Begriff mit der genannten Bedeutung verwenden, erscheinen bereits seit mindestens 2020;[39] 2022 wurde zum ersten Mal eine internationale Konferenz in Amsterdam veranstaltet, die fortan jährlich in wechselnden Städten stattfindet.[40]

Die Entstehung des Begriffs lässt sich auf die jüngsten Fortschritte in den KI-Technologien zurückführen, insbesondere auf die Entwicklung von Systemen wie ChatGPT, Midjourney und deren generativen Verwandten. Diese Systeme haben

37 Vgl. z. B. Inge Molenaar, *Towards hybrid human-AI learning technologies*, 2022 in [onlinelibrary.wiley.com/doi/full/10.1111/ejed.12527] (Zugriff: 30. 09. 2023); Filipe Dwan Pereira et al., *Towards a Human-AI Hybrid System for Categorising Programming Problems*, 2021, in [dl.acm.org/doi/10.1145/3408877.3432422] (Zugriff: 30. 09. 2023); Kenneth Holstein et al., *A Conceptual Framework for Human–AI Hybrid Adaptivity in Education*, 2020, in [www.doi.org/10.1007%2F978-3-030-52237-7_20] (Zugriff: 30. 09. 2023).

38 Vgl. HHAI2023, „Call for Papers. Hybrid Human-Artificial Intelligence", 2023, in [https://hhai-conference.org/2023/cfp/] (Zugriff am10. 10. 2023).

39 Vgl. Chen et al. 2020.

40 Vgl. HHAI, „International Conference Series on Hybrid Human-Artificial Intelligence", 2023, in [https://hhai-conference.org/] (Zugriff am 10. 10. 2023).

gezeigt, dass KI-Anwendungen nicht nur dazu verwendet werden können, menschliche Aufgaben zu automatisieren, sondern auch, um menschliche Fähigkeiten zu erweitern und zu bereichern.

Eine übliche Kategorisierung von Mensch-Maschine-Hybriden hat sich noch nicht hervorgetan, weswegen im Folgenden eine solche Kategorisierung anhand der Rolle des KI-Systems gegenüber dem Menschen vorgeschlagen wird. Diese Rollen lauten „Dominator", „Facilitator" und „Operator". Um diese Kategorien zu veranschaulichen, werden sie mithilfe von Beispielen aus der Forschung bzw. Wissenschaft erläutert.

Tabelle 1: Kategorisierung des Human-AI-Hybrid anhand der Rolle des KI-Systems gegenüber dem Menschen.

Rolle	Erklärung	Beispiel
Dominator	In dieser Kategorie dominiert das KI-System das Zusammenspiel und trifft Entscheidungen eigenständig, basierend auf den ihm zur Verfügung stehenden Daten. Hierbei agiert der Mensch vorwiegend als Beobachter oder Korrektor, hat demnach wenig bis gar keinen direkten Einfluss auf die Aktionen der KI. Bezogen auf die Wissenschaft würde es bedeuten, dass in dieser Kategorie die KI die Hauptrolle im Forschungsprozess übernimmt, indem sie Daten autonom analysiert und Interpretationen anbietet. Wissenschaftler:innen fungieren hier hauptsächlich als Beobachter oder Korrektoren, um die Ergebnisse zu überprüfen und gegebenenfalls zu korrigieren.	Ein KI-System analysiert autonom große Datensätze von astronomischen Beobachtungen, um neue Exoplaneten zu identifizieren. Die Wissenschaftler:innen überprüfen die Ergebnisse und nehmen gegebenenfalls Korrekturen vor, haben aber wenig direkten Einfluss auf den Entdeckungsprozess.
Facilitator	In der Kategorie „Facilitator" arbeiten Mensch und KI-System Hand in Hand, wobei beide Parteien gleichberechtigt Entscheidungen treffen und Aktionen durchführen. Es handelt sich um eine echte Partnerschaft, in der die Stärken und Schwächen beider Parteien berücksichtigt werden. Hier arbeiten KI und Wissenschaftler:innen also gemeinsam daran, Forschungsfragen zu untersuchen und zu beantworten. Die Partnerschaft ermöglicht es, die Stärken beider Parteien zu nutzen und zu einer	Ein KI-System arbeitet mit Geowissenschaftler:innen zusammen, um Erdbebenrisiken zu analysieren. Das System liefert eine vorläufige Analyse der seismischen Daten, die Wissenschaftler:innen bringen ihre Expertise in die nachfolgende Interpretation der Ergebnisse ein. Gemeinsam entwickeln sie ein umfassendes Verständnis der Erdbebenrisiken in einer bestimmten Region.

Tabelle 1: Kategorisierung des Human-AI-Hybrid anhand der Rolle des KI-Systems gegenüber dem Menschen. *(Fortsetzung)*

Rolle	Erklärung	Beispiel
	fundierten Entscheidung oder Interpretation zu gelangen.	
Operator	Hier fungiert das KI-System als Unterstützer:in oder Assistenz des Menschen. Es bietet Vorschläge, Analysen und Einblicke, basierend auf seinen Algorithmen und Daten, während der Mensch die endgültige Entscheidung trifft. In dieser Kategorie agiert die KI als Unterstützer:in der Wissenschaftler:innen, indem sie Datenanalysen und Interpretationssupport anbietet, aber die endgültigen Schlussfolgerungen den Wissenschaftler:innen überlässt.	Ein KI-System unterstützt Chemiker:innen bei der Analyse von Molekülstrukturen und bietet Vorschläge für mögliche neue Synthesewege. Die endgültige Entscheidung über die zu verfolgenden Synthesewege und die Interpretation der Ergebnisse liegt jedoch bei den Wissenschaftler:innen.

Diese Einteilung spiegelt die verschiedenen Grade der Interaktion und Integration zwischen Mensch und KI wider und wird daher als Einteilung für den weiteren Diskurs im Bereich des Human-AI Hybrids empfohlen.

Die Einteilung des Human-AI Hybrids in die Kategorien „Dominator", „Facilitator" und „Operator" wirft eine Reihe von ethischen, sozialen und wissenschaftlichen Fragen auf. Sie impliziert beispielsweise, dass in der „Dominator"-Kategorie die Gefahr bestünde, dass KI-Systeme Entscheidungen treffen, die ethisch fragwürdig oder sogar gefährlich sein könnten. Wenn z. B. ein KI-System, das medizinische Diagnosen stellt, ohne ausreichende menschliche Überprüfung stellt, könnte dies zu einem fehlerhaften Behandlungsplan führen. In den Kategorien „Facilitator" und „Operator" liegt die ethische Verantwortung mehr beim Menschen, aber auch hier gibt es Fallstricke, wie die mögliche Verzerrung von Daten oder Diskriminierung. Die Integration von KI-Systemen in den menschlichen Entscheidungsprozess könnte ebenfalls zu sozialen Veränderungen führen, insbesondere in Bezug auf die Rolle von Fachleuten. In der „Operator"-Kategorie könnte der Einsatz von KI-Assistenten dazu führen, dass weniger qualifizierte Personen in der Lage sind, komplexe Aufgaben zu erfüllen, was die Rolle von Experten verändert. Im „Facilitator"-Modus könnten KI-Systeme als gleichwertige Partner angesehen werden, was die soziale Dynamik in Teams und Organisationen verändern könnte.

In einem wissenschaftlichen Kontext könnte die Möglichkeit, KI in Diskursen als „Facilitator" oder „Operator" einzusetzen, neue Wege für die Forschung bieten. KI-Systeme könnten dazu verwendet werden, wissenschaftliche Daten zu analysieren, Hypothesen zu generieren oder sogar bei der Formulierung von For-

schungsfragen zu helfen. Allerdings ergeben sich auch Fragen zur Validität und Reproduzierbarkeit von KI-generierten Daten und Erkenntnissen.

Wird ChatGPT mit dem Modell GPT-4 selbst dazu befragt, welche Implikationen sich aus generativen KI-Systemen für den wissenschaftlichen Diskurs ergeben, so gibt es folgende Antwort: „Generative KI-Systeme, wie GPT-4, haben das Potenzial, den wissenschaftlichen Diskurs in vielerlei Hinsicht zu beeinflussen […].“[41] Es fährt fort mit der Nennung einiger Implikationen, die sich teils mit den zuvor beschriebenen decken. Beispielsweise könne generative KI einerseits dabei helfen, das Forschungstempo zu beschleunigen (s. auch Abschnitt 2). Andererseits steige auch das Risiko, dass sich Fehlinformationen verbreiten, die „insbesondere dann problematisch werden, wenn diese Inhalte als wissenschaftliche Tatsachen präsentiert werden.“ (s. auch Abschnitt 2, Risiken von KI-Sprachmodellen, sowie Abschnitt 4). Wie bereits in Abschnitt 2 erwähnt wurde, ist mit KI-Sprachmodellen auch das Risiko einer kriminellen oder absichtlich irreführenden Nutzung verknüpft. Auch auf dieses Risiko weist der Chatbot im wissenschaftlichen Zusammenhang hin. Nicht nur Fragen zu „Urheberschaft und Anerkennung von Beiträgen“ könnten auftreten, generative KI könnte auch zur gezielten Produktion von falschen Studienergebnissen genutzt werden.

Zusätzlich zu den in diesem Beitrag bereits aufgezeigten Implikationen hält ChatGPT aber auch weitere Ideen für Implikationen bereit. So könnten „Generative Systeme […] dazu verwendet werden, wissenschaftliche Erkenntnisse in einer leicht verständlichen Form für ein breiteres Publikum zu präsentieren.“ Dies könnte die Leistungsfähigkeit und Geschwindigkeit der Wissenschaftskommunikation erhöhen. In diesem Zusammenhang wird ebenfalls der Begriff der „Zugänglichkeit“ genannt. Die Systeme könnten „den Zugang zu wissenschaftlichen Informationen erleichtern“ und insofern zur Demokratisierung in der Wissenschaft beitragen. Ein weiterer Vorteil, den ChatGPT für den wissenschaftlichen Diskurs bei der Nutzung von KI-Systemen sieht, ist ein kollaborativer Ansatz. Durch die Nutzung generativer KI könnten Perspektiven anderer Forschungsdisziplinen direkt mit in die eigenen Forschungserkenntnisse einfließen, ohne auf tatsächlichen Kontakt oder Austausch mit Expert:innen aus dem betreffenden Feld angewiesen zu sein.

Am Ende des Dialogs weist der Chatbot noch auf Folgendes hin:

> Schließlich, wie bei jeder Technologie, hängen die tatsächlichen Auswirkungen von generativen KI-Systemen auf den wissenschaftlichen Diskurs von ihrer Anwendung und Implementierung ab. Es ist wichtig, sowohl die Vorteile als auch die potenziellen Risiken zu berück-

41 Der gesamte Chatverlauf kann hier eingesehen werden: https://chat.openai.com/share/43a68546-5ea9-4eae-9673-fcfb35f4f42d, generiert am 09.08.2023.

sichtigen und Richtlinien und Best Practices zu entwickeln, um sicherzustellen, dass sie ethisch und effektiv eingesetzt werden.

Entsprechend dieser Aussage wurden im Laufe des Beitrags im Kontext des Human-AI Hybrids sowohl Chancen als auch Herausforderungen aufgezeigt. Die genannten Implikationen führen zu weiteren entscheidenden Fragen bezüglich der Verantwortung und möglicher Auswirkungen auf unsere Informationslandschaft (vgl. auch Abschnitt 4).

Die Frage, ob gemäß dem Begriff in Abschnitt 4 eine realitätsverfälschende „Infokalypse“ bevorsteht oder ob Wissenschaftler:innen bzw. die Gesamtgesellschaft sich bereits in dieser befinden, bleibt noch offen. Die Bewältigung der Herausforderungen, die durch KI-Systeme auch im wissenschaftlichen Kontext entstehen, wird vermutlich eine breite interdisziplinäre Zusammenarbeit und ein kollektives Engagement erfordern.

6 Schlussfolgerungen und Ausblick

Im vorliegenden Beitrag wurden einige der Einsatzmöglichkeiten und Auswirkungen der KI auf den gesellschaftlichen und wissenschaftlichen Diskurs untersucht. Durch das Aufzeigen unterschiedlicher Perspektiven und Konzepte wurde deutlich, dass die rasante Entwicklung von KI-Systemen die Art und Weise, wie wir Wahrheit definieren und erkennen, grundlegend verändert und dabei ethische, soziale und wissenschaftliche Fragen entstehen.

Die Diskussion um die Konzeption von Wahrheit in der KI-Ära offenbart die Problematik einer alleinigen Abhängigkeit von technologischen Systemen zur Wahrheitsfindung. Die Betrachtung von z. B. Elon Musks LLM-Projekt xAI und die Analyse des Human-AI Hybrids verdeutlichen die Notwendigkeit einer ausgewogenen Mensch-Maschine-Interaktion, um die Stärken beider zu nutzen. Die Einteilung der Mensch-Maschine-Hybriden in „Dominator“, „Facilitator“ und „Operator“ bietet eine mögliche Grundlage für die weitere Erforschung und Entwicklung von KI-Systemen, die auf eine harmonische Zusammenarbeit mit menschlichen Akteuren abzielen. Dabei ist es von zentraler Bedeutung, ethische Richtlinien und Best Practices zu etablieren, um die Integrität der Wahrheitsfindung und die soziale Verantwortung zu gewährleisten.

In Zukunft könnte die Weiterentwicklung von KI-Systemen, insbesondere im Bereich der generativen Modelle, neue Horizonte für die wissenschaftliche Forschung und Praxis eröffnen. Die Möglichkeit, KI als „Facilitator“ oder „Operator“ in wissenschaftlichen Diskursen einzusetzen, könnte die Forschung beschleunigen und zu interdisziplinären Entdeckungen führen. Gleichzeitig ist es unerlässlich, die

Risiken von Fehlinformationen und Datenverzerrungen ernst zu nehmen und Mechanismen zur Qualitätssicherung und menschlichen Überprüfung zu implementieren.

Es bleibt viel Raum für weitere Forschung, insbesondere im Hinblick auf die langfristigen Auswirkungen dieser Technologien auf die Gesellschaft und die Wissenschaft. Zukünftige Studien könnten sich auf die Entwicklung von Best Practices für die Ethik und Governance von Human-AI Hybrids konzentrieren, um sicherzustellen, dass diese Technologien auf eine Weise eingesetzt werden, die dem Gemeinwohl dient.

Autor:innen-Erklärung

Für diese Ausarbeitung hat das Team der Autor*innen folgende Werkzeuge aus dem Bereich generativer KI bzw. KI-gestützter Schreibwerkzeuge genutzt:

- ChatPDF zur Ideengenerierung.
- ChatGPT zur Ideengenerierung, für erste Entwürfe und als Formulierungshilfe.

Nach der Nutzung dieser Tools/Dienste haben die Autor:innen den Inhalt nach Bedarf überprüft und bearbeitet und übernehmen die volle Verantwortung für den Inhalt der Veröffentlichung.