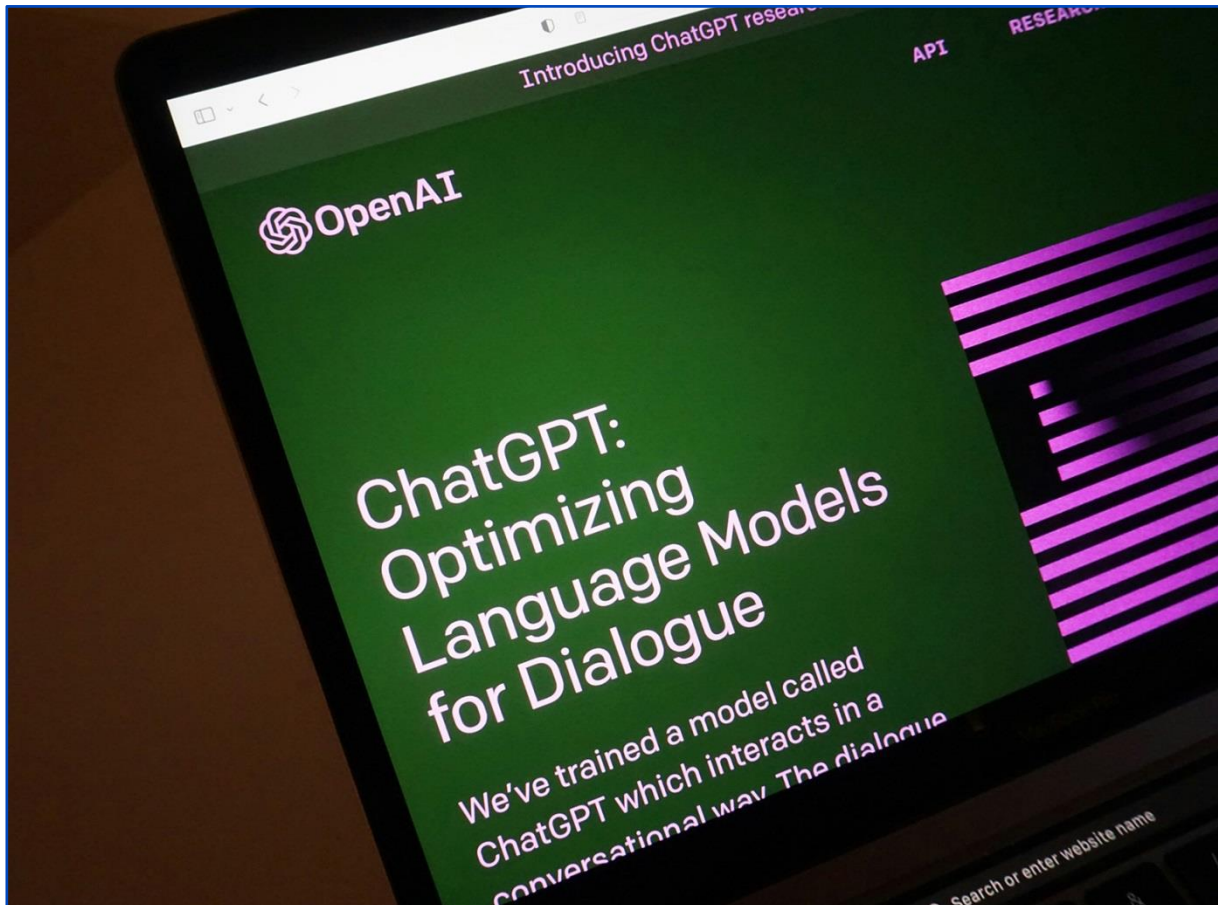


GENERATIVE KÜNSTLICHE INTELLIGENZ



© unsplash.com/Rolf van Root

ZUSAMMENFASSUNG

Seit ChatGPT im November 2022 veröffentlicht wurde, sorgen Generative KI-Systeme für Aufsehen und alte Träume, Hoffnungen und Ängste in Bezug auf die Leistungsfähigkeit Künstlicher Intelligenz (KI) werden ebenso angefacht wie ganz konkrete Befürchtungen zu den Auswirkungen auf demokratische Gesellschaften. Generative KI verspricht enorme Arbeitserleichterungen und uneingeschränkten Zugang zu Informationen. Bei genauerer Betrachtung offenbaren sich jedoch viele nach wie vor ungelöste Schwierigkeiten, die einen Einsatz in den meisten professionellen Umgebungen nicht ratsam erscheinen lassen. Zusätzlich sehen Expert:innen ein großes Sicherheitsproblem auf die Gesellschaft zukommen, da Generative KI-Systeme bspw. auch genutzt werden könnten, um tausendfach leicht variierten Schadcode zu schreiben oder auf Millionen Bürger:innen individuell zugeschnittene Betrugsversuche zu starten.

*Generative KI –
Hype, Gamechanger
oder Gefahr für die
Gesellschaft?*

ÜBERBLICK ZUM THEMA

Die Möglichkeiten, Verständnis und Denken zu simulieren, haben sich seit der Aufklärung deutlich verbessert, als 1769 in Wien der „Schachtürke“, ein angeblich selbst denkender Schachautomat, in dessen Innerem ein menschlicher Spieler verborgen war, vorgestellt wurde. Als Vorgänger der aktuellen generativen KI-Systeme kann man Joseph Weizenbaums Programm ELIZA aus dem Jahr 1966 sehen, das über verschiedenen Skripte unterschiedliche Gesprächspartner darzustellen vermochte. Die Idee denkender, intelligenter Maschinen war immer mit hohen Erwartungen, Träumen und Utopien verbunden, aber auch mit bedrohlichen Szenarien und Ängsten. Es ging oft darum, den Menschen mächtiger zu machen, seine „Makel“ auszugleichen, bis hin zu seiner Rettung, wenn wir uns aktuelle Vorstellungen davon ansehen, wie KI-Systeme uns von den Folgen des Klimawandels befreien sollen.¹ Ist das jetzt die Generation Künstlicher Intelligenz, auf die seit ELIZA alle warten? Oder ist es, wie Bareis und Katzenbach (2021) schreiben, mehr „Talking AI into Being“, also mehr das Narrativ als die tatsächlichen Möglichkeiten Künstlicher Intelligenz, die den Hype befeuern?

Den Wunsch nach denkenden, dienstbaren Maschinen gibt es schon lange

Eng verbunden mit den Überlegungen zur Nützlichkeit denkender Systeme ist die Frage nach dem Vertrauen. Vertrauen wir Maschinen, wenn es um lebenswichtige Entscheidungen geht? Zugleich neigen Menschen dazu, in Technik menschliche Wesenszüge zu erkennen. Das wird nicht zuletzt in unserem Umgang mit digitalen Assistenten sichtbar (Schaber et al., 2019). Maschinen, besonders Rechenmaschinen und Computern, wird unterstellt, dass sie sich nicht irren könnten, somit ihre Ergebnisse zweifelsfrei richtig und – übertragen auf KI-Systeme – auch ihre Entscheidungen richtig seien.² Das liegt daran, dass Menschen zu wenig nachvollziehen können, wie diese Systeme funktionieren, und deshalb vielleicht vertrauen, wo kein Vertrauen angebracht ist. Dazu muss man sich auch vor Augen halten, dass diese Systeme von Menschen entworfen und gebaut werden, und KI mit bestimmten Daten trainiert wird, die auch nicht neutral sind. Wir wissen heute, dass Technik allgemein nicht neutral ist. Sie trägt die Werte und Wertvorstellungen ihrer Entwickler:innen in sich. Diese Erkenntnis ist ganz besonders dann relevant, wenn es um Automation und KI geht. Diese Systeme reproduzieren die (diskriminierenden) Strukturen, von denen sie (selbst wieder) lernen.

Wem vertrauen wir?

Es gibt auch warnende Stimmen, die davon ausgehen, dass eine lernende, sich selbst optimierende Maschine, in weiterer Folge mit eigenem Bewusstsein (vielleicht sogar einer Persönlichkeit), die Entwicklung der Intelligenz und denken der Maschinen derart beschleunigen würde, dass der Mensch in so einer Welt keinen Platz mehr hätte. Maschinen würden Menschen in allen Belangen übertrumpfen und sukzessive ersetzen. Diesen Punkt, an dem die Maschinen in allem weitaus besser wären als der Mensch, bezeichnet man als „technische Singularität“. (Siehe auch die Bücher von Kurzweil 2005, oder Bostrom 2016) Und wäre diese Vorstellung nicht schon bedrohlich genug, gibt es natürlich auch die Phan-

Befürchtungen rund um (Generative) KI-Systeme

¹ derstandard.at/story/2000142852407/wie-kuenstliche-intelligenz-im-kampf-gegen-die-klimakrise-hilft.

² Mehr Informationen zum sog. Automation Bias bspw. von Kate Goddard (2014).

tasien, die diese Maschinen als feindlich, dem Menschen böse gesinnt denken. Von diesen Entwicklungen sind wir beim aktuellen Stand der Forschung jedoch noch weit entfernt.

Die KI-Systeme, die seit der Vorstellung von Chat-GPT Menschen auf der ganzen Welt faszinieren, sind sog. Generative KI-Systeme, die über Texteingaben weitere Texte erstellen oder andere audiovisuelle Inhalte produzieren. Sie simulieren mit Wahrscheinlichkeiten und extrem großen Mengen an Trainingsdaten Intelligenz und Verständnis.³ Es werden statistische Methoden und (Trainings-)Daten, sowie die daraus gewonnen Informationen über Wahrscheinlichkeiten in einem stochastischen Prozess genutzt, um Anfragen zu bearbeiten und darauf Antworten zu entwickeln, die auf Grund statistischer Verteilungen und Wahrscheinlichkeiten Wort- oder Satzteile hintereinander reihen (oder Bildelemente, Programmcode zusammenfügen usw.). Dabei entstehen sinnvoll klingende Sätze, Bilder, Videos oder Teile eines Programmcodes, die sich auf die gestellte Aufgabe beziehen. Bei den Nutzer:innen entsteht dadurch der Eindruck, verstanden zu werden und eine inhaltlich passende Antwort erhalten zu haben. Das System hat jedoch zu keinem Zeitpunkt wirklich ein Verständnis entwickelt, nachgedacht oder darauf basierend eine Antwort formuliert. Wie Tests gezeigt haben, sind die Antworten auch oft nicht faktisch korrekt. Sie erscheinen mitunter frei erfunden, obwohl auch dieser kreative Akt eigentlich nicht stattfindet, sondern Produkt eines statistischen Zufalls ist.

Diese Funktionsweise ist für Menschen, die eher in deterministischen Strukturen denken, schwer nachzuvollziehen und lässt keine verlässliche Auskunft über z. B. wissenschaftliche Fragestellungen erwarten. Die Stringenz und der mittelmäßige Wahrheitsgehalt der generierten Antworten reichen jedoch aus, um für die meisten Menschen täuschend echt erscheinende Antworten bis hin zu Fake News zu produzieren. Hierin liegt eine große Gefahr für die Gesellschaft, deren Aushandlungsprozesse und die Demokratie, da von der Maschine produzierte Fake News nicht als solche erkannt werden, und der Manipulation der öffentlichen Meinung damit Tür und Tor geöffnet sind.

Auch wenn Sprache/Text schon zuvor missbräuchlich, manipulativ und/oder in Täuschungsabsicht verwendet wurde, und *auch Bilder und Videos gefälscht werden*, sind beide nach wie vor zentrale Basis unserer Kommunikation. Insbesondere Bildmaterial ist ein gefühlter Beleg für den Wahrheitsgehalt einer Behauptung. Expert:innen meinen, die Gesellschaft steuere hier auf eine „Sicherheitskatastrophe nicht gekannten Ausmaßes“ zu.⁴ Einerseits fehlt noch die Praxis im Umgang mit von Maschinen erstellten Inhalten, andererseits ist es auf Anwender:innenseite nahezu voraussetzungslos geworden, Falschnachrichten in hoher Qualität, Anzahl und/oder Individualität herzustellen. Auch in der Wissenschaft tauchen erste Publikationen auf, die nicht von Menschen erstellt wurden. Gelingt es nicht, diese zu „enttarnen“, hätte das katastrophale Auswirkungen für die Wissenschaft (Haider et al 2024). Darüber hinaus werden auf Generativer

Die Funktionsweise

...

*... und daraus
resultierende Gefahren*

*Weitere
Missbrauchsszenarien*

³ Beispiele für derartige Systeme sind: *Bard*, *ChatGPT* (Login erforderlich), *Dall-E2*, *Jasper*, *Kaiber*, *LaMDA* (Beschreibung) und *Midjourney*.

⁴ heise.de/hintergrund/Drei-Gruende-warum-KI-Chatbots-eine-Sicherheitskatastrophe-sind-8933941.html.

KI basierende Chatbots als kostengünstiger Ersatz für Psychotherapie eingesetzt (Kahwaja et al 2023) und um Gespräche mit Verstorbenen zu simulieren⁵ (siehe dazu auch *Soziale Agenten*).

Abgesehen von anderen Gefahren ist es auch demokratiepolitisch wichtig auf individueller Ebene zu wissen, ob man mit einem anderen Menschen oder eine Maschine kommuniziert – um deren Antworten einordnen zu können und zu wissen, wem man sich anvertraut. Auf Grund der Funktionsweise ist es nur für wenige – wenn überhaupt – nachzuvollziehen, welchen Bias Systeme durch Programmierung und Training erhalten. Im Spannungsfeld geopolitischer Wirtschaftsinteressen und Machtbestrebungen können KI-Systeme, die als Blackboxes funktionieren (Udrea et al., 2022), und eine Abhängigkeit von diesen, durch bewusste Manipulation der Ergebnisse auch Einfallstore für die Durchsetzung fremder Interessen sein, womit deren Einsatz auch aus Gründen der *digitalen Souveränität* zu hinterfragen wäre.

*Gefahren für
liberale Demokratien*

Vor allem die Unkenntnis über den Umgang mit den gesammelten Daten von ChatGPT, die sich durch die verpflichtende Registrierung bestimmten Personen zuordnen lassen, hat dazu geführt, dass die Anwendung derzeit bspw. in Deutschland von den Datenschutzbehörden näher analysiert wird.⁶ In China wurden 2023 Interimsregelungen für die Nutzung von KI erlassen, die zukünftig angepasst werden sollen.⁷ Auf EU-Ebene wurde die KI-VO um einen Passus zu „general purpose AI models“ ergänzt.⁸

Regulierung

Weil die Angst besteht, dass Schüler:innen und Studierende Aufsätze von der Software schreiben lassen, diese als ihre ausgeben und die Lehrenden den Betrug nicht bemerken könnten, wurde der Einsatz der Software, unabhängig von der Durchsetzbarkeit solcher Regeln, an manchen Schulen und Universitäten verboten. In Österreich wurde als Reaktion die Erstellung der sog. Vorwissenschaftlichen Arbeit für die Matura an einer AHS durch eine Abschlussarbeit ersetzt, die nicht textlich sein muss und hohe Auflagen in Bezug auf die Dokumentation über die Erstellung mit sich bringt.⁹ Dadurch wurde eine wichtige gesellschaftliche Diskussion über den Umgang mit Generativen KI-Systemen auf unbestimmte Zeit verschoben. An anderen Bildungseinrichtungen wiederum, werden die Studierenden zum Gebrauch der Software ermuntert, um die kritische Auseinandersetzung mit den Ergebnissen zu fördern.

*Generative KI
und Bildung*

Positive Nutzungsmöglichkeiten für solche Systeme ergäben sich (v. a. bei verlässlicher und transparenter Funktionsweise) u. a. in den Bereichen Journalismus/Recherche, Zusammenfassung von Texten, Kuratierung von Inhalten, in Werbung und Marketing, bei der Erstellung von Programmcode oder der *Diagnoseerstellung im Gesundheitswesen* (siehe auch Albrecht, 2023).

⁵ In der Dokumentation „Eternal You“ (2024) ausführlich dargestellt:

de.wikipedia.org/wiki/Eternal_You_%E2%80%93_Vom_Ende_der_Endlichkeit.

⁶ faz.net/aktuell/feuilleton/medien/datenschuetzer-haben-fragen-an-open-ai-18839095.html.

Von der Konferenz der unabhängigen Datenschutzaufsichtsbehörden in Deutschland gibt es mittlerweile einen ersten *Leitfaden* zum Einsatz von LLMs in Unternehmen.

⁷ nzz.ch/technologie/china-verabschiedet-weltweit-erste-regeln-fuer-chat-gpt-und-co-ld.1747327.

⁸ artificialintelligenceact.eu/article/51/.

⁹ orf.at/stories/3361802/.

Auch bspw. der Bereich generischer Bebilderung von Publikationen würde sich stark verändern, was neue Businesskonzepte für Bilddatenbanken erforderlich machen würde. Generell scheint die Zusammenarbeit von Mensch und Maschine der vielversprechendste Weg zu sein. Die dafür nötigen Kompetenzen müssten jedoch besser gestern als heute in den Bildungseinrichtungen vermittelt werden. In diese Richtung entwickeln sich auch Produkte und Dienstleistungen auf Basis Generativer KI. Es wird ausgelotet, wo KI den Menschen unterstützen kann, und wo sich damit Geld verdienen lässt.

Die ethischen Anforderungen an Generative KI-Systeme sind weitgehend mit denen an andere KI-Systeme vergleichbar. Es sollte der Mensch im Mittelpunkt stehen, Entscheidungen müssten transparent, nachvollziehbar, begründbar und reproduzierbar sein.¹⁰ Besonders bemängelt werden die Machtasymmetrien, die durch mangelnde Transparenz über Funktionsweise und Datenverwendung entstehen, sowie die Machtkonzentrationen durch große Datensammlungen und proprietäre Systeme von amerikanischen IT-Konzernen. Ebenso problematisch erscheinen die Arbeitsbedingungen, unter denen Menschen in Niedriglohnländern die Systeme trainieren, so wie der nicht nachhaltige Betrieb und der hohe Ressourcenverbrauch der Systeme.¹¹

*Neue Chancen
zur Vereinfachung
bestimmter
Arbeitsprozesse*

*Ethische
Anforderungen*

RELEVANZ DES THEMAS FÜR DAS PARLAMENT UND FÜR ÖSTERREICH

Um bei dem Thema nicht den Anschluss zu verlieren, ist eine regelmäßige Evaluierung des Umfangs und der Mittelverwendung im Bereich der nationalen Forschungsförderung für KI durchzuführen, um Fehlentwicklungen, die sich 2023 abzeichneten¹², zu vermeiden und die Stärken der österreichischen KI-Entwicklung weiter zu fördern. Dabei darf jedoch nicht außer Acht gelassen werden, dass bei allen Bemühungen, nationale Standortvorteile geltend zu machen, letztendlich nur ein abgestimmtes Vorgehen auf europäischer Ebene dazu führen kann, dass sich in Europa ein Gegengewicht zu den Entwicklungen in Nordamerika und China herausbilden und erhalten kann. Mindestens ebenso wichtig wie die Forschung an und die Entwicklung von KI-Systemen und deren Anwendungen ist die Forschung über KI und die damit verbundenen Veränderungen in Forschung und Gesellschaft. Hier wäre es dem Parlament möglich, durch entsprechende Schwerpunktsetzung bei Budgetierungen lenkend einzugreifen, und diesen sehr wichtigen Bereich zu unterstützen, der sehr zum Verständnis und zur Reflexion über die stattfindenden Veränderungen beiträgt, und letztendlich einen erfolgreichen Einsatz von KI-Systemen erst möglich macht.

*Schwerpunktsetzungen
in der
Forschungspolitik
durch das Parlament*

¹⁰ Stellvertretend für viele Stellungnahmen zu dem Thema sei hier die Stellungnahme des deutschen Ethikrates angeführt: Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz, ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf.

¹¹ Eine gute Zusammenfassung der Literatur dazu findet sich etwa im Themenpapier zu ChatGPT von TA-Swiss: ta-swiss.ch/chatgpt; Zusammenfassung zum Thema Wasserverbrauch: heise.de/news/Wasserbilanz-von-KI-Modellen-Halber-Liter-Wasser-pro-Unterhaltung-mit-ChatGPT-8973680.html.

¹² science.orf.at/stories/3218956/ und science.orf.at/stories/3218860/.

VORSCHLAG FÜR DAS WEITERE VORGEHEN

Neben der laufenden, vom Parlament beauftragten Studie zu den Auswirkungen von Generativen KI-Systemen auf die Demokratie in Österreich könnten Forschungsprojekte zu weiteren konkreten Aspekten angedacht werden, um die wissenschaftsbasierte Grundlage für zukünftige politische Entscheidungen vorzubereiten: zur Schwerpunktsetzung bei der Forschungsförderung; zu den Kriterien für einen gelingenden Einsatz von KI in der staatlichen Verwaltung; oder im Bildungsbereich. Darüber hinaus wäre es wichtig, die laufenden Bemühungen in Bildung und Forschung, bspw. die Projekte, die im Rahmen der Initiative AIM-AT gestartet wurden, und ihre Ergebnisse zeitnah zu evaluieren, um ggf. im Forschungs- und Bildungsbereich nachjustieren zu können. Für die Zukunft wird es wichtig sein, den Arbeitskräften in Österreich das nötige Wissen mitzugeben, um mit KI-Systemen zu arbeiten, diese produktiv zu nutzen, und deren Vor- und Nachteile besser beurteilen zu können.

*Neue
Forschungsprojekte*

*Bildung als Schlüssel
zur erfolgreichen
Nutzung von KI*

ZITIERTE LITERATUR

- Albrecht, St. (2023). ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen. TAB-Hintergrundpapier Nr. 26. publikationen.bibliothek.kit.edu/1000158070/150614893.
- Bareis, J., & Katzenbach, C. (2022). Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics. *Science, Technology, & Human Values*, Vol. 47 (Issue 5), pages 855–881.
- Bostrom, N. (2016). Superintelligenz. Szenarien einer kommenden Revolution. Suhrkamp, Frankfurt am Main.
- Goddard K, Roudsari A, Wyatt JC (2014). Automation bias: empirical results assessing influencing factors. *International Journal of Medical Informatics*. 2014 May; Vol 83 (Issue 5): pages 368-75. [doi: 10.1016/j.ijmedinf.2014.01.001](https://doi.org/10.1016/j.ijmedinf.2014.01.001).
- Haider, J., et al. (2024). GPT-fabricated scientific papers on Google Scholar: Key features, spread, and implications for preempting evidence manipulation. *Harvard Kennedy School Misinformation Review*.
- Helberger, N., Diakopoulos, N. (2023). ChatGPT and the AI Act. *Internet Policy Review* (2023) 12 (1). doi.org/10.14763/2023.1.1682.
- Khawaja Z, Bélisle-Pipon JC. (2023). Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots. *Front Digit Health*, 5:1278186. [doi: 10.3389/fdgth.2023.1278186](https://doi.org/10.3389/fdgth.2023.1278186).
- Kurzweil, R. (2005). *The Singularity Is Near. When Humans Transcend Biology*. Viking, New York.
- Schaber, F., Krieger-Lamina, J., Peissl, W. (2019). *Digitale Assistenten*. ITA-Bericht 2019-01. epub.oeaw.ac.at/ita/ita-projektberichte/2019-01.pdf.
- Udrea, T. et al. (2022). *Künstliche Intelligenz – Verstehbarkeit und Transparenz*. ITA-Bericht 2022-01. epub.oeaw.ac.at/ita/ita-projektberichte/ITA-2022-01.pdf.