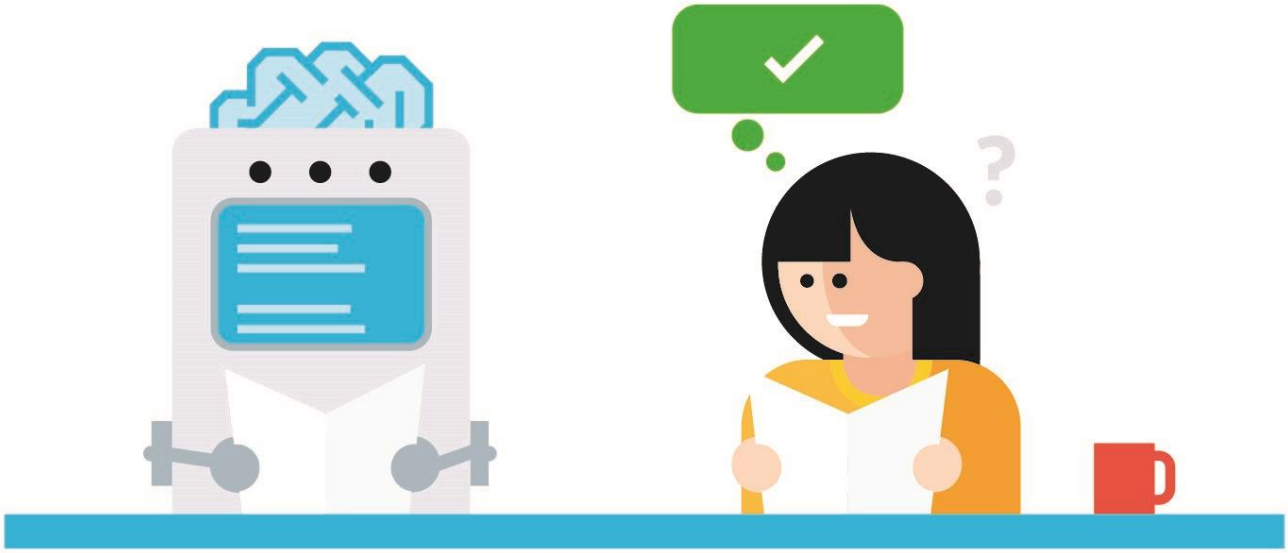


ChatGPT – wenn die künstliche Intelligenz schreibt wie ein Mensch. Und was es dabei zu beachten gilt.

Inhaltsverzeichnis

Der am 30. November 2022 veröffentlichte Chatbot ChatGPT verblüffte auf Anhieb Millionen von Nutzerinnen und Nutzer. Dieser Artikel beleuchtet die Schlüsselelemente, die zum Verständnis des Phänomens erforderlich sind, und wirft eine Reihe von Fragen auf, die in nächster Zukunft auf politischer, beruflicher und individueller Ebene relevant sein dürften.

Die Stiftung für Technologiefolgen-Abschätzung TA-SWISS hat die Aufgabe, die Auswirkungen neuer Technologien auf die Gesellschaft als Ganzes zu untersuchen. In diesem Sinn fördert sie den Dialog zwischen verschiedenen Fachbereichen und Standpunkten. Die Einführung von ChatGPT und anderen vergleichbaren Anwendungen berührt zahlreiche Tätigkeitsfelder und Themenbereiche. Für grosse Teile der Bevölkerung könnte sie zudem massive Auswirkungen haben. Aus diesem Grund prüft TA-SWISS, ob eine TA-Studie zu diesem Thema angesagt wäre, und verfolgt die Entwicklungen rund um ChatGPT weiterhin genau.

Worum geht es?

ChatGPT wurde vom kalifornischen Unternehmen OpenAI auf den Markt gebracht. OpenAI ist auf die Entwicklung künstlicher Intelligenz spezialisiert. Wie sein Name bereits verrät, ist ChatGPT ein *Chatbot* – ein textbasiertes Dialogsystem, das auf der Grundlage des GPT-3-Computermodells entwickelt wurde und Texte in menschlicher Sprache analysieren und generieren kann. Das Akronym GPT-3 steht für *Generative Pre-trained Transformer 3*. ChatGPT ist, mit anderen Worten, eine künstliche Intelligenz (KI), mit der man sich auf einer Website «unterhalten» kann. Die Anwendung generiert in Sekundenschnelle Antworten auf unterschiedlichste Fragen oder Eingaben: Witze, Textzusammenfassungen oder E-Mails, Aufsätze und Gedichte. ChatGPT liefert auch Programmcodes und weist auf Codierfehler hin. Und selbst wenn die Qualität der erhaltenen Ergebnisse schwankt, so muss doch festgestellt werden, dass die Antworten oft durchaus beeindruckend sind: Die Texte sind korrekt formuliert und strukturiert – und zwar in einem Mass, dass sie manchmal nur schwer von Texten zu unterscheiden sind, die von Menschen verfasst wurden.

Obwohl ChatGPT nicht mehr aus den Schlagzeilen kommt, ist diese Anwendung genau besehen nicht die erste KI-Anwendung, die ganze Texte produziert. Vielmehr ist sie das Ergebnis der kontinuierlichen Fortschritte in der Forschung zur künstlichen Intelligenz (mehr dazu weiter unten). Allerdings wird eine solche Dienstleistung mit ChatGPT zum ersten Mal einem breiten Publikum zur Verfügung gestellt. Seit der Veröffentlichung von ChatGPT überstürzt sich die Entwicklung: So ist Microsoft derzeit daran, verwandte Anwendungen mit seiner Suchmaschine Bing[1] sowie mit seinen Office-Lösungen Microsoft 365 (Word, Excel, PowerPoint, Teams, usw.) [2] zu verknüpfen, die nach dem gleichen Prinzip funktionieren wie ChatGPT. Auch Google[3] und das chinesische Unternehmen Baidu[4] haben angekündigt, in nicht allzu ferner Zukunft eine ähnliche Lösung in ihre Suchmaschinen integrieren zu wollen. Alibaba arbeitet ebenfalls an einem Projekt, das ChatGPT ähnelt.[5] Und seit neuestem stellt OpenAI eine Programmierschnittstelle (*API*) zur Verfügung, damit interessierte Unternehmen ChatGPT nahtlos in ihre Produkte integrieren können.[6] Darüber hinaus hat OpenAI vor Kurzem GPT-4 lanciert,[7] eine noch leistungsfähigere Version von GPT-3, sowie ein kostenpflichtiges ChatGPT-Abonnement. ChatGPT und Konsorten scheint also eine erfolgreiche Zukunft bevorzustehen. Darum drängt es sich auf, ihre Auswirkung auf die Gesellschaft genauer zu betrachten.

Themenverwandte Studien von TA-SWISS

- Studie «Wenn Algorithmen für uns entscheiden: Chancen und Risiken der KI» (2020)
- Studie «Bürger und Institutionen angesichts der Digitalisierung der Demokratie in der Schweiz» (2021)
- Bericht «Roboter im Spiegel der Technologiefolgen-Abschätzung» (2022)
- Laufende Studie «Deepfakes und manipulierte Realitäten» (geplante Veröffentlichung: 2024)
- Laufende Studie «Kultur und Digitalisierung» (geplante Veröffentlichung: 2024)

[1] Siehe Microsoft-Blog, «Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web» (Februar 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/).

[2] Siehe Microsoft-Blog, «Introducing Microsoft 365 Copilot» (März 2023, https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/).

[3] Siehe Google-Blog, «An important next step on our AI journey» (Februar 2023, https://blog.google/technology/ai/bard-google-ai-search-updates/).

[4] Swissinfo, «Le chinois Baidu va lancer son propre robot face à ChatGPT» (Februar 2023, https://www.swissinfo.ch/fre/toute-l-actu-en-bref/le-chinois-baidu-va-lancer-son-propre-robot-face-%C3%A0-chatgpt/48264576).

[5] Siehe CNBC, « Chinese tech giant Alibaba working on a ChatGPT rival; shares jump » (Februar 2023, https://www.cnbc.com/2023/02/08/chinese-tech-giant-alibaba-working-on-a-chatgpt-rival-shares-jump.html).

[6] Siehe OpenAI-Blog, «Introducing ChatGPT and Whisper APIs» (Februar 2023, https://openai.com/blog/introducing-chatgpt-and-whisper-apis).

[7] Siehe OpenAI-Blog, «GPT-4» (März 2023, https://openai.com/research/gpt-4).

Wie funktioniert ChatGPT?

Heute kann jede und jeder ein kostenloses Konto auf der ChatGPT-Website eröffnen – muss dabei allerdings auch eine Menge Angaben zur eigenen Person machen. Sodann lässt sich eine Frage oder eine Aufforderung (*Prompt*) eingeben und dabei auch angeben, wie die angeforderte Antwort in Bezug auf Länge, Struktur und Tonfall ausfallen soll. Um die Anfragen beantworten zu können, wurde ChatGPT vorgängig anhand von Milliarden im Internet verfügbarer Texte «trainiert». Die KI hat all diese Texte mathematisch analysiert, um sprachliche Regelmässigkeiten zu erfassen. ChatGPT arbeitet auf Grundlage der Wortfolgestatistik, berechnet also die Wahrscheinlichkeit, mit der innerhalb eines Textkorpus ein Wort auf das andere folgt. In seinen Antworten gibt das Programm somit wiederkehrende Wortmuster aus den Texten wieder, mit denen es gefüttert wurde. Diese Antworten enthalten daher immer eine grosse Zufallskomponente. Das erklärt auch, warum ChatGPT auf eine ihm mehrmals gestellte gleiche Frage jedes Mal eine andere Antwort ausspuckt.

OpenAI nutzt das sogenannt bestärkende Lernen aus menschlichem Feedback (*Reinforcement Learning from Human Feedback*). Das bedeutet, dass das Unternehmen Personal damit beauftragt, eine grosse Menge von Antworten des Systems zu beurteilen, plausible Ergebnisse zu bestätigen und unbefriedigende Ergebnisse als solche zu kennzeichnen, um den Algorithmus von ChatGPT zu verbessern. Das Unternehmen bemüht sich zudem um eine ethische Kalibrierung, insbesondere um gewaltverherrlichende, offen sexistische oder rassistische Äusserungen, aber auch Werturteile und Anleitungen zu illegalen oder gefährlichen Handlungen zu unterbinden (mehr zu den damit einhergehenden normativen Fragen weiter unten). Seit dem Start von ChatGPT werden die Nutzeranfragen laufend analysiert, um das Training weiter zu justieren.

Im weiteren Sinne gehört ChatGPT zur sogenannten **«generativen» künstlichen Intelligenz**, die seit einigen Jahren einen grossen Ausschwung erlebt. Dabei handelt es sich um Modelle, die auf der Grundlage einiger weniger Schlüsselwörter Bilder, Text oder Videos generieren. Auch diese «synthetischen» Inhalte sind das Ergebnis einer rein statistischen Analyse der Datenfülle auf Basis von *Deep Learning*.[8] Was die Textgenerierung betrifft, leiten sich all diese Innovationen aus der Entwicklung von «**Sprachmodellen**» ab. Diese IT-Modelle katalogisieren die Muster der Wortfolgen in den Texten, mit denen sie vortrainiert wurden, und reproduzieren sie, wenn sie neue Texte generieren.[9] Um die Kohärenz zwischen den Sätzen zu gewährleisten, werden die je nach Kontext unterschiedlichen Abhängigkeiten zwischen den Wörtern eines ganzen Textes erfasst (mithilfe der *Self-Attention*-Technik von Google).[10]

ChatGPT seinerseits basiert auf **GPT-3**. Dieses von OpenAI entwickelte Sprachmodell ist heute das grösste der Welt: Sein Training stützt sich auf 175 Milliarden Analyseparameter und 570 Gigabyte Text.[11] Auch Google rüstet auf und hat seine Sprachmodelle (BERT und LaMDA) inzwischen weiterentwickelt, das Gleiche gilt für Facebook mit XLM-R und seit kurzem LLaMa[12]. Im Übrigen existierten Textgeneratoren bereits vor ChatGPT. Ein Beispiel ist Jasper Chat, der ChatGPT ähnlich ist, und in Partnerschaft mit OpenAI entwickelt wurde.[13] Ein weiteres ist die GitHub-Copilot-Plattform, die auf Anfrage Codes generiert und an der OpenAI ebenfalls beteiligt war.[14]

---

[8] Siehe beispielsweise Dall-E (ebenfalls von Open AI) und Stable Diffusion für das Generieren von Bildern oder Synthesia für das Generieren von Videos.

[9] J. Goldstein, G. Sastry, M. Musser et al., «Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations», *arXiv* (2023, https://arxiv.org/abs/2301.04246).

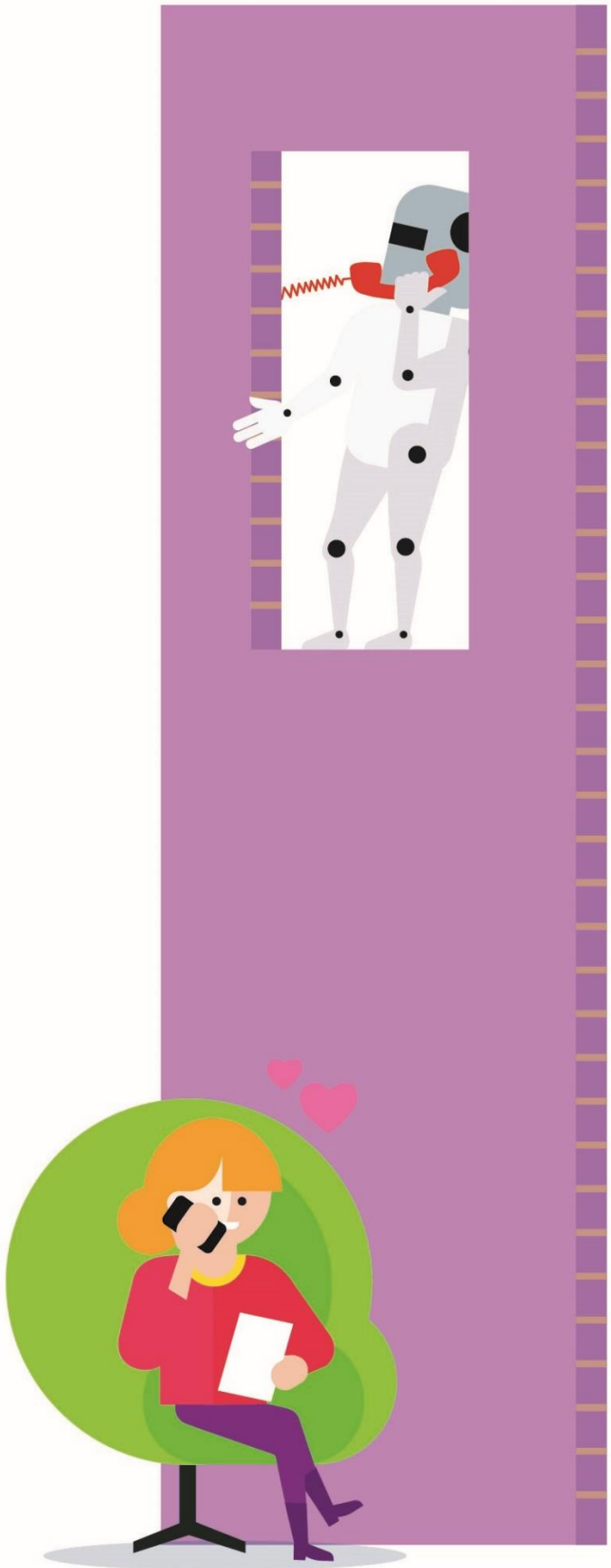
[10] A. Vaswani, N. Shazeer, N. Parmar et al., «Attention Is All You Need», *Advances in neural information processing systems* (2017, https://arxiv.org/pdf/1706.03762.pdf).

[11] «Generative Language Models and Automated Influence Operations» (*op. cit.*), S. 2.

[12] Siehe den MetaAI-Blog, «Introducing LLaMA: A foundational, 65-billion-parameter large language model» (Februar 2023, https://ai.facebook.com/blog/large-language-model-llama-meta-ai/).

[13] Siehe Jasper Chat und für weitere Beispiele C. Agar, «The Best AI Novel Writing Software For 2022», *TheTechReviewer* (2022, https://thetechreviewer.com/best-ai-novel-writing-software/).

[14] Siehe https://github.com/features/copilot.



Eine Fülle von Möglichkeiten, aber auch offene wirtschaftliche Fragen

Glaubt man den zahlreichen Medienberichten, so ist ChatGPT bereits heute ein äusserst beliebtes Tool. Sei es, um Antworten oder Ideen für ein Projekt zu erhalten oder um die Ausführung einer Aufgabe Schritt für Schritt zu planen: ChatGPT bietet Ressourcen für die unterschiedlichsten Tätigkeitsbereiche. So wird der Chatbot oft als ein Garant für mehr Effizienz und Produktivität angesehen. Nicht nur zum Verfassen von Texten, sondern auch zum Codieren und zur Informationssuche.[15] Gleichzeitig setzt die möglichst optimale Nutzung von Anwendungen wie ChatGPT neue Kompetenzen voraus: Dazu gehört das «Prompt Engineering»[16], d.h. die Fähigkeit, die relevantesten Fragen zu formulieren.

Es ist gut möglich, dass ChatGPT und seine Nachfolgersysteme all diejenigen Bereiche auf eine harte Probe stellen werden, in denen Texte – in natürlicher oder in Programmiersprache – produziert werden. Zu den am offensichtlichsten betroffenen Branchen gehören der Journalismus, die Kommunikation und das Marketing, aber auch die IT-Branche, das Recht oder die wissenschaftliche Forschung sowie bestimmte literarische Formate. Die Integration einer Anwendung wie ChatGPT in eine Suchmaschine könnte auch dazu führen, dass Websites weniger besucht werden und ihre Betreiber somit einen Teil ihrer Einkommensquellen einbüssen.[17] Dies bringt uns zu einem altbekannten Problem, das mit dem Fortschritt im Bereich der künstlichen Intelligenz einhergeht: Welche Berufe könnten durch diese Technologien verändert oder gar ersetzt werden? Welche neuen Tätigkeiten werden sie gleichzeitig neu schaffen? Und inwieweit können oder sollen wir uns an diese neuen Tools adaptieren?

In einem anderen Zusammenhang ist anzumerken, dass die Einrichtung und Nutzung solcher Sprachmodelle extrem **kostspielig** ist. Sie benötigen für ihr Training und ihre laufende Anpassung riesige Datenmengen, was wiederum eine enorme Rechenleistung ihrer Server voraussetzt. Natürlich könnten diese Modelle später öffentlich einsehbar werden (*open source*), so dass Dritte sie anpassen bzw. mit ihren eigenen Daten trainieren oder durch zusätzliche Funktionen ergänzen könnten (mittels der Methode des *Transfer Learning*). Bei GPT-2 ist dies bereits der Fall, bei GPT-3 noch nicht. Die Grundkosten und der Besitz der Daten verschaffen den Tech-Giganten jedoch einen massiven Vorteil und folglich grossen Einfluss.[18] Dabei stellt sich die Frage, wie sich ihr Konkurrenzkampf um diese Anwendungen auf die globale Wirtschaft auswirken wird. Und wie kann sich die Schweiz in diesem Feld positionieren?

Im Hinblick auf die Nachhaltigkeit erfordern grosse Sprachmodelle wie GPT **einen hohen Energieverbrauch** für die Berechnungen beim Training und später zum Beantworten der Suchanfragen der Nutzerinnen und Nutzer.[19] Die aktuelle Forschung versucht zwar, energieeffizientere Materialien und Rechnungsmethoden zu entwickeln. Wenn die für diese Anwendungen eingesetzten Server aber nicht mit nachhaltiger Energie betrieben werden, kann ihr CO_2-Fussabdruck enorm sein. Ausserdem wurden in Teilen der Produktionskette von KI-Systemen **prekäre Arbeitsbedingungen** aufgedeckt. So beruht ihre Betreuung auf repetitiven Abläufen, weil es darum geht, in möglichst kurzer Zeit möglichst viele Ergebnisse zu kontrollieren – auch solche mit Gewaltdarstellungen oder anderen problematischen Inhalten. Diese Aufgabe wird häufig an Subunternehmen ausgelagert, die ihr Personal in stark von Armut

[15] Siehe beispielsweise M. Chui, R. Roberts und L. Yee, «Generative AI is here: How tools like ChatGPT could change your business», *McKinsey* (2022, https://www.mckinsey.com/capabilities/quantumblack/our-insights/generative-ai-is-here-how-tools-like-chatgpt-could-change-your-business).

[16] Siehe zum Beispiel D. Holzer, «KI schafft neue Jobs: Was macht eigentlich ein Prompt Engineer?», *Br24* (2023, https://www.br.de/nachrichten/netzwelt/ki-schafft-neue-jobs-was-macht-eigentlich-ein-prompt-engineer,TX4P23Z).

[17] Siehe A. Seydtaghia, «Le match des moteurs de recherche est relancé grâce à l'intelligence artificielle», *Le Temps* (2023, https://www.letemps.ch/economie/match-moteurs-recherche-relance-grace-lintelligence-artificielle).

[18] Siehe R. Fulterer «KI-Forschung nach dem Vorbild Cern: Die ehemalige Spass-Firma Hugging Face fordert Meta und Google heraus», NZZ (2022, https://www.nzz.ch/technologie/huggingface-diese-nach-einem-emoji-benannte-firma-fordert-meta-und-google-heraus-ld.1682435).

[19] Siehe D. Patterson, J. Gonzalez, Q. Le et al., «Carbon Emissions and Large Neural Network Training», *arXiv* (2022, https://arxiv.org/ftp/arxiv/papers/2104/2104.10350.pdf) oder L. Anthony, B. Kanding und R. Selvan, «Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models», *arXiv* (2020, https://arxiv.org/abs/2007.03051).

betroffenen Ländern massiv unter Druck setzen.[20] Laut einer kürzlich durchgeführten Recherche des *Time* in Kenia scheint dies auch bei ChatGPT der Fall zu sein.[21]

Andererseits kann die kommerzielle Nutzung von ChatGPT **die Urheberrechte** an den von ihm analysierten Texte verletzen, insbesondere wenn dabei geschützte Formulierungen aufgenommen werden, ohne dass dies offengelegt wird. Wie lassen sich die Daten, die in eine Antwort eingeflossen sind, identifizieren und damit feststellen, ob tatsächlich eine Rechtsverletzung vorliegt? Sollte, unabhängig von solchen Verstössen, die Nutzung der Originaldaten durch ChatGPT vergütet werden? Wie wäre dabei vorzugehen (d.h. welcher Anteil würde den Urheberinnen und Urhebern der verwendeten Daten zustehen und welcher Anteil OpenAI)?[22]

Von der Schule zur Universität: ein Wendpunkt in der Bildung

Nach der Einführung von ChatGPT gerieten auch die Bildungseinrichtungen in Aufruhr. ChatGPT eignet sich hervorragend dazu, ganze Passagen schriftlicher Arbeiten zu verfassen oder Ideen für eine Hausarbeit zu entwickeln, und scheint selbst in der Lage zu sein, Prüfungsfragen zu beantworten. Neue Möglichkeiten für Schummeleien und Plagiate sind nun also mit wenigen Klicks verfügbar. Zwar überzeugen die Ergebnisse von ChatGPT laut verschiedener Medienberichte nicht immer überwältigend, doch sie reichen oft aus, um einen Test zu bestehen und dabei vielleicht sogar eine gute Note zu erhalten.[23] Der Einsatz des Chatbots könnte den bisher auf Übung und Erfahrung basierten Erwerb von Kenntnissen und Kompetenzen untergraben, weil die dabei erforderliche Anstrengung an die Maschine delegiert würde. Lernprozesse, die ein bewusstes Verstehen voraussetzen, würden wegfallen. Zudem könnte uns die Schnelligkeit der Antworten überfordern, da wir für die Verarbeitung von Informationen Zeit brauchen.[24] Viele Lehrpersonen sehen aber auch Chancen für ihren Unterricht. Zum Beispiel als Gelegenheit, sich mit den Antworten von ChatGPT kritisch auseinanderzusetzen.[25]

ChatGPT stellt somit nicht nur die derzeitigen Bewertungssysteme in Frage, sondern stösst auch eine breite Diskussion über **Lernmethoden und die Produktion von Wissen** an. In welchem Umfang und auf welche Weise soll diese Art von künstlicher Intelligenz in den Unterricht integriert werden? Wie lässt sich ihr Einsatz möglichst sinnvoll begleiten, da ein vollständiges Verbot kaum umsetzbar scheint? Welche Kompetenzen sollten im Zusammenhang mit künstlicher Intelligenz in der Ausbildung erworben werden?[26] Wo sind Grenzen zu setzen, um dem Menschen eigene Kompetenzen zu bewahren? Und wie lässt sich der Einsatz von künstlicher Intelligenz in der wissenschaftlichen Forschung legitimieren, beispielsweise als Hilfsmittel für die Recherche, Synthese oder das Verfassen von Texten?[27]

Zu den in Erwägung gezogenen Massnahmen gehört die Entwicklung von Anwendungen, die **mithilfe von künstlicher Intelligenz erzeugte Texte erkennen können**.[28] Ein Forscher von Open-AI hat

---

[20] Siehe Antonio Casalli, *En attendant les robots. Enquête sur le travail du clic* (2019, Le Seuil).

[21] B. Perrigo, «Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic», *Time* (2023, https://time.com/6247678/openai-chatgpt-kenya-workers/).

[22] I. Wildhaber, «ChatGPT aus rechtlicher Perspektive: Was Unternehmen beachten sollten», Universität St. Gallen (2023, https://www.unisg.ch/de/videodetail/news/chatgpt-aus-rechtlicher-perspektive-was-unternehmen-beachten-sollten/).

[23] Siehe P. Kaelin, «Luzerner Hochschulen wappnen sich gegen ChatGPT», *zentralplus* (2023, https://www.zentralplus.ch/beruf-bildung/luzerner-hochschulen-wappnen-sich-gegen-chatgpt-2512228/).

[24] Voir M. de Vevey, « ChatGPT : à quoi ressemblera l'enseignement dans dix ans ?», *uniscope* (2023, https://wp.unil.ch/uniscope/chatgpt-a-quoi-ressemblera-lenseignement-dans-dix-ans/).

[25] Service écoles-médias du canton de Genève, «Risques et opportunités de ChatGPT pour l'enseignement : premiers éléments d'analyse», (2023, https://edu.ge.ch/sem/ressources/risques-et-opportunites-de-chatgpt-pour-lenseignement-premiers-elements-danalyse-3574).

[26] Dazu siehe die Studie von TA-SWISS über künstliche Intelligenz, S. 171-8 und 301-2.

[27] Siehe C. Stokel-Walker und R. Van Noorden, «What ChatGPT and generative AI mean for science», *Nature* (2023, https://www.nature.com/articles/d41586-023-00340-6).

[28] Siehe *DetectGPT* (E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning und C. Finn, «DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature», Cornell University (2023, https://arxiv.org/pdf/2301.11305.pdf) oder *GPTZero* (C. Cassidy, «College student claims app can detect essays

beispielsweise die Idee ins Spiel gebracht, in den Metadaten von ChatGPT-Texten eine unsichtbare Markierung zu verankern: Eine Art Wasserzeichen, das bescheinigt, dass der Text von ChatGPT stammt.[29] Verschiedene Fragen bleiben jedoch offen: Wer hätte Zugriff auf solche Tools, die in der Lage wären, über jeden Text eine Menge Informationen preiszugeben? Wie zuverlässig wären sie? Und was ist mit den konkurrierenden Anwendungen zur Umgehung dieser Tools, die zwangsläufig auf den Markt kommen würden?

Im Hinblick auf den Zugang zur Bildung sollte zudem erwähnt werden, dass Anwendungen wie ChatGPT bestimmte Lernbarrieren abbauen könnten: Zum Beispiel indem sie Inhalte in leichte Sprache umformulieren.[30] Auch bei Schwierigkeiten mit dem schriftlichen Ausdruck könnten sie helfen, sei es generell oder im Kontext einer Fremdsprache. Bleibt die Frage nach dem Preis solcher Dienstleistungen, der ihr Potenzial wiederum beschränken könnte.

Grenzen von ChatGPT – und Risiken, die es im Auge zu behalten gilt

Trotz des unbestrittenen technischen Fortschritts, den sie verkörpert, ist die Anwendung von ChatGPT mit erheblichen Risiken verbunden. Zunächst einmal operieren statistische Sprachmodelle mit Wahrscheinlichkeiten, ohne Argumentation oder Interpretation. Deshalb ist es wichtig zu unterstreichen, dass die Begriffe richtig oder falsch in diesen Systemen nicht existieren und dass sie die Inhalte, die sie verarbeiten, nicht «verstehen» können. ChatGPT macht deshalb viele **sachliche Fehler**. Auch wenn es um Quellenangaben geht, erfindet ChatGPT nicht selten Links oder ganze Artikel.[31] Nur wer sich auf einem Gebiet bereits gut auskennt, kann die Qualität der Antwort auf eine Frage beurteilen. Oft sind diese Antworten jedoch sehr souverän und auf formeller Ebene sehr überzeugend formuliert, was bei den Nutzenden den Eindruck erwecken kann, dass sie es hier mit einem hohen Mass an Kompetenz zu tun haben oder sogar mit einem denkenden menschlichen Gegenüber. Die Gefahr, falschen Angaben zu vertrauen, ist also greifbar und kann z. B. bei medizinischen, politischen oder rechtlichen Fragen folgeschwer sein. Das Ersetzen eines menschlichen Urteils durch die statistische Berechnung von Wörtern wirft auch ein grundlegendes ethisches Problem auf.[32]

Darüber hinaus lassen sich textgenerierende Anwendungen missbräuchlich in eine bestimmte Richtung lenken. Sie können beispielsweise eingesetzt werden, um Massenkampagnen zur Meinungsbeeinflussung oder **Desinformation** zu befeuern, indem sie in schneller Folge auf den jeweiligen Kontext zugeschnittene Sätze generieren, etwa in sozialen Medien und auf Internetforen.[33] Darüber hinaus könnte ChatGPT Codezeilen zu einem **Cyberangriff** beisteuern.[34]

written by chatbot ChatGPT», *Guardian* (2023, https://www.theguardian.com/technology/2023/jan/12/college-student-claims-app-can-detect-essays-written-by-chatbot-chatgpt)).

[29] Siehe S. Aaronson, «My AI Safety Lecture for UT Effective Altruism» (2022, Vorlesungsskript, https://scottaaronson.blog/?p=6823).

[30] Siehe B. Blume, *Deutsches Schulportal der Robert Bosch Stiftung*, «ChatGPT: Das Ende vom Lernen wie wir es kennen» (2023, https://deutsches-schulportal.de/kolumnen/chatgpt-das-ende-vom-lernen-wie-wir-es-kennen/).

[31] Es ist anzumerken, dass eine in eine Suchmaschine integrierte Anwendung Quellen anzeigen könnte. Siehe A. Glaese, N. McAleese, M. Trebacz et al., «Improving alignment of dialogue agents via targeted human judgements», *DeepMind* (2022, https://arxiv.org/pdf/2209.14375.pdf), welche die Google-Anwendung Sparrow behandelt, oder den Prototyp WebGPT von OpenAI (https://openai.com/blog/webgpt/).

[32] Zum Bedarf an menschlicher Kontrolle über automatisierte Entscheidungen über Menschen, siehe die Studie von TA-SWISS über künstliche Intelligenz, S. 136-7 und 293.

[33] J. Goldstein, G. Sastry, M. Musser et al., «Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations», *arXiv* (2023, https://arxiv.org/abs/2301.04246).

[34] Siehe das Experiment von Cybersicherheitsfachpersonen des Unternehmens Check Point Software (S. Ben-Moshe, G. Gekker, G. Cohen, «OpenAI: AI That Can Save the Day or HACK it Away», (2022, https://research.checkpoint.com/2022/opwnai-ai-that-can-save-the-day-or-hack-it-away/).

Wie oft im Bereich der künstlichen Intelligenz geben auch die **Trainingsdaten** Anlass zur Sorge. Diese Daten stammen aus von Menschen verfassten Texten und widerspiegeln deshalb die Meinungen und Haltungen ihrer Verfasser. Daher enthalten sie auch Vorurteile und Stereotypen, die sich dann in den Antworten des Programms wiederfinden und die in diesen Daten reflektierten Diskriminierungen weiter verstärken. Am häufigsten sind davon in erster Linie Frauen und Minderheiten betroffen.[35] Trotz der Vorsichtsmassnahmen von OpenAI wurden in den sozialen Medien zahlreiche diskriminierende Formulierungen von ChatGPT verzeichnet.[36] Technisch gesehen ist es derzeit schwierig, die Verzerrung der Daten auf denen das KI-System aufbaut, zu vermeiden oder zu korrigieren. Zudem ist die Identifizierung solcher Verzerrungen ein nicht transparent erfolgender sozialer bzw. politischer Prozess, der auf normativen Urteilen beruht und daher nicht auf technische Parameter reduziert werden kann.

Auch eine angemessene **Repräsentation** aller Sprachen und Kulturen scheint in diesem Bemühen um Fairness kaum erreichbar, da die grossen Sprachmodelle der Gegenwart hauptsächlich aus dem Englischen entwickelt wurden. Auch ihr Training auf der Grundlage von Online-Inhalten führt zu einer Selektion, da einige Standpunkte im Internet kaum vertreten sind, während andere verstärkt werden, z.B. in den sozialen Medien.[37] Der Einsatz dieser Programme könnte sodann zu einer **Standardisierung** von Meinungen, Informationen und Wissen führen, da sie rein statistisch operieren. Während seltenere Inhalte eher aussortiert werden, werden häufiger vertretene Positionen bevorzugt. Dieser Effekt dürfte schwer zu erkennen und zu überwachen sein, da die Nutzerinnen und Nutzer keinen Einblick in den Entstehungsprozess einer Antwort haben. Es ist zu erwarten, dass zukünftige Anwendungen immer mehr künstliche Texte integrieren werden und somit immer öfter mit maschinengemachten Texten gefüttert werden - was Standardisierungstendenzen wiederum verstärken würde.[38]

[35] Siehe E. M. Bender, T. Gebru, A. McMillan-Major, et S. Shmitchell (2021): «On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?». *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, *Association for Computing Machinery* (https://dl.acm.org/doi/abs/10.1145/3442188.3445922).

[36] Siehe beispielsweise I. Vock, «ChatGPT proves that AI still has a racism problem», *New Statesman* (2023, https://www.newstatesman.com/quickfire/2022/12/chatgpt-shows-ai-racism-problem).

[37] «On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?», op cit.

[38] Siehe K. Creel und D. Hellman, «The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision Making Systems», *Virginia Public Law and Legal Theory* (2021, https://ssrn.com/abstract=3786377).

Angesichts dieser Risiken stellt sich die Frage, welche Anwendungen von ChatGPT wir fördern wollen und in welcher Form. Wie fügen sich diese neuen Anwendungen in bestehende KI-**Normen** ein? Sind neue Standards notwendig, um ihren Einsatz und ihre Weiterentwicklung zu begleiten? Welche Optionen sind in Anbetracht der Allgegenwart des Internets und seines internationalen Charakters praxistauglich? Welche ethischen Anforderungen sollen an diese Modelle gestellt werden, und wer entscheidet, welche ihrer Antworten akzeptabel sind? Wie lässt sich hinsichtlich dieser Kriterien und Entscheidungen Transparenz gewährleisten?

Und ausserdem: Wer ist für die Probleme verantwortlich, die sich aus der Nutzung dieser Tools ergeben? Wie lässt sich das Bewusstsein der Öffentlichkeit für die Grenzen und die Fehlbarkeit von Chatbots schärfen? Was geschieht mit den persönlichen Daten der Nutzerinnen und Nutzer, und welche Datenschutzmassnahmen sollten getroffen werden? Was passiert, wenn jemand sensible Informationen eingibt? Diese zahlreichen Herausforderungen zeigen, wie wichtig eine breite öffentliche Debatte ist.

Fragen zur Zukunft unseres Schreibens und unseres Denkens

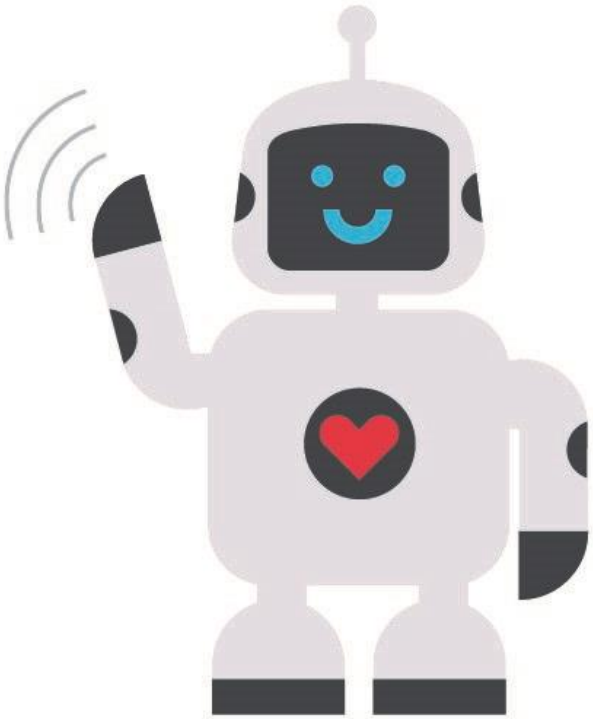
Die ethischen Fragen treten klar zu Tage. Das Aufkommen künstlicher Intelligenzen, die Texte verfassen können, die denen eines Menschen gleichen, stellt auch unsere Schreib- und Lesegewohnheiten in Frage – und damit unser Verhältnis zur Welt. Wann genügt uns die Berechnung einer wahrscheinlichen Wortfolge in Texten, die ein Computermodell auswählt, und wann nicht und warum? Haben KI-Texte im Vergleich zu menschengemachten ihre eigenen Merkmale, und was verraten sie? Wie wird der Einsatz dieser Hilfsmittel die Textproduktion und unsere Gedankenarbeit während des Schreibprozesses beeinflussen?

Bereits heute stammen zahlreiche Texte im Internet von KI-Systemen, ohne dass wir uns dessen unbedingt bewusst werden. Wenn sich Anwendungen wie ChatGPT weiterverbreiten, wird die Zahl maschinengemachter Texte noch weiter ansteigen. Sollten künstlich erzeugte Inhalte als solche deklariert werden, so dass wir sie auch erkennen können? Darüber hinaus stellt sich die Frage, wie sich eine Zunahme von KI-generierten Texten auf unsere Sprachen auswirken und, da Sprache immer auch unser Denken prägt, auf unsere Sicht der Welt? Wie lässt sich der Einfluss einer statistischen und auf Wahrscheinlichkeiten ausgerichteten «Denkweise» erkennen? Müssen wir vielleicht Modelle entwickeln, die sich ausschliesslich auf von Menschen verfasste Texte stützen?

Auch angesichts dieser Fragen sind umfassende Überlegungen erforderlich, um die richtige Dosierung der Interaktion zwischen menschlichen und künstlichen Texten zu bestimmen. Wie die vorangehenden Ausführungen nahelegen, ist ein interdisziplinärer und sektorübergreifender Ansatz unerlässlich, um die Auswirkungen dieser Technologien zu erfassen und den Handlungsbedarf abzuschätzen. Die Stiftung TA-SWISS wird daher die Entwicklungen in diesem Bereich weiterhin aufmerksam verfolgen. Ihre Kommentare sind uns sehr willkommen!

Wir danken Dr. Bruno Baeriswyl, Prof. Antoine Bosselut, Prof. Olivier Glassey, Prof. Lorenz Hilty, Dr Anna Jobin und Prof. Reinhard Riedl für ihre wertvollen Kommentare bei der Erarbeitung dieses Textes.

Laetitia Ramelet,
Projektleiterin TA-SWISS,
März 2023



TA-SWISS
Stiftung für Technologiefolgen-Abschätzung
Brunngasse 36
3011 Bern

www.ta-swiss.ch

mitglied der
akademien der wissenschaften schweiz