ChatGPT auf dem Weg zur Weltherrschaft? Risiken Künstlicher Intelligenz

Barbara Budrich und *Jannik L. Esser*

Einleitung

Künstliche Intelligenz (KI) begleitet uns schon seit vielen Jahren, doch seit Anfang 2023 schlagen die Wellen der öffentlichen Aufmerksamkeit hoch: Zwischen Weltuntergangsszenarien und Fortschrittsbegeisterung bekommt Rechenkapazität menschliche Züge zugeschrieben. Bei der Beurteilung dieser menschlichen Züge überwiegen die negativen Eigenschaften von Menschen, vor allem Machthunger und der Wunsch danach, die eigenen Interessen gegen die aller anderen durchzusetzen. Im Folgenden schauen wir uns einige der Szenarien an, in denen KI derzeit unter Generalverdacht steht, das Ende der Menschheit einläuten zu wollen. Als Beispiel möge uns dafür ChatGPT dienen. Zunächst aber werfen wir einen kurzen technischen Blick darauf, womit wir es eigentlich zu tun haben.

Was ist KI und wie funktioniert sie?

ChatGPT ist, wie alle „sprechenden" KI, ein maschinell lernendes System für die Analyse und Vorhersage von Wörtern. Derartige KI sind auf der Basis einer bestimmten Datenmenge trainiert und können basierend auf diesen Daten z.B. das nächste Wort eines Satzes vorhersagen. Programme dieser Art sind aus dem Alltag seit Jahren nicht mehr wegzudenken, helfen sie doch z.B. bei der Autovervollständigung von Wörtern und Sätzen beim Gebrauch von Smartphones und Suchmaschinen. Bei ChatGPT gehen die Fähigkeiten über diese Vervollständigung von Wörtern und Sätzen hinaus. Seine Programmierer:innen haben diese KI so ausgestattet, dass sie selbst in unterschiedlichen Sprachen zusammenhängende Texte generieren kann.

ChatGPT ist dennoch an die Grenzen seiner Datenbasis gebunden und berechnet auf dieser Grundlage Wahrscheinlichkeiten, die seine Antworten leiten. Im Jahre 2023 bezieht sich die weitverbreitete kostenlose Version von ChatGPT auf Daten bis ins Jahr 2021. Das Programm nutzt beim Berechnen seiner Ergebnisse ausgereiftere Methodiken, wie z.B. neuronale Netze, aber auch die sind nichts weiter als mehrdimensionale Vektoren (sog. Tensoren), die miteinander multipliziert und kombiniert werden. Im Hintergrund ist

Barbara Budrich, M.A.,
Verlegerin, Verlag Barbara Budrich, Opladen, Berlin, Toronto.

Jannik L. Esser, B.Sc.,
Masterstudent im Fach Informatik an der Heinrich-Heine-Universität Düsseldorf, Backend Developer bei PwC Standort Düsseldorf.

Gesellschaft • Wirtschaft • Politik (GWP) 72. Jahrg., Heft 4/2023, S. 415-419
https://doi.org/10.3224/gwp.v72i4.03

jede KI nichts anderes als ein automatisiertes Programm, das mithilfe von Vektorrechnungen und stochastischen Auswertungen Wahrscheinlichkeiten kombiniert. Wir kommen im Verlauf unserer Überlegungen noch darauf, welche Herausforderungen diese inhärente Dominanz von Quantität haben kann.

ChatGPT und seine Kollegen wie Bard von Google oder Jasper usw. sind allerdings auf einer sehr großen Datenmenge trainiert worden, weshalb sie auch LLM, Large Language Model, genannt werden. Sie verfügen also über viele Daten, die sie kombinieren können, um natürliche Sprache zu verstehen und zu generieren. Somit können LLMs Usern zu vielen Themen Informationen liefern.

Schüler:innen und Studierende haben sehr schnell die Vorzüge dieser Anwendungen verstanden und nutzen sie für bequeme Abkürzungen: Sie lassen sich Zusammenfassungen von Texten ausgeben, die sie eigentlich in Gänze hätten selbst lesen sollen, oder bei schwierigen Prüfungen helfen – wie 2023 bei den Abiturprüfungen in Hamburg gesehen.

Vom Ende der Menschheit durch KI

Im März 2023 äußerten sich zahlreiche Persönlichkeiten aus dem Tech-Bereich, darunter auch Elon Musk (Space X) und der Apple-Mitbegründer Steve Wozniak, in einem Offenen Brief des Future of Life Institute (2023) kritisch und forderten ein Pausieren in der rasanten, unregulierten KI-Entwicklung. Die große Sorge: Die ungebremste Entwicklung in einem scheinbar rechtsfreien Raum führe in die Katastrophe. Die Katastrophe erscheint zwar in unterschiedlichen apokalyptischen Szenarien, doch der Kern ist derselbe: Kontrollverlust der Menschen, Kontrollübernahme durch die KI. Und damit: das Ende der Menschheit.

Szenarien gottgleicher Megarechner, die ihre eigenen Machtfantasien ausleben, sind nicht ganz neu. Doch gehen die größten Gefahren von der KI selbst aus? Oder vielmehr von ihrem missbräuchlichen Einsatz durch Menschen? Daraus würde sich nicht die Machtübernahme durch die KI ergeben. Sondern Machtkonzentration und absolute Kontrolle könnten mit der Unterstützung entsprechend ausgerichteter KI mit weniger Aufwand durchgesetzt werden.

Allerdings ergeben sich aus den Möglichkeiten der LLMs derzeit auch kleinere Probleme als der Untergang der Menschheit oder der Auf- und Ausbau von Terrorregimes. Fangen wir klein an, bei der Betrachtung von Urheberrechtsfragen und dem Aushebeln von Prüfungen.

Schummeleien und Urheberrecht

Wenn Schüler:innen ihre eigene Cleverness einsetzen, um sich Arbeit zu ersparen, nutzen sie LLMs im Sinne der Erfinder:innen. Den Schüler:innen ist es einerseits nicht so wichtig, wer einen Text ursprünglich verfasst hat, den sie zum Beispiel als Hausarbeit im eigenen Namen einreichen. Und tatsächlich verhalten sie sich damit den aktuellen urheberrechtlichen Regelungen entsprechend korrekt, sofern sie den KI-generierten Text überarbeitet haben. Eine KI kann nach geltendem Urheberrecht selbst keinen Status als Urheberin zuerkannt bekommen. Urheber:innen bleiben also die ursprünglichen Autor:innen. Es sei denn, ein Mensch überarbeitet den KI-generierten Text. Dann wird Letzterer zum Urheber.

Für die Wissenschaft hat dies derzeit wenig amüsante Konsequenzen. Hier ein Beispiel aus der Erfahrungswelt der Verlegerin: Eine generative KI wurde von Dritten aufgefordert, einen spezifischen Text zu übersetzen und zusammenzufassen. Diese „Auftragsvergabe“, Prompt genannt, war der einzige menschliche Anteil an dem auf diese Art neu entstehenden Text. Wir erinnern uns: Die KI kann keine Urheberrechte erwerben. Anschließend wurde der überarbeitete Text in einer anderen Sprache folgerichtig unter dem Namen des ursprünglichen Urhebers veröffentlicht. Ein Text, den dieser Autor nie gesehen hatte. Der Autor stieß auf diese Publikation – und war sehr verärgert. Abhängig von der Qualität der Texte, die auf diese Art produziert werden, drohen Reputationsschäden oder gar größere Probleme, wenn auf diesem Wege gar Falschinformationen entstehen.

Erschwerend kommt hinzu, dass ChatGPT ein äußerst bemühter Schreibdiener ist: Nicht immer erkennt er die Quelle, aus der er zitiert. Da

sich das Programm aber bemüht, jede Frage innerhalb seines Rahmens zu beantworten, liefert es brav eine Quelle. Ob es diese Quelle jedoch wirklich gibt, bleibt zu prüfen. Warum „halluziniert" die KI an dieser Stelle? Wir erinnern uns: Es werden Wahrscheinlichkeiten errechnet. Und auch bei derartigen Quellenangaben errechnet ChatGPT, welcher Name, welcher Aufsatztitel, welche Zeitschrift am wahrscheinlichsten zu benennen sind.

Die Entwickler:innen von ChatGPT haben die urheberrechtlichen Aspekte ignoriert, die das Training des LLMs hätten behindern können. Es sind Texte in die gigantische Datensammlung eingeflossen, die ohne Genehmigung seitens der Urheber:innen niemals hätten genutzt werden dürfen. Nun könnte man behaupten, dass Urheberrechte in Zeiten von KI keine Rolle mehr spielen dürften. Das würde allerdings auch dazu führen, dass alle Codes und Daten von ChatGPT und Co. offengelegt werden müssten: Denn auch die Leistung der Programmierer fällt in den Bereich des Urheberrechts. Allerdings hat OpenAI, das Unternehmen, dem ChatGPT gehört, wenig Interesse an derart durchgängiger Transparenz. Hier wird klar, dass es die (wirtschaftlichen) Interessen von Menschen sind, die zu Rechtsbrüchen führen, nicht etwa „eigene Interessen" einer KI.

Wenn also beklagt wird, dass es keine Regulierung für KI gebe, ist das nicht ganz korrekt. Natürlich tut Regulierung Not. Allerdings: Wenn sich Unternehmen nicht an geltendes Recht halten, hilft auch Regulierung nicht.

Freilich erklingt der Ruf nach Regulierung nicht allein auf der Basis von Rechtsbrüchen beim Urheberrecht. Und auch in diesem Bereich sind längst nicht alle Dinge geklärt. Aktuell (Stand September 2023) wird die Europäische KI-Verordnung (AI Act) diskutiert und unter vielen anderen fordern auch Urheberrechts- und Kreativverbände eine stärkere Beachtung und den Schutz ihrer Interessen (Initiative Urheberrecht 2023).

Kritisches Denken und Desinformationskampagnen

Wenn Yuval Noah Harari (2023) warnt, KI habe „das Betriebssystem der menschlichen Zivilisation gehackt", dann geht es ihm um Sprache, um Geschichten. So, wie KI Sprache beherrscht, können Menschen nicht mehr zwischen Mensch und Maschine unterscheiden. KI könne Menschen mit Sprache manipulieren, weil über Sprache Emotionen hervorgerufen und gelenkt werden können.

Erste Ansätze dieser Manipulation haben wir im Netz bereits beobachten können. Social-Media-Algorithmen haben dafür gesorgt, dass die eigene Meinung immer wieder bestärkt wurde. Das Gesetz der Masse führt dazu, dass User in der eigenen Timeline „mehr vom Selben" serviert bekommen. So entstehen Echokammern, in denen jede:r wieder und wieder die eigene Stimme vernimmt – reflektiert durch andere, die die jeweilige Meinung teilen. Daraus erwachsen Illusionen von Einigkeit, von „wir" und „die". Gefährliche Ansätze von Polarisierung, die das Zeug haben, Gesellschaften zu spalten. Anschaulich wurde dies während der Corona-Pandemie, als sehr ausgeprägte „Impfgegner"- und „Impfbefürworter"-Spaltungen entstanden. Ein Beispiel, das auch Harari in seinem Vortrag 2023 bemüht.

Schon ohne eine bösere Absicht damit zu verbinden, als Kund:innen länger auf der eigenen Plattform zu halten, können auf diesem Wege Verzerrungen entstehen, die aus Meinungsvielfalt „Richtig" und „Falsch", „Freund" und „Feind" machen. Doch wenn es nur noch eine Informationsquelle gäbe, eine einzige KI, der wir nahezu uneingeschränkt „glauben", ohne selbst kritisch zu überprüfen und nachzudenken, birgt dies eine noch größere Sprengkraft.

Schlimm genug, wenn ChatGPT wie andere KI-Software bei seinen Antworten dem Gesetz der Wahrscheinlichkeit folgt. Diese Wahrscheinlichkeiten können allein aufgrund von Quantität beispielsweise vorherrschende Vorurteile reproduzieren. So machte der Bewerbungsroboter von Amazon 2018 Schlagzeilen, weil er spezifische Merkmale im Rekrutierungsprozess vorzog – womit er bestehende Diskriminierung zum Beispiel von Frauen verstärkte (Wilke, 2018). Solange Programmierer:innen diesen Vorurteilen auf die Schliche kommen, können sie sie ändern. Jedoch leben Menschen immer in spezifischen soziohistorischen Kontexten und werden damit blind für die je eigenen Blindheiten.

Gleichwohl gibt es Bestrebungen und Techniken, faire Entscheidungsalgorithmen zu trainie-

ren. Das Problem der eigenen soziohistorischen und kulturellen Blindheit bleibt bestehen. Doch die reine Orientierung an der großen Zahl kann relativiert werden. Bei ChatGPT ist dies bislang allerdings nicht implementiert worden (Ferrara, 2023).

Noch schlimmer wird es, wenn gezielte Desinformationskampagnen eingesetzt werden. So sollen russische Hacker 2016 versucht haben, Einfluss auf die US-Präsidentschaftswahlen zu nehmen. Stehen derartigen gezielten, systematischen Kampagnen stärkere Tools zur Verfügung – wie ausgereifte KI – um massenhaft Fehlinformationen in wenigen Minuten zu verfassen, wird großflächige Desinformation umso einfacher.

Wir sprechen wiederum nicht davon, dass die KI autonom eigene Ziele verfolgt. Die größte Gefahr in den hier thematisierten Kontexten liegt im Missbrauch dieses hochentwickelten Werkzeugs durch Menschen, die ihre Interessen durchsetzen wollen.

Vom lebendigen demokratischen Diskurs bis hin zur absoluten Informationskontrolle

Die Anfangszeiten des Internet waren mit großer Euphorie verknüpft: Der große Vorteil war – und ist –, dass es keine zentrale und beherrschende Stelle gibt. Das Internet war quasi gleichzusetzen mit einer multilokalen Weltgesellschaft, bei der alle Menschen gleichen Zugang zu Informationen und gleiche Ausdrucksmöglichkeiten haben. Eine Weltdemokratie mit globaler Meinungsfreiheit und weltweitem Diskurs. Lange schon haben wir uns begleitet von „Shitstorms", Cyberangriffen, „Hate Speech" und „Deep Fake" von der naiven Vorstellung des digitalen Paradieses verabschiedet. Der notwendige lebendige demokratische Diskurs, die Auseinandersetzung um unterschiedliche Interessen, Perspektiven und das Ringen um gute Wege für möglichst viele findet nur in wenigen – zumeist regulierten, also geschützten – Bereichen statt.

Mit zu „Aufpassern" trainierten KIs ist es freilich möglich, freien Gedankenaustausch im Internet zu manipulieren. Gepaart mit Deep Fake, bei dem z.B. Influencer oder Politiker:innen in Videos mit zielgerichteter Wahrheitsverdrehung manipuliert werden, können jedem Menschen nach Belieben Aussagen zugeschrieben werden. Video- oder Audio-Mitschnitten ist mit wachsendem Misstrauen zu begegnen, weil Fälschungen mit Hilfe von KI immer leichter und besser werden. Gezielte Angriffe auf echte Webseiten, die mit anderen Informationen bestückt nachgebaut werden – all diese Ansätze können genutzt werden, um Informationen besser zu kontrollieren und Menschen in die Irre zu führen.

Doch auch in diesem Szenario ist es nicht die KI selbst, die „böse" geworden ist: Wieder sind es Menschen, die eigene Interessen durchsetzen. Und dafür mittels KI weniger Aufwand betreiben müssen. Ein machtvolles Tool, missbräuchlich eingesetzt.

Fazit

Die Gefahren, die im Kontext von KI beschworen werden, sind durchaus real. Allerdings erscheint in unseren hier gewählten Beispielen die Quelle der Gefahr nicht die KI selbst zu sein. Es sind Menschen, die das machtvolle Tool einsetzen, um ihre eigenen Interessen durchzusetzen. Die hier skizzierten Szenarien der Eskalation beruhen auf bewusst begangenen Rechtsbrüchen z.B. durch Betrug und zielgerichtete Täuschung mit Hilfe von KI. Regulierung kann Kriminalität per se nicht verhindern. Doch wird Regulierung gebraucht, um nicht alle Daten dem freien Gebrauch für wirtschaftliche Partikularinteressen zur Verfügung zu stellen.

Jenseits der Risiken und der Überlegungen zu notwendigen Regulierungen kann aber die rasant an Intelligenz zunehmende KI der Menschheit wertvolle Dienste leisten. Jedenfalls dann, wenn sie nicht bloß zum Vorteil einiger Weniger eingesetzt wird, sondern zum Wohle Vieler.

Literatur

Emilio Ferrara: Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. April 2023. arXiv:2304.03738. https://doi.org/10.48550/arXiv.2304.03738

Future of Life Institute: Pause Giant AI Experiments: An Open Letter. 22.3.2023 https://

futureoflife.org/open-letter/pause-giant-ai-experiments/ [Letzter Zugriff: 1.9.2023].

Yuval Noah Harari: AI and the future of humanity. Vortrag beim Frontiers Forum, April 2023. YouTube-Video https://www.youtube.com/watch?v=LWiM-LuRe6w [Letzter Zugriff: 1.9.2023].

Dan Hendrycks/Mantas Mazeika/Thomas Woodside: An Overview of Catastrophic AI Risks. Xiv:2306.12001v4 [cs.CY] 22 Aug 2023 https://www.safe.ai/ai-risk [Letzter Zugriff: 1.9.2023].

Initiative Urheberrecht: Urheber:innen und Künstler:innen fordern Maßnahmen zum Schutz vor generativer KI in der Europäischen KI Verordnung. https://urheber.info/media/pages/diskurs/ruf-nach-schutz-vor-generativer-ki/03e4ed0ae5-1681902659/finale-fassung_de_urheber-und-kunslter-fordern-schutz-vor-gki_final_19.4.2023_12-50.pdf 19.4.2023 [Letzter Zugriff: 1.9.2023].

Felicitas Wilke: Bewerbungsroboter. Künstliche Intelligenz diskriminiert (noch). Die Zeit. 18.10.2018. https://www.zeit.de/arbeit/2018-10/bewerbungsroboter-kuenstliche-intelligenz-amazon-frauen-diskriminierung [Letzter Zugriff: 1.9.2023].