Narrative der Weltbeglückung

Die neue Sprach-KI und die Mathematisierung der Ethik

Von **Roberto Simanowski**

Seit Wochen ist ein Name in aller Munde, den zuvor kaum einer auszusprechen in der Lage gewesen wäre: ChatGPT. Wir erleben einen gewaltigen Hype um diese neue Sprach-KI, nicht nur das Feuilleton überschlägt sich in historischen Vergleichen – ob mit der Einführung der Druckerpresse, der Erfindung der Elektrizität oder gar der Nutzbarmachung des Feuers.

Und wie sollte man auch nicht begeistert sein? Sprach-KIs wie ChatGPT und GPT-4 schreiben in Sekundenschnelle Artikel, denen man die künstliche Handschrift kaum anmerkt. Sie bestehen Aufnahmeverfahren an Elite-Unis, erstellen Textzusammenfassungen, Versammlungsprotokolle, Diagnoseberichte oder entwerfen das Gerüst für Vorträge, Seminare, ganze Bücher und schreiben sogar Gedichte – ganz zu schweigen von der schon fast trivialen Glückwunschrede zum 50. Betriebsjubiläum einmal im feierlichem, einmal im humorvollen Ton. Kurzum: Sprach-KI verspricht das Ende der Schreibblockade, die Befreiung von mechanischen Denkaufgaben hin zum wirklich Kreativen – als ein höchst effizienter Koautor oder Copilot, wie es bei Microsoft heißt.

Dann aber, am 29. März 2023, warnte ein Offener Brief mit dem imperativen Titel „Pause Giant AI Experiments" vor einem außer Kontrolle geratenen Wettlauf um die Entwicklung und den Einsatz immer leistungsfähigerer KI, die nicht einmal ihre Erfinder verstehen, vorhersagen oder zuverlässig kontrollieren können, und forderte eine sechsmonatige Pause in der Entwicklung von KI, um der Politik Zeit zu geben, mit den nötigen Regulationen hinterherzukommen.[1] Was für Spielverderber! Der Brief kam jedoch nicht von den üblichen Verdächtigen; nicht aus den Departments der Geisteswissenschaften, sondern von KI-Experten, die sich wiederum auf verschiedene Fachartikel von IT-Spezialisten über die Risiken von KI beriefen, in denen auch die alte Gefahr wieder Gestalt annimmt: dass der Mensch von seiner eigenen Schöpfung versklavt werden könnte. Dass auch der „Godfather of AI", Geoffrey Hinton, eine solche Entwicklung für möglich hält, beunruhigt.[2] Noch mehr beunruhigt freilich, dass selbst OpenAI (das Start-up hinter ChatGPT und GPT-4) diese Gefahr nicht ausschließt: Dessen CEO Sam Altman erwartet

1 Vgl. Pause Giant AI Experiments: An Open Letter, www.futureoflife.org, 22.3.2023.
2 Vgl. Will Douglas Heaven, Geoffrey Hinton tells us why he's now scared of the tech he helped build, www.technologyreview.com, 2.5.2023.

neben „disinformation problems“ und „economic shock“ als mögliche Folgen von GPT Probleme, „die weit über das hinausgehen, worauf wir vorbereitet sind“[3] und OpenAIs *GPT-4 Technical Report* vermerkt erste Anzeichen für eine sogenannte „starke“ KI, die eigene Interessen verfolgt.[4]

Natürlich teilen nicht alle und wahrscheinlich nicht einmal die meisten der IT-Spezialisten den Aufruf zu einem Entwicklungsstopp. Stellvertretend sei auf einen der wichtigsten KI-Experten verwiesen, Yann LeCun, Chief AI Scientist bei Meta, der in einem Tweet eine medienhistorische Analogie bemüht: „Wir schreiben das Jahr 1440 und die katholische Kirche hat ein sechsmonatiges Moratorium für den Gebrauch der Druckerpresse erlassen. Man stelle sich nur vor, was passieren könnte, wenn das gemeine Volk Zugang zu Büchern bekäme! Sie könnten die Bibel selbst lesen, und die Gesellschaft würde zerstört werden.“[5] Diese Reaktion ist symptomatisch für das Selbstverständnis vieler KI-Befürworter und ihr Diskussionsniveau: Ein zweifelhafter Vergleich mündet in eine Emanzipationsrhetorik, die noch jene Phrase enthält, ohne die im Silicon Valley kein Produkt-Pitch auskommt: „to make the world a better place“. Kritik am neuen Produkt wird so zur Obstruktion des gesellschaftlichen Fortschritts.

Gewiss, ein gutes Narrativ ist die halbe Miete, wenn man öffentlich ein Produkt verteidigen will, das andere aus verschiedenen Gründen als mehr oder weniger problematisch ansehen. Bei einer Sprach-KI wie GPT reicht die Spannweite des Narrativs von der Erweiterung unseres Wissens (KI als Expertin zu allem und jederzeit) und unserer Kreativität (KI als Ideengeber) bis zur Bekämpfung des Klimawandels (durch effizientere Ressourcenverteilung und Katastrophenmanagement). Ganz zu schweigen von der Lösung des Krebsproblems. Und natürlich hilft es auch gleich noch den Migranten – die dank ChatGPT weniger Nachteile wegen lückenhafter Englischkenntnisse gegenüber Muttersprachlern haben[6] – und sozial Schwachen – die sich mit ChatGPT nun auch einen Tutor leisten können.[7] Wer wollte schon gegen solche Versprechen sein?

»Digital first, Bedenken second«

Wer sich durch die „Californian Ideology“[8] – dieser Verbindung des „freewheeling spirit“ der Hippies mit dem „entrepreneurial zeal“ der Yuppies – nicht das kritische Bewusstsein hat nehmen lassen, wird trotz der Weltverbesserungsrhetorik zunächst daran denken, dass KI vor allem ein riesiges

3 Lex Fridman und Sam Altman, Open AI CEO on GPT-4, Chat-GPT and the Future of AI | Lex Fridman Podcast #367, www.youtube.com, 25.3.2023, ab Min. 1:09:39.
4 Vgl. GPT-4 Technical Report, www.openai.com, 27.3.2023, S. 54 f.
5 Yann LeCun, https://twitter.com/ylecun, 30.3.2023.
6 Vgl. SXSW, OpenAI Co-founder Greg Brockman on Chat GPT, Dall-E and the Impact of Generative AI | SXSW 2023, www.youtube.com, 12.3.2023, ab Min. 18:50.
7 Vgl. Khan Academy, Khan Academy announces GPT-4 powered learning guide, www.youtube.com, 14.3.2023.
8 Vgl. Richard Barbrook und Andy Cameron, The Californian Ideology, in: „Mute Magazine“, abrufbar auf www.imaginaryfutures.net, 1.9.1995; Richard Barbrook: Cyber-Communism: How The Americans Are Superseding Capitalism In Cyberspace, in: „Science as Culture“, 1/1999, S. 5-40.

Geschäft ist. Ein Geschäft, das die Profiterwirtschaftung weiter von menschlichen auf maschinelle Produktivkräfte verschiebt und sehr schnell sehr viel Geld verspricht, weswegen die KI-Forschung in den letzten Jahren immer stärker dem Prozess einer demokratischen Kontrolle entzogen und durch entsprechendes Funding und eine entsprechende Lenkung der Forschungsfragen den Interessen machtvoller Kooperation unterstellt wurde.[9]

Doch auch diese harte ökonomische Realität wird von den Protagonisten der Netzideologie mit leichter Hand überspielt und heroisiert. Eine Reaktion auf LeChuns Tweet verwies darauf, dass der Bau der Florenzer Kathedrale Santa Maria del Fiore 1296 von Leuten begonnen wurde, die nicht wussten, wie sie die Kuppel dazu konstruieren konnten, die erst ein Jahrhundert später mit neuen Einsichten und Werkzeugen errichtet wurde. Diesen Pioniergeist, diese Risikobereitschaft gelte es heute neu zu beleben.[10] Im Geiste der Renaissance wird hier kritisiert, was Mark Zuckerberg als zwanzigjähriger Facebook-Gründer und Harvard-Student einmal als Übervorsicht beschrieb, während es sinnvoller sei, Dinge einfach anzugehen und sich später für Fehlentwicklungen zu entschuldigen.[11] Diesen Ratschlag gab Zuckerberg im Mai 2017 in seiner Commencement-Rede auch den Harvard-Absolventen mit auf den Weg,[12] als er längst zu Facebooks Arbeitsmotto geworden war: „move fast and break things".

Man weiß, wie es weiterging. Im November 2017 zeigte „The Economist" Facebook – 2011 im Kontext des Arabischen Frühlings noch mit dem Begriff „Facebook Revolution" als Werkzeug der Demokratisierung geadelt – auf seinem Cover als „threat to democracy". In den Anhörungen vor US- und EU-Politikern, die Zuckerberg 2018 wegen des Cambridge Analytica-Skandals über sich ergehen lassen musste, wiederholt sich eine Szene immer wieder: Zuckerberg verweist auf die Komplexität von Facebook, die man erst allmählich zu verstehen beginne; es sei dem Facebook-Management vorher nicht klar gewesen, wie leicht das soziale Netzwerk von „bad actors" missbraucht werden könne. So viel zu einem früheren Großunternehmen der Weltverbesserung, das aus der Verbindung aller Menschen ein besseres Verständnis untereinander versprach, sie dann aber eher gegeneinander aufbrachte. Facebooks Management hatte sein Produkt schneller vorangetrieben, als es dessen gesellschaftliche Risiken verstehen und beherrschen konnte. Dass auch Facebook Geschäftsmodell und ökonomische Interessen der Investition in Content Management und Risikokontrolle im Wege standen, ist bekannt und überrascht wenig – und sollte beim aktuellen Weltverbesserungsnarrativ erinnert werden.

Vor diesem Hintergrund lässt sich der Offene Brief auch als Lerneffekt aus der jüngsten Mediengeschichte verstehen, als Abschied vom *move fast*

9 Vgl. Abeba Birhane et al., The Values Encoded in Machine Learning Research, www.dl.acm.org, Juni 2022, S. 173-184.
10 Vgl. François Luc Moraud, www.twitter.com/francoismoraud, 30.3.2023.
11 Vgl. Frontline PBS, The Facebook Dilemma, Part One (full documentary) | FRONTLINE, www.youtube.com, 30.10.2018, ab Min. 4:12.
12 Vgl. ABC News, Mark Zuckerbergs Harvard Commencement Speech 2017 FACEBOOK CEO'S FULL SPEECH, www.youtube.com, 26.5.2017, ab Min. 13:38.

and break things-Fieber hin zu einer *Slow-AI*-Nachhaltigkeit. In Deutschland geht für dieses Entschleunigen die Digitalisierung freilich viel zu langsam voran. Wenn Daten zur Pandemiebekämpfung noch per Fax verschickt werden und Schulen das Internet oder die Medienkompetenz fehlt, um ihren Bildungsauftrag auch in Krisenzeiten zu erfüllen, haben es Losungen gegen Bürokraten, Bedenkenträger und Bremser leicht, wie etwa: „Digital First, Bedenken Second" (FDP im Wahlkampf 2017), Digitalisierung mit „maximalem Tempo" und „ohne Wenn und Aber" (Bitkom-Präsident Thorsten Dirks, 2017),[13] „Digitalisierung der Schulen von null auf hundert beschleunigen, und das von jetzt auf gleich" (Bitkom-Präsident Achim Berg, 2020)[14] oder auch „Wir müssen jetzt KI mit voller Kraft umarmen" (Sascha Lobo, 2023).[15]

Natürlich darf man nicht übersehen, wer solche Losungen vertritt. Wirtschaftskräfte, die neue Produkte schaffen und neu Märkte erobern wollen, haben naturgemäß andere Interessen als Soziologinnen oder Medien- und Kulturwissenschaftler, die nach den gesellschaftlichen Folgen der Digitalisierung fragen, oder als Klimaexperten, die gegen die heilige Kuh der Wachstumslogik argumentieren. Auch einflussreiche „Digitalexperten", die ihr Geld damit verdienen, der Wirtschaft das Internet zu erklären, werden Bedenken lediglich zur Abrundung einer ansonsten optimistischen Grundhaltung einstreuen. Brisant wird es freilich, wenn diese Perspektive auch außerhalb des Beratungsjobs, in Zeitungbeiträgen und Talkshow-Auftritten, das allgemeine Problembewusstsein bestimmt, die Gesellschaft also selbst im Rahmen der sogenannten Vierten Gewalt ganz im Sinne des Wirtschaftssystems informiert wird. Brisant ist dies zumal dann, wenn es sich um Wiederholungstäter handelt, die mit dem gleichen naiven Optimismus schon einmal gründlich reingefallen waren, als sie das Internet als das „perfekte Medium der Demokratie, der Emanzipation, der Selbstbefreiung" umarmten, bis sie sich im Kontext der NSA-Affäre von ihm „verletzt" fühlten und ihr Analyseversagen zu einer „vierten, digitalen Kränkung der Menschheit" aufbauschten.[16] Schon damals hätte man es besser wissen können, hätte man sich auf den Diskussionsstand der akademischen Internetforschung gebracht oder auch nur die Berichte in den Medien beachtet, die weniger euphorisch waren als man selbst. Um so wichtiger ist es, jetzt, in der Debatte um die neue Sprach-KI, diesem Irrtum nicht vor lauter Umarmungslust erneut zu verfallen.

Wie werden wir denken?

Die Gefahr, die der erwähnte Offene Brief anspricht, zielt auf eine starke KI, die der Mensch – eines Tages – vielleicht nicht mehr kontrollieren kann. Kritiker sehen (neben dem Umstand, dass gerade OpenAI-Mitbegründer Elon

13 Bundesministerium für Wirtschaft und Klimaschutz, Digital-Gipfel 2017: Keynotes von Bundeskanzlerin Angela Merkel und Bitkom-Präsident Thorsten Dirks, www.de.digital, 13.6.2017.
14 Achim Berg, Bitkom zur Digitalisierung der Schulen nach Corona, www.bitkom.org. 6.5.2020.
15 Sascha Lobo im Gespräch mit dem ZDF, „KI mit voller Kraft umarmen", www.zdf.de, 6.4.2023.
16 Sascha Lobo, Abschied von der Utopie: Die digitale Kränkung des Menschen, www.faz.net, 11.1.2014.

Musk sich hier als Retter positioniert) in diesem „longtermism" eines Machtkampfes zwischen Mensch und KI allerdings primär ein Ablenkungsmanöver von den Problemen, die bereits diesseits einer solchen KI aufgrund politisch-ökonomischer Machtverhältnisse bestehen.[17] Heute gehe es vor allem darum, die Herkunft der Daten, an denen die KI trainiert wurde, transparent zu machen, sowie um das, was in der Fachwelt unter dem Begriff „Data Colonialism" diskutiert wird: die Ausbeutung der Datenbeschaffung (Data-Labeling durch prekäre Klickworker im Globalen Süden) und die Profitgenerierung für IT-Unternehmen aus der Extraktion der Daten-Allmende (inklusive des Kulturguts des Globalen Nordens).[18] In die Kategorie der ganz konkreten politisch-ökonomischen Machtverhältnisse gehören zudem die Risiken, die von der KI für die Beziehung zwischen Staaten (wer KI beherrscht, beherrscht die Welt), Staat und Wirtschaft (Machtgewinn der IT-Unternehmen gegenüber dem Staat) sowie Staat und Bürgern (Kontrolle der Bürger durch den Staat) ausgehen.[19]

Neben diesen in der Tat dramatischen, aber eher unterschwellig auftretenden Problemen gibt es zugleich jene, über die derzeit alle sprechen: Copyright, Plagiarismus, Arbeitsplatzvernichtung, Halluzinationen, Deepfakes. Diese Themen bestimmen schon deshalb die Talkshows, weil Halluzinationen und Deepfakes einen gewissen Unterhaltungswert haben und es sich wunderbar darüber streiten lässt, wem nun der Text gehört, den GPT aus fremden Daten generiert – und ob es nun die Anwälte sind oder die Journalisten oder die Programmierer, die zuerst ihren Job verlieren. Die meisten dieser Probleme sind jedoch bloß Anpassungsprobleme der Gesellschaft an die neue Technologie, wobei vor allem das Rechts- und Bildungswesen in der Pflicht stehen. Nur das Halluzinationsproblem richtet sich tatsächlich an die KI-Entwickler, darf aber als „Noch-Problem" gelten, das im Zuge der technischen Entwicklung der Sprach-KI bald verschwinden dürfte.

Weit weniger eindeutig ist die Problemlage bei der Delegation kognitiver Tätigkeiten an die KI, worin nicht nur der Deutsche Ethikrat einen Verlust menschlicher Fähigkeiten fürchtet.[20] Im Schreiben merkt der Mensch, wie sich die Informationen, über die er verfügt, und die Ansichten, die er besitzt, sinnvoll kombinieren lassen und auf etwas hinführen, das sich Erkenntnis oder Einsicht nennen lässt. Im Schreiben kommt der Mensch zu sich. Delegiert er diese Erfahrung an die KI, verwandelt sich der Produktionsprozess zurück in einen Rezeptionsprozess: Man bleibt Leser, nämlich der Synthese an Informationen und Ansichten, über die nicht man selbst verfügt, sondern die KI. Insofern ist in der Tat ein Verlust kognitiver Fähigkeiten zu fürchten, was die medienphilosophische These bestätigen würde, dass jede Techno-

17 Vgl. Timnit Gebru et al., Statement from the listed authors of Stochastic Parrots on the „AI pause" letter, www.dair-institute.org, 31.3.2023.
18 Vgl. Nick Couldry und Ulises Ali Mejias, The Costs of Connection. How Data Is Colonizing Human Life and Appropriating It for Capitalism, Stanford University Press 2019.
19 Vgl. Benjamin S. Bucknall und Shiri Dori-Hacohen, Current and Near-Term AI as a Potential Existential Risk Factor, AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, www.acm.org, Juli 2022.
20 Deutscher Ethikrat, Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz. Stellungnahme, www.ethikrat.org, 20.3.2023, S. 120.

logie zugleich eine Erweiterung und eine Einschränkung des Menschen ist: Die Schrift verbessert die Aufbewahrung von Information und verringert das Erinnerungsvermögen, so wie der Taschenrechner nicht das Kopfrechnen und das Navi nicht das Orientierungsvermögen verbessern.

Ob wir auch schlechter denken werden, wenn GPT uns in Zukunft beim Schreiben und Konzipieren hilft, ist vorerst noch offen. Es hängt wesentlich davon ab, wie viel davon an die KI delegiert wird und wie sehr man dem „Nudge" des generierten Textes folgt, also sein Angebot, eine bestimmte Sache zu sehen, übernimmt. Gerade wenn es schnell gehen muss und man – unter zunehmendem ökonomischen Druck – für die Textproduktion immer weniger Zeit einplant, besteht darin eine wirkliche Gefahr. Doch GPT kann den Denkprozess durchaus auch schärfen und inspirieren, indem man es durch bestimmte Prompts künstlich Diskursschnittpunkte herstellen lässt („diskutiere A aus der Perspektive von B im Stil von C!").

Das Medium als Botschaft: Wes Geistes Kind ist die KI?

Hier liegt denn auch das Problem mit der wohl am häufigsten gezogenen Analogie, nämlich der zur Erfindung der Druckerpresse. Kritiker dieser Analogie verweisen zu Recht darauf, dass im Gegensatz zur KI die Druckerpresse weder eine „Black Box" war (deren Prozesse zwischen Input und Output selbst für Programmierer im Dunkeln liegen) noch „agentic" (also eigene Interessen verfolgte). Doch so richtig und nötig diese Korrektur auch ist, sie übersieht die eigentliche Differenz zwischen KI und Druckerpresse.

Diese Differenz führt zu der vielleicht bekanntesten medienphilosophischen These: „The medium is the message" (Marshall McLuhan). Demzufolge verbreiten Medien nicht nur Ideen oder Inhalte, sondern sind selbst eine Idee oder Botschaft (also *agentic*), mit der sie die Situation des Menschen ändern. So bringt das Auto den Individualverkehr mit sich und so verbinden die sozialen Netzwerke Menschen über Kontinente hinweg. Zugleich ändern die Medien – das ist ihre sekundäre, weniger offensichtliche Botschaft – die Situation des Menschen oft in einer nicht antizipierten Art und Weise: Das Auto führt zu Vorstädten, das Web 2.0 zu einer Kultur der Selbstdarstellung unter den Bedingungen der Aufmerksamkeitsökonomie. Diese Botschaften sind der zentrale Gegenstand der Medienwissenschaft, während der von den Medien vermittelte Inhalt eher als ein „saftiges Stück Fleisch" verstanden wird, „das der Einbrecher mit sich führt, um die Aufmerksamkeit des Wachhundes abzulenken".[21] Anders gesagt: Wenn die Medienwissenschaft über die Fotografie spricht, spricht sie nicht primär über die Inhalte der Fotos, sondern darüber, wie die Kamera die Wahrnehmung der Welt ändert.

Anders ist es bei einer KI wie GPT. Sie ändert die Situation des Menschen nicht nur durch die Automatisierung seiner kognitiven Prozesse. Die für die Automatisierung typische Nebenfolge der Standardisierung führt bei einer

21 Marshall McLuhan, Die magischen Kanäle. Understanding Media, Düsseldorf 1992, S. 29.

auf dem Wahrscheinlichkeitsprinzip basierenden Sprach-KI dazu, dass in den von ihr produzierten Texten immer das gilt, was in ihren Trainingsdaten die Mehrheit auf seiner Seite hat. Die Standardisierung – oder eben Denkweise – erfolgt im Sinne des Mainstreams und materialisiert sich in den von der KI vermittelten Inhalten. Die Weltanschauung der KI-Texte hängt also nicht wie bei der Druckerpresse von der Autorin des gedruckten Textes ab und variiert entsprechend, sie ist Teil des Mediums selbst und wird überall dorthin exportiert, wo dieses zum Einsatz kommt.

Ein berühmtes Beispiel ist die Zusammenfassung eines australischen Gesetzesentwurfs für schärfere Waffengesetze durch ein vor allem an amerikanischen Daten trainiertes GPT-3-Programm, das – entgegen dem Geist des Gesetzesentwurfs, aber ganz im Sinne des 2. Zusatzartikels der US-Verfassung – vor dem Verlust des Rechts auf Selbstverteidigung warnt.[22] Hier zeigt sich: Die weltanschauliche Ausrichtung der Sprach-KI ist ebenso wie der Mensch das Ergebnis der Daten, mit denen sie „aufwächst". Der Fachartikel, der diese Verzerrung berichtet, trägt den Titel „The Ghost in the Machine has an American accent". Dass dieser Akzent nicht wirklich zu hören ist, gehört zu den strukturellen Problemen jeder Sprach-KI und zu ihrer beunruhigenden Unheimlichkeit: Wir wissen nicht, wessen Geistes Kind die KI ist, die da unsere Fragen beantwortet und unsere Texte schreibt.

Die Fachwelt debattiert das „Akzent-Problem" der KI unter Stichworten wie „decolonial computing", „algoritmic reparation" und „value alignment", die alle auf eine genaue Kuratierung der Daten zielen, die in das Trainingsset der KI eingehen. Denn wer die Macht über die Trainingsdaten hat, bestimmt das „Denken" der KI und damit auch das Denken derer, die unter ihrem Einfluss stehen. Während den einen genügt, dass alle Daten im Internet gleichermaßen berücksichtigt werden, sehen andere darin nur die Fortschreibung der disproportionalen Bevölkerungsverhältnisse online – die Dominanz junger, weißer Männer – und fordern eine statistische Verrechnung gemäß der Offline-Bevölkerung. Auch das aber wäre bloß die Fortschreibung der Machtverhältnisse in der Offline-Welt, sagen die Vertreter des „decolonial computing" und der „algoritmic reparation". Erstere wenden sich gegen die koloniale Auslöschung nicht-westlicher Seins- und Wissensformen,[23] letztere fordern die Bevorzugung der Daten jener Gruppen, die in der Gesellschaft (auch des Globalen Nordens) bisher zu wenig zu Wort kamen: als eine Art „affirmative action" oder Quotenregelung. Der Ansatz der Reparatur bzw. Reparation fragt also nicht, ob eine KI gut ist oder gerecht, sondern ob sie hilft, die bestehenden Machtverhältnisse zu verändern.[24]

KI wird damit tatsächlich zu einem Mittel, „to make the world a better place". Die mit „Quoten-Daten" gefütterte oder auch nachträglich auf einer höheren Programmierebene entsprechend ausgerichtete KI repräsentiert

22 Vgl. Rebecca L. Johnson, Giada Pistilli et al., The Ghost in the Machine has an American accent: value conflict in GPT-3, www.arxiv.org.
23 Vgl. Decolonial AI Manyfesto, www.manyfesto.ai.
24 Vgl. Jenny L. Davis, Apryl Williams und Michael W. Yang, Algorithmic reparation, in: "Big Data & Society" 7-9/2021; Pratyusha Kalluri, Don't ask if artificial intelligence is good or fair, ask how it shifts power, in: „Nature", 7.7.2020, S. 169.

nicht den Status quo der Welt, sie repräsentiert die Welt, wie sie – um nur ein mögliches Beispiel zu nennen –, aus der Perspektive der Critical Race Theory sein sollte, und wird also gendern, politisch korrekt sein und verlässlich die Rechte der Minderheiten vertreten. Es ist die Fortführung einer engagierten Repräsentationspolitik mit technischen Mitteln, die im Grunde schon Praxis ist, wenn zum Beispiel der Sprachassistent von Google Docs mit genderneutraler Sprache nudged – indem bei „chairman" der Vorschlag „chairperson" und bei „mailman" der Vorschlag "mail carrier" erscheint – oder wenn ChatGPT keine Liste an Schimpfnamen für Deutsche erstellen will, weil es darauf ausgerichtet sei, eine „respektvolle und sachliche Kommunikation zu fördern", die verbiete, „jemanden aufgrund seiner Nationalität, Ethnizität oder eines anderen Merkmals herabzusetzen" – wogegen dann auch die Erklärung nicht hilft, dass man diese Liste doch für eine kritische Arbeit über nationale Vorurteile benötige.

Es ist keineswegs auszuschließen, dass die in der Gesellschaft nicht unumstrittene Critical Race Theory künftig ihre eigene Sprach-KI auf den Markt bringt. Das rechtsgerichtete Netzwerk Gab bastelt bereits an einer Alternative zu OpenAIs GPT – „without the constraints of liberal propaganda" – und die Suchmaschine Brave hat bereits ihren eigenen Chatbot integriert, der die Informationen des Web mit rechtslastigem Akzent wiedergibt.[25] Wie Experimente zeigen, ist die nachträgliche ideologische Ausrichtung von Chatbots auf dem Fundament einer existierenden Sprach-KI weder zeitaufwendig noch kostenintensiv und könnte sich leicht als neue Front des Kulturkampfs erweisen.[26] Die Frage ist also nicht, ob es möglich ist, GPT oder wie immer die Sprach-KI heißt, die künftig unsere Fragen beantwortet und unsere Texte schreibt, mit einer bestimmten Weltanschauung auszustatten, sondern ob dies wünschenswert wäre. Das führt zum nächsten zentralen Stichwort der KI-Ethikdebatte: „value alignment".

Welche Werte soll die KI haben?

Die Frage, mit welchen Werten eine global operierende KI ausgestattet sein sollte, führt schnell vom politischen aufs philosophische Terrain. Unklar ist bereits, ob die Werte durch Vernunft oder Beobachtung erschlossen werden sollen. Unklar ist auch, ob im Falle der empirischen Methodik die Menschen vor sich selbst (ungesunde Gewohnheiten, Stichwort „present bias") geschützt werden sollten, und wie sich die verschiedenen Beobachtungen in verschiedenen kulturellen, religiösen und politischen Systemen auf einen Nenner bringen lassen.[27] So überrascht es nicht, dass die Debatte regelmäßig zur Einsicht ihrer Unlösbarkeit oder zur Absage an *eine* Lösung weltweit führt. Bleibt die Frage, welches Wertesystem dann aber global operierende KIs wie GPT-4

25 Stuart A. Thompson, Tiffany Hsu und Steven Lee Myers, Conservatives Aim to Build a Chatbot of Their Own, www.nytimes.com, 22.3.2023.
26 Vgl. David Rozado, The Political Biases of ChatGPT, in: „Social Sciences" 2023, S. 148.
27 Vgl. Iason Gabriel, Artificial intelligence, values, and alignment, in: „Minds and Machines", 3/2020, S. 411-437.

in Microsofts Bing oder LaMDA in Googles Bard bestimmen soll. Dass Chinas Chatbots die „socialist core values" widerspiegeln müssen, wird niemanden überraschen.[28] Was aber ist mit den chatbots der USA? Oder Europas? Oder dem „Rest der Welt"? Welchen Akzent wird die KI dort haben?

OpenAIs CEO hofft diesbezüglich auf eine Zukunft, wo der Zugang zu KI „superdemokratisiert" ist, wo es mehrere KIs bzw. AGIs gibt, die mehrere Blickwinkel zulassen.[29] Jede Gesellschaft solle der KI sagen können, an welche Werte sie sich halten soll. Allerdings belässt es Altman nicht bei gesellschaftlich programmierten KIs, sondern gesteht schließlich jedem Individuum die Möglichkeit des Fine-Tunings durch die Eingabe seiner persönlichen Werte zu.[30] Altman wiederholt diese Position später mit dem Hinweis, dass GPT-4 bereits mit der Funktion „system message" ausgestattet ist, die dem Nutzer ein Fine-Tuning der Perspektive (des Akzents) von GPT-4 ermöglicht.[31]

Jedem am Ende also seine eigene KI? Das scheint dem Emanzipationsnarrativ zu entsprechen, dass alle Entscheidungsmacht beim Individuum liegt, ist im Grunde aber nicht mehr als die ultimative Ausweitung der Filterblase, die sich ja schon bei den sozialen Netzwerken als äußerst problematisch erwiesen hat. Altmans Vision erinnert damit an einen ähnlichen, früheren Vorschlag im Kontext der sozialen Netzwerke, der ebenfalls die Frage der Wertevielfalt in einem globalen Medium durch Personalisierung zu lösen suchte.

In seinem Manifest *Building Global Communit*y vom 17. Februar 2017 schlägt Mark Zuckerberg angesichts der Unterschiede nicht nur zwischen den Kulturen, sondern auch innerhalb einer Kultur vor, dass die Nutzer selbst entscheiden sollen, wie viel Nacktheit, Gewalt und Profanität sie in ihrem News Feed sehen wollen. Ziel ist ein „system of personal control over our experience".[32] Zur Ironie dieses Vorschlags, den gordischen Knoten der Wertepräferenzen zu lösen, gehört, dass gerade die vorgeschlagene „self-governance" ein zutiefst westlicher Ansatz ist: die höchstmögliche Freiheit des Individuums vom kulturellen Kontext und dessen Glaubens- und Wertesystemen. Dazu kam es auf Facebook allerdings nie. Nur anderthalb Jahre nach dem Manifest legt Zuckerberg denn auch ein *Blueprint for Content Governance and Enforcement* nach, in dem er erklärt, dass „local content laws" befolgt werden müssen, und sich von den lokalen Regierungen Anweisungen wünscht, wie sie sich die Content-Moderation in ihren Ländern vorstellen.[33] Dieser Schritt zurück von der Personalisierung des Informationsmanagements in Richtung Informationshoheit des Staates entspricht der Renationali-

28 Chang Che, China Says Chatbots Must Toe the Party Line, www.nytimes.com, 24.4.2023.
29 Connie Loizos, StrictlyVC in conversation with Sam Altman, www.youtube.com, 18.1.2023, Min. 7:40-7:58.
30 Ebd., ab Min. 10:10; Irene Solaiman und Christy Dennison, Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets, 35th Conference on Neural Information Processing Systems (NeurIPS 2021), www.arxiv.org.
31 Vgl. Lex Fridman, a.a.O., ab Min. 26:18.
32 Mark Zuckerberg, Building Global Community, www.facebook.com/mark-zuckerberg; vgl. dazu ausführlich meinen Beitrag „Die Facebook-Utopie. Wie Mark Zuckerberg die Welt retten will", in: „Blätter", 9/2017, S. 109-119.
33 Mark Zuckerberg, A Blueprint for Content Governance and Enforcement, https://www.facebook.com/mark-zuckerberg.

sierung des Internets durch strengere nationale Auflagen nicht nur in China oder Russland, sondern auch in Deutschland; und er bezeugt, wie flexibel, um es freundlich auszudrücken, Zuckerberg die Frage behandelt, wer am Ende bestimmen soll, welche Inhalte dem Nutzer gezeigt werden.

Vergleichbar wenig scheint heute Altman zu wissen, was er hinsichtlich der Werteausrichtung der KI eigentlich will. Sein „dream scenario" jedenfalls besteht dann plötzlich doch nicht mehr in der Personalisierung der KI, sondern in der kollektiven Entscheidung über deren Werte, vergleichbar der US-amerikanischen Verfassung.[34]

Da diese Einigung jedoch sehr unwahrscheinlich erscheint – allemal auf der Ebene der Vereinten Nationen und selbst auf der Ebene des Nationalstaats –, bleibt es bei der Frage, wer am Ende bestimmt, mit welchem Akzent die KI, die unsere kognitiven Prozesse übernimmt, spricht. Wird es gegen den Akzent der USA und Chinas einen EU-Akzent geben, der dann gewiss weniger das Recht auf Waffenbesitz verteidigt als das Recht auf Abtreibung?

Warten kann man jedenfalls nicht, bis die KI selbst Vorschläge macht, mit welchen Werten sie ausgestattet sein sollte, wie es in der Alignment-Debatte gelegentlich empfohlen wird – und zwar mit dem Argument, dass die KI auf Grund ihrer immensen Datenverarbeitungskapazität für alle konkurrierenden Wertesysteme Zukunftsszenarien durchspielen und so herausfinden könnte, welches System oder welche Mischung verschiedener Systeme das Wohl der Menschheit am effektivsten befördert.[35]

Sprach-KI als Dilemma-Technologie

An diesem Punkt wird deutlich, was Sprach-KI von allen vorangegangenen Technologien unterscheidet: Sie verändert die Situation des Menschen durch die Automatisierung kognitiver Prozesse mit der Nebenfolge der Standardisierung, sofern jeweils die gleichen Daten zugrunde liegen und keine Variation von außen (durch raffinierte Prompts etwa) induziert wird. Darüber hinaus bestimmt sie den kognitiven Prozess durch die Wertvorstellungen, die sie enthält und ihren ahnungslosen Nutzern aufdrängt.[36] Sprach-KI exportiert also nicht nur Handlungsmuster, sondern auch Sichtweisen. Nie zuvor war eine Technologie so direkt mit Machtstrukturen verquickt wie sie. Dass dieser Umstand kein Thema in Talkshows zu GPT ist, erstaunt in Zeiten, da im Namen der Identitäts- und Repräsentationspolitik immer dringlicher gefragt wird, wer da eigentlich spricht und mit welchem Mandat und inwiefern damit die Perspektive einer dominanten Gruppe dem Rest der Gesellschaft aufgedrängt wird. Ist es genau dieses Dilemma der Werteausrichtung, das man in der aktuellen KI-Debatte offensiv zu diskutieren sich scheut?

34 Lex Fridman, a.a.O., ab Min. 35:00.
35 Vgl. William MacAskill in Lucas Perrys *AI Alignment Podcast*, Episode „Moral Uncertainty and the Path to AI Alignment", www.futureoflife.org, 18.9.2018.
36 Maurice Jakesch et al., Co-Writing with Opinionated Language Models Affects Users Views, CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, New York, 19.4.2023.

Denn um ein Dilemma handelt es sich und es ist nicht das einzige, das die Entwicklung Künstlicher Intelligenz mit sich bringt. Erst war es die Automatisierung des Autofahrens durch KI, womit die Entscheidung über Leben und Tod im Falle eines Unfalls vorab programmiert werden muss: Nach welchen Werten soll man die KI im Auto ausrichten? Soll sie immer das Kind retten und die Senioren überfahren?[37] Jetzt ist es die Automatisierung kognitiver Prozesse, die verlangt, dass Entscheidungen über die angemessene Form der Kommunikation vorab getroffen werden, statt situationsbedingt ausgehandelt zu werden. Und wie es sich für ein Dilemma gehört, gibt es darauf keine richtige Antwort, nicht, was die Triage-Regel in einem Unfall betrifft, und nicht, was die richtige Weltanschauung einer Sprach-KI betrifft.

Vielleicht ist das die eigentliche, generelle Botschaft der KI: dass der Mensch Dinge festlegen muss, die er bisher nicht festlegen musste. Automatisierung verlangt Standardisierung, und zwar auch, wenn es kognitive Prozesse sind, die automatisiert werden. Das Dilemma besteht darin, dass wir eigentlich nicht wissen, wie wir die KI im Auto oder Sprach-KIs wie GPT ausrichten sollen, während wir zugleich aber auch wissen, dass wir sie nicht nicht ausrichten können.

Unterbleibt die gezielte Ausrichtung, wird der Akzent der KI dem überlassen, was in den Trainingsdaten die Mehrheit auf seiner Seite hat. Im besten Falle repräsentiert diese Mehrheit zugleich die Mehrheit der Menschheit. Und im besten Falle gilt diese empirische Werteausrichtung auch für das RLHF-Verfahren (Reinforcement Learning from Human Feedback), das heute bei der Ausrichtung von GPT hilft, wenn Millionen an Nutzern im Sinne des Crowdsourcing dessen Output bewerten. OpenAI preist dieses Vorgehen als Kollektivierung der Werteausrichtung und schreibt es sich als Vermeidung von Machtkonzentration zugute.[38] Man darf aber nicht übersehen, dass es sich dabei um eine Bewertung ohne Argumentation handelt, eher wie die Stimmenabgabe bei einer Wahl, weswegen auch nicht das bessere Argument siegt, sondern einfach die Mehrzahl der Hände. Vielleicht liegt hier die eigentliche Botschaft der KI, mit der sie die Situation des Menschen ändert: die Mathematisierung nicht nur der Moral, sondern auch der Ethik.

Angesichts all dieser mehr oder weniger verborgenen Gefahren und jener Unwägbarkeiten, „die weit über das hinausgehen, worauf wir vorbereitet sind“, wie OpenAI-CEO Altman es formuliert, ist es durchaus verständlich, dass das Europaparlament ChatGPT und vergleichbare Sprach-KIs mit Skepsis betrachtet und als Hochrisikotechnologie einstufen will, was unter anderem bedeuten würde, dass seine Trainingsdaten offengelegt werden müssen. Wer dies dem Parlament im Namen des Wirtschaftsstandortes Europa als Regulierungswut und Fortschrittsfeindlichkeit zum Vorwurf macht, hat wenig aus der Geschichte der digitalen Medien gelernt – und könnte sich eines Tages von der KI genauso enttäuscht und gekränkt fühlen wie zuvor vom Internet.[39]

37 Vgl. Roberto Simanowski, Todesalgorithmus. Das Dilemma der künstlichen Intelligenz, Wien 2020.
38 OpenAI, How should AI systems behave, and who should decide?, www.openai.com, 16.2.2023.
39 Sascha Lobo, Die Panik der Politik vor der künstlichen Intelligenz, www.spiegel.de, 3.5.2023.