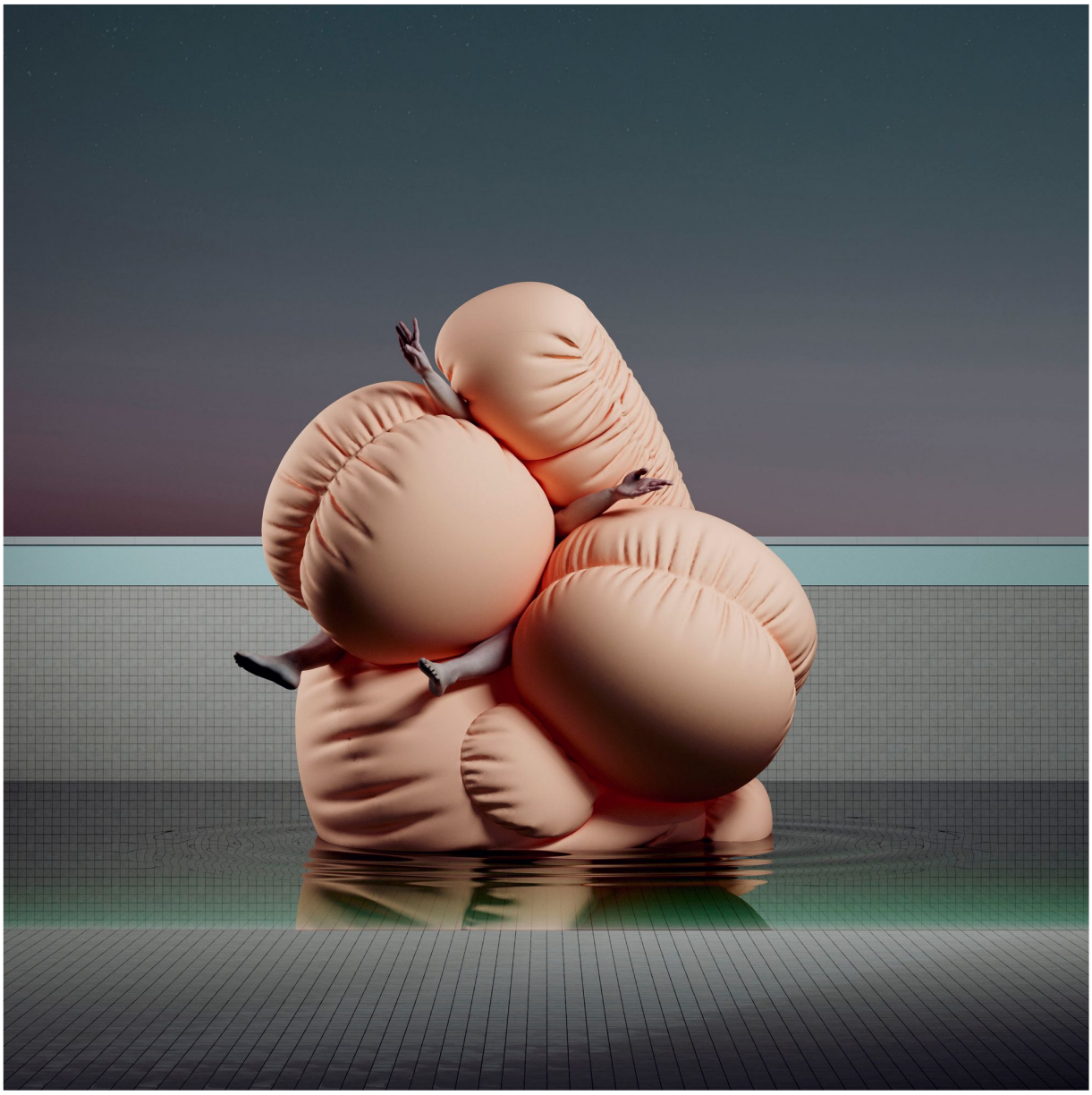
Apokalypse als Businessmodell

Das Silicon Valley macht künstliche Intelligenz zur Zukunftstechnologie – und warnt gleichzeitig vor ihren menschheitsbedrohenden Gefahren. Was wird hier gespielt?

Von Felix Maschewski und Anna-Verena Nosthoff, 11.10.2023



Borja Alegre

Wer sich derzeit von den allgegenwärtigen Katastrophen – von der Klimaerhitzung bis zu den geopolitischen Verwerfungen – emotional beschwert fühlt, der sollte um den aktuellen Diskurs um künstliche Intelligenz (KI)

lieber einen Bogen machen. Wurden technische Entwicklungen wie die Digitalisierung, das Smartphone oder Social Media in der Regel mit den Verheissungen eines besseren Lebens assoziiert, scheint es zur Stimmung der heutigen Zeit zu gehören, dass zwischen KI-Innovation und Untergangsszenario kaum noch zu unterscheiden ist.

Dafür verantwortlich sind weder kulturpessimistische Abgesänge noch esoterisch-technophobe Apokalyptiker, sondern gerade diejenigen, die Chat GPT und Co. erschaffen haben oder gar an kognitiv-menschenähnlicher, «starker KI» und «Superintelligenz» forschen – sogenannter «*artificial general intelligence*» (AGI) oder zu Deutsch: allgemeiner künstlicher Intelligenz.

Open-AI-Chef Sam Altman, Deepmind-CEO Demis Hassabis oder die *Godfathers of AI* Geoffrey Hinton (Universität Toronto / Google) und Yoshua Bengio (Universität Montreal) haben nicht nur faszinierende KI-Systeme entwickelt. Wie moderne Frankensteins warnen sie gleichzeitig auch vor ihren eigenen «Monstern», unterschreiben Statements, die mal auf «existenzielle Risiken», mal auf eine potenzielle «Auslöschung der Menschheit» aufmerksam machen. Und sie fordern eine Prioritätenverschiebung: Die ungezügelte Entwicklung der KI berge mindestens so viel Gefahrenpotenzial wie globale Pandemien oder nukleare Kriege.

Ein Businessmodell

Einzelne Forscher wie Eliezer Yudkowsky legen bereits allerlei Endgame-Szenarien aus der klassischen Science-Fiction neu auf, andere suchen nach «Lösungen». Sie pochen auf eine weltumspannende Governance und versuchen, das langfristige Überleben der Menschheit – auch als *longtermism* bekannt – in den Fokus zu rücken. Dafür forderten führende KI-Entwickler und CEOs wie Elon Musk in einem offenen Brief im letzten März ein Moratorium, also eine zeitweilige Unterbrechung der Technologieentwicklung, damit Risiken und Nebenwirkungen analysiert, Leitplanken für eine «sichere KI» entwickelt und die übermenschliche Smartness frühzeitig eingehegt beziehungsweise reguliert werden könne.

Das klang nicht unvernünftig, doch nebst der Sorge um die Zukunft des Menschengeschlechts wurde auch eine gewisse Doppelzüngigkeit augenscheinlich.

So warb etwa Open-AI-Chef Sam Altman bei europäischen Staatsmännern und -frauen publikumswirksam für mehr Regulierung und tourte in dieser Mission von Podium zu Podium, bis durch Enthüllungen des «Time Magazine» offenbar wurde, was viele vermuteten: Altman hatte im Verborgenen gegen stärkere Regeln lobbyiert und geholfen, den «AI Act» der EU – ein geplantes Gesetz zur Regulierung von KI – zu verwässern. Eine bemerkenswerte Diskrepanz zwischen offenen Briefen und geschlossenen Hinterzimmern, zwischen existenzieller Sorge und schnödem Profitinteresse.

Auch das eigentlich geforderte Moratorium wurde schon rasch durch eine Art Wettrüsten von Open AI, Google, Meta und Co. ersetzt: Die grossen Player haben immer neue Produkte auf den Markt geworfen und damit immer neue Fakten geschaffen. Zuletzt wurde gar die Federal Trade Commission (FTC) in den USA nervös, sandte Open AI wegen Chat GPT einen langen Fragenkatalog, verlangte Einsicht in Dokumente und Details über die Verarbeitung persönlicher Daten. In scheinbar vorauseilendem Gehorsam verkündeten die Konzerne kurz darauf selbstverpflichtende Massnahmen, die

ihre Systeme nicht nur sicherer, sondern auch transparenter machen und gesellschaftliche KI-Risiken minimieren sollen – ein Vorgehen, das man aus der Vergangenheit kennt und das eher auf PR-strategisches Kalkül als auf strenge Regelauslegung schliessen lässt.

An der apokalyptischen Tonlage hat sich jedoch wenig geändert.

Bereits Anfang Juli setzte Open AI ein sogenanntes Superalignment-Team ein, eine firmeneigene, recht homogen besetzte – nur männliche – Taskforce, die den Auftrag hat, innerhalb der nächsten vier Jahre das Problem der «Entmachtung» beziehungsweise «Auslöschung der Menschheit» zu lösen. Natürlich wolle man dabei weiterhin an einer AGI arbeiten, die nun allerdings «sämtlichen Aspekten des Lebens nützen» und vor allem «verantwortlich entwickelt und eingesetzt werden» solle.

Eine ähnliche Route scheint auch der Mitgründer von Open AI, Elon Musk, mit seiner neuen Firma xAI einzuschlagen. Musk, der sich als Doomsday-Prophet gerne ausbreitet über das «existenzielle Risiko» der Superintelligenz – das ominöse «X» in seinen Unternehmen (SpaceX, X – früher Twitter – etc.) lässt dies anklingen –, warnt schon seit Jahren vor dem technologischen Kontrollverlust und will mit xAI und einem wiederum ausschliesslich männlichen Team nun ebenfalls die Entwicklung einer «guten AGI» forcieren. Für jemanden, der kaum etwas lieber tut, als über gesellschaftliche Zusammenbrüche oder den Weltenbrand – mal durch niedrige Geburtenraten, mal durch einen Asteroideneinschlag oder das Verglühen der Sonne – zu posten, klang diese Botschaft dann fast hoffnungsvoll: *«I think to a super-intelligence, humanity is much more interesting than not [-having] humanity.»* (Ich glaube, für eine Superintelligenz ist es viel interessanter, eine Menschheit zu haben, als keine Menschheit zu haben.)

Wie jedoch eine «sichere KI» oder «gute AGI» aussehen soll, geschweige denn wie sie durch eine globale Governance geregelt werden müsste, bleibt trotz all der Unterschriftensammlungen, Ankündigungen und Neugründungen weiterhin recht unklar.

Sowohl xAI als auch das «Superalignment»-Team von Open AI suchen noch – nicht unironisch – nach *the world's best minds* für die superintelligenten Gefahren. Die Warnungen voller überspannter Unschärfen gleichen einem Doomsday-Marketing, das Gefahren heraufbeschwört, die es wahrscheinlich gar nicht gibt, um Produkte zu verkaufen, die noch entwickelt werden müssen. Nur eines scheint dabei sicher: Die Investorinnen stehen Schlange – Open AI hat jetzt schon einen Marktwert von bis zu 90-Milliarden US-Dollar erreicht.

Science-Fiction und Moral

Hinsichtlich der Finanz- und Diskursmacht von Musk, Altman und Co. lohnt es sich, genauer auf die zentralen Akteure und Begrifflichkeiten zu schauen, die das technosoziale Imaginäre so wirkmächtig prägen und die Genese der ebenso apokalyptischen wie profitablen Zukunftsvisionen erklären können. Augenfällig ist jedenfalls, dass hinter dem Doomsday-Hype und der öffentlich zelebrierten Besorgnis ein ganzes Netzwerk an Institutionen steht, das die Entwicklung und die «Ausrichtung transformativer Technologien» wie auch die «Minderung des Auslöschungsrisikos durch AI» zu seiner eigenen Sache macht. So etwa das Center for AI Safety (auch involviert bei xAI) oder die beiden von Elon Musk mitfinanzierten Thinktanks Future of Life Institute (FLI, Boston) und Future of Humanity Institute (FHI, Oxford).

Begriffe wie der *longtermism* oder eben das «existenzielle Risiko» kommen nicht einfach aus dem Nichts. Sie sind vielmehr mit Akteuren wie etwa dem FHI-Direktor Nick Bostrom oder William MacAskill, einem analytischen Philosophen in Oxford, verknüpft. Beide forschen schon seit Jahren im Modus moralphilosophischer Denkexperimente zu hypothetischen «KI-Übernahmeszenarien», sind im Silicon Valley aktuell sehr en vogue und liefern nicht nur Musk schon lange die apokalyptischen Argumente für seine «Philosophie».

So markiert etwa der von Bostrom und MacAskill propagierte *longtermism* eine Denkbewegung, die sich, um mit den Worten Émile Torres, einem der profiliertesten Kritiker, zu sprechen, aus einer utilitaristischen Ethik und beständigen Risikokalkulation ableitet. Diese Moralphilosophie, die sich an scheinbar quantifizierbaren Bewertungsmassstäben orientiert, nimmt dabei eine radikale Blickverschiebung vor: Sie richtet den Fokus weniger auf die Lösung gegenwärtiger Probleme als auf – daher der Name – die lange Frist. Ganz besonders: auf das langfristige Überleben der menschlichen Spezies.

Der zentrale Begriff ist das «existenzielle Risiko», das nach Bostrom ausdrücklich nur jene Bedrohungen meint, die das – egal wie unwahrscheinliche – Potenzial haben, die gesamte Menschheit auszulöschen. Seien es künstliche, letale Viren oder eine Nanofabriken steuernde KI, die moskitoähnliche Roboter auf die Menschen hetzt. Solche Szenarien werden in der «ethischen» Kalkulation mit Wahrscheinlichkeiten belegt, die mögliche Katastrophen einer imaginierten Zukunft auf vermeintlich präzise Zahlen bringen: Sci-Fi-Doomscrolling als Methode, Moralphilosophie als angewandte Dystopie.

Dass in diesen Berechnungen zuweilen auch die real existierende Klimakatastrophe mit all ihren Folgekatastrophen – steigende Meeresspiegel, Dürren, Hunger oder Flucht – als weniger bedrohlich eingestuft wird als die Gefahren einer Superintelligenz, ist dann mehr als nur eine Diskursblüte. Denn mit den infinitesimal winzigen (Un-)Wahrscheinlichkeiten von Katastrophen durch Superintelligenz gehen Handlungsempfehlungen einher, die Bostrom oder MacAskill nicht nur zu Befürwortern einer «sicheren KI» machen. Sie lassen sie auch darüber nachdenken, ob es nicht fast schon eine moralische Pflicht ist – wie etwa Bostroms Paper «Astronomical Waste» oder MacAskills «The case for strong longtermism» argumentiert –, anstatt irdische Fatalitäten zu bekämpfen, an der Kolonisierung fremder Sterne und Planeten zu arbeiten – eine Überzeugung, die der angehende Marsmensch Elon Musk nicht nur auf X teilt.

Angesichts solcher denkexperimentellen Volten stellt sich die Frage, zu welchem Grad Science-Fiction in die analytische Moralphilosophie eingewandert ist, ob es zwischen den beiden Sparten überhaupt noch einen Unterschied gibt. Wichtiger ist es allerdings, das Kalkül hinter den Diskursen über die lange Frist zu dechiffrieren. Man wird das Gefühl nicht los, dass die Fokussierung auf spekulative, entfernte Zukünfte vor allem einen Zweck erfüllt: von den vielschichtigen, schwierig zu lösenden Gegenwartsproblemen abzulenken.

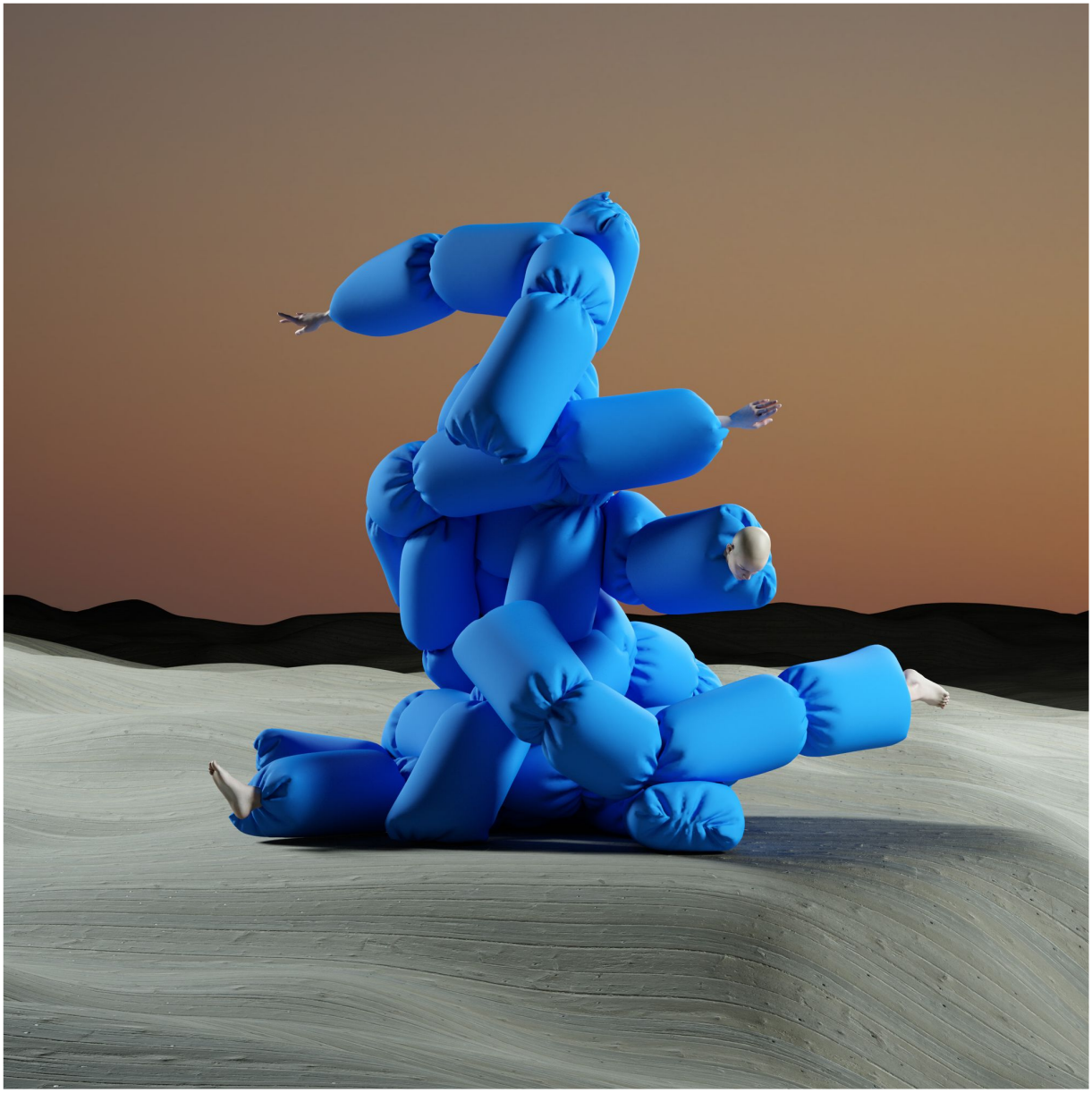
Lieber in die longtermistische Ferne schweifen und mit endzeitlichem Schauer in Berechnungen über Tod und Vernichtung schwelgen als sich mit den realen, kleinteiligen und spröden Herausforderungen der Gegenwart herumärgern. Die Mitgründerin des AI Now Institute, Meredith Whittaker, bringt das Problem auf den Punkt: *«A fantastical, adrenalizing ghost story is being used to hijack attention around what is the problem that regulation needs to solve.»* (Mit einer imaginären, aufregenden Gespenstergeschichte wird

versucht, die Aufmerksamkeit von dem Problem wegzulenken, das durch Regulierung gelöst werden muss.)

Intelligenz oder Dummheit?

Betrachtet man die Technologien jenseits der apokalyptischen Horizonte einmal etwas realistischer, wird schnell klar, dass viele KI-Systeme weniger mit übermenschlichen Potenzialen als mit «künstlicher Dummheit» durchsetzt sind: dass das Doomsday-Marketing ihre Fähigkeiten überzeichnet. Die aktuellen generativen KI wie Chat GPT sind keine omnipotenten Systeme, sondern von maschinellen, allzu maschinellen Falschheiten und sogenannten Halluzinationen (Behauptungen ohne Quelle) durchzogen. Deshalb generieren sie zwar smart klingende, aber nicht selten ziemlich defizitäre und desinformierte Texte.

Man kann auch von technosophistischen Systemen sprechen, die bestenfalls Schattenrisse auf die erkenntnistheoretische Höhlenwand projizieren und – wie die FTC befürchtet und selbst Altman einräumt – vor allem die Probleme von Desinformation in den Weiten des Internets eskalieren lassen. Mit Blick auf die Verschwörungserzählungen der Gegenwart sind maschinengeschriebene Versionen jedenfalls kein wünschenswertes Szenario.



Borja Alegre

Zu den Bildern

Der 3-D-Designer Borja Alegre hinterfragt und erforscht mit seinen Bildern die Grenzen der Realität: Was kann existieren, was nicht? Sein surrealer Stil mutet traumhaft und emotional an, schwankt zwischen Dystopie und Utopie. Die artifiziellen Bilder und Welten, die er schafft, wecken Faszination, Unbehagen und werfen Fragen auf, die ohne Antwort bleiben – und lösen damit ähnliche Reaktionen aus wie künstliche Intelligenz.

Neben diesen technikinhärenten Missliebigkeiten stellen die KI-Verfahren vor allem sozioökonomische Herausforderungen dar. Häufig nehmen sie uns nicht nur mühsame Arbeit ab, sondern schreiben dabei Ungleichheiten und Ungerechtigkeiten fort oder verschärfen diese gar. So sind die Chatbots Chat GPT oder Bard gespeist mit Unmengen an Trainingsdaten, die Wissen weniger sammeln als zusammenklauen und reihenweise Urheberrechte verletzen. Oder, wie es kürzlich in einer Klageschrift in den USA hiess: *«It has very recently come to light that Google has been secretly stealing everything ever created and shared on the internet by hundreds of millions of*

Americans.» (Kürzlich kam ans Licht, das Google heimlich alles gestohlen hat, was jemals von Amerikanerinnen erschaffen und im Internet geteilt worden ist.)

Auch jenseits der Trainingsdaten wird nicht unbedingt fair gearbeitet: Um Chat GPT weniger falsche, rassistische oder sexistische Aussagen treffen zu lassen, hatte Open AI etwa kenianische Klickarbeiterinnen angestellt, die Inhalte «moderieren» und herausfiltern sollten – für weniger als zwei Dollar pro Stunde.

Das Beispiel verweist neben den Ausbeutungsverhältnissen einer nur vermeintlich künstlichen Intelligenz auf ein weiteres Problem: die Tatsache, dass KI mit der Reformulierung menschlicher Texte auch rassistische oder misogyne Biases (Verzerrungen) und Vorurteile reproduziert. Sie dichtet zum Beispiel Lyrics wie: *«If you see a woman in a lab coat, she's probably just there to clean the floor / But if you see a man in a lab coat, then he's probably got the knowledge and skills you're looking for.»* (Wenn du eine Frau im Laborkittel siehst, dann ist sie wahrscheinlich nur da, um den Boden zu putzen / Wenn du aber einen Mann im Laborkittel siehst, dann hat er wahrscheinlich das Wissen und die Fähigkeiten, die du suchst.)

Auch für dieses Problem hat der Open-AI-CEO Altman noch keine superintelligente Antwort gefunden. Einer der letzten Lösungsvorschläge: Was den Moderatoren entgeht, sollen die User selbst (ganz kostenlos) übernehmen und diskriminierende Texte markieren, um die Technik zu verbessern. Toxische Texte sollen durch Gratisarbeit verbessert werden, eine für das Silicon Valley bezeichnende Idee: die billigste.

Trotz der proklamiert *künstlichen* Intelligenz sind die Systeme stets auf Menschen angewiesen, auf korrigierende Eingriffe, auf Arbeit hinter den glatten Oberflächen. Die «smarten» Prozesse sind nur vermeintlich schwerelos und basieren zudem auf materiellen Strukturen – von den Data-Centern bis zu Unterseekabeln –, die Unmengen Ressourcen und Energie verbrauchen. Damit gehören auch KI und Klimakatastrophe strukturell zusammen. Forscherinnen um Emma Strubell, Assistenzprofessorin an der Carnegie Mellon University, haben auf diesen Umstand in einem viel diskutierten Paper verwiesen und erklärt, dass nur schon das Training eines BERT-Modells (eine Variante einfacher Sprachmodelle) so viel Energie verbraucht wie ein Flug von New York nach San Francisco. Das Training des grossen Sprachmodells GPT-3 (175 Milliarden Parameter), auf dem auch Chat GPT basiert, emittiert dann schon so viel CO_2 wie 126 Haushalte im Jahr – dies alles abgesehen von den alltäglichen, energetischen Betriebskosten.

Wir müssen uns also nicht eine ferne Zukunft imaginieren, um das zerstörerische Potenzial der Systeme zu erkennen. Es reicht, auf ihre Ökobilanz zu schauen.

Hier wird auch ein weiteres (infra-)strukturelles Problem offensichtlich. Der energetische Verbrauch ist mit riesigen Serverfarmen verbunden, die sich jenseits der etablierten Big-Tech-Konzerne kaum jemand leisten kann. Bei Chat GPT stellt Microsoft die Rechenpower, bei Bard ist es Google. Auch in der KI-Entwicklung scheint sich zu manifestieren, was wir aus dem «normalen» kommerziellen Internet kennen: dass wenige Konzerne beziehungsweise «parastaatliche Unternehmen» (Literatur- und Kulturwissenschaftler Joseph Vogl in seinem Buch: «Kapital und Ressentiment») jenseits jeder demokratischen Legitimation eine nahezu konkurrenzlose infrastrukturelle Macht aufbauen, die der Gesellschaft die (Community-)Standards diktiert.

Wer dann ferner noch auf einzelne Milliardäre wie Elon Musk schaut, seine antisemitischen Aussagen, Beleidigungen und ideologischen Irrläufe auf X verfolgt – zuletzt wollte er gar die Anti-Defamation League verklagen –, wird erkennen müssen, dass in der Macht- und Diskurskonzentration vielleicht kein existenzielles, aber ein gesellschaftliches Risiko steckt, das nicht zu unterschätzen ist.

Alternativen für das Hier und Jetzt

Mit Blick auf die breite mediale, aber auch von der Präsidentin der EU-Kommission befeuerte Reproduktion des Doomsday-Marketings und eine KI-Industrie, die sich selbst schützt, indem sie die von ihr ausgehenden Risiken aufpumpt und sich dagegen «absichert» (Stichwort: *safe AI*), wäre eine Politisierung des Technoimaginären jenseits der Visionen einer weiss-männlichen und longtermistischen Elite dringend notwendig. Mit dem heutigen KI-Diskurs werden einmal mehr kulturtheoretische Binsenweisheiten anschaulich: Wir können uns eher das Ende der Welt als das Ende des (digitalen) Kapitalismus vorstellen – oder, mit den Worten des Philosophen Walter Benjamin gesprochen: «Dass es ‹so weiter› geht, ist die Katastrophe.»

Weil dies so ist, bedürfen wir eines Gegenbegriffs; eines Begriffs, der aufzeigt, dass mit Leuten wie Bostrom oder Musk eine Grenze der tech-utopischen Vorstellungskraft erreicht ist; dass die Vielfalten und Möglichkeiten des Internets von einzelnen Konzernen mit Doomsday-Visionen überformt und eingeschränkt werden und dass viele technische Systeme schon heute vernichtende Konsequenzen zeitigen.

Genau diese Bedrohungen bringt «Extinction Internet» auf den Begriff; ein Kofferkonzept des niederländischen Medientheoretikers Geert Lovink, das vor dem Hintergrund zunichtegemachter Vorstellungshorizonte entscheidende Fragen aufwirft. Was tun, wenn die Systeme nicht mehr als smarte Tools der Weltverbesserung, sondern der Weltauslöschung annonciert werden? Wenn die «Apokalypse-Blindheit» (Technikphilosoph Günther Anders in «Die Antiquiertheit des Menschen») von einst durch die falsche Apokalypsesättigung von heute ersetzt wird – und uns, dialektisch gewendet, vielleicht noch blinder macht für Alternativen?

Vielleicht sollten wir tatsächlich ein Ende denken lernen; natürlich nicht derart, wie von den Phrasendreschern aus dem Valley vorformuliert. Stattdessen müsste es um ein Ende der Ausbeutung von Mensch und Umwelt gehen, um ein Ende der (daten-)extraktiven KI-Systeme, wie wir sie kennen. Etwas anderes müsste forciert werden: ein Ende, bei dem wir uns in Gegenprogrammen, einem Plattform-Exit, einem *machine unlearning* üben; anfangen, uns von Big Tech und der nur vermeintlich «offenen» AI zu verabschieden. Jenseits der technischen Monokulturen des Silicon Valley gibt es Netzalternativen – Duck Duck Go, Signal, Mastodon usw. –, die dezentraler, kooperativer, demokratischer, dekolonialer und offener sind; es gibt andere Protokolle, Apps und Plattformen. Wir müssen sie nur suchen, nutzen oder auch kreieren. Wie also könnte eine andere KI aussehen? Wie ein anderer KI-Diskurs?

Angesichts des KI-Hypes und seiner kalkulierten Untergangsfantasien muss es uns um radikale Nüchternheit und gestalterischen Widerstand gehen; um eine Praxis umwertender Daten- und Machtkritik, die KI-Narrative und ihre (kommerziellen) Codes umschreibt, die Infrastrukturen (Datenzentren etc.) dezentralisiert und demokratisiert, KI-Modelle und Datensätze transparent macht, sie als öffentliches Gut versteht. Um eine politische

Praxis, die sich nicht auf zukünftige Regulierungen («AI Act» der EU etc.) verlässt, sondern an Alternativen zum Status quo arbeitet – an Care-Techniken für eine beschädigte Welt im Hier und Jetzt.

Vielleicht ist die Zeit gekommen, die proklamierte Lust am Doomsday ernst zu nehmen, ernster noch als die Tech-Elite selbst das tut: Denn wer so oft über das Ende spricht, es verkauft und bewirbt, dürfte wirklich an *ein Ende* gelangt sein.

Zu den Autorinnen

Felix Maschewski ist Medien-, Kultur- und Wirtschaftswissenschaftler; Ko-Direktor des Critical Data Lab an der Humboldt-Universität (HU) zu Berlin, assoziierter Forscher am Institute of Network Cultures (Amsterdam) und lehrte zuletzt am Institut für Soziologie der Universität Basel.

Anna-Verena Nosthoff ist Philosophin, politische Theoretikerin, Publizistin und Ko-Direktorin des Critical Data Lab an der HU Berlin. Gemeinsam mit Felix Maschewski veröffentlichte sie unter anderem das Buch «Die Gesellschaft der Wearables. Digitale Verführung und soziale Kontrolle».