

BUSINESSMODELL WELTUNTERGANG

Warum ausgerechnet die größten KI-Konzerne nach Regulierung rufen und welche tatsächlichen Probleme dahinter stecken.

[Matthias Müller]

KI ist sprichwörtlich in aller Munde. Wer dieses Thema noch nicht wenigstens aus den Nachrichten kennt, der muss sich wahrscheinlich zu Recht vorwerfen lassen, überhaupt gar nichts mehr von der Realität mit zu bekommen – so dauerpräsent ist die „Künstliche Intelligenz“ auf allen Kanälen. Aber nicht nur hier. In Deutschland haben mittlerweile rund 58 bis 65 Prozent der Menschen ab 16 Jahren schon einmal ein Produkt mit generativer Künstlicher Intelligenz wie ChatGPT, Google Gemini oder DeepL ausprobiert oder nutzen es regelmäßig. Unter den 16- bis 29-Jährigen liegt die Nutzungsrate extrem hoch – hier haben über 90 Prozent bereits Erfahrung mit KI-Tools gesammelt. Damit ist KI die technologische Revolution, die innerhalb kürzester Zeit eine Marktdurchdringung erreicht hat, wie keine andere Innovation in der Geschichte der Menschheit zuvor – denn, was wir aufgrund der Geschwindigkeit der Ereignisse komplett vergessen: „KI“ als öffentlich verfügbaren Dienst gibt es erst seit November 2022. Da wurde der Dienst „ChatGPT“ vom US-amerikanischen Konzern „Open AI“ erstmals zugänglich gemacht.

Von Beginn an sah sich diese Technologie extrem kontroversen Diskussionen ausgesetzt. Der so genannte „Doomer-Hype“ – den wir später noch detailliert sezie-

ren werden – zeichnete äußerst medienwirksam die düstersten Visionen von einer nahen Zukunft, in der die „AGI“ (Artificial General Intelligence = Allgemeine Künstliche Intelligenz, also eine hypothetische Superintelligenz, die sich selbst verbessern und vervielfältigen kann) die Kontrolle über den gesamten Planeten einschließlich der Menschheit übernimmt und uns, die Menschen, aufgrund eiskalter mathematisch-logischer Schlussfolgerungen wahlweise versklavt oder tötet.

Und gerade in den letzten Monaten melden sich ausgerechnet die CEOs praktisch sämtlicher großer KI-Konzerne, wie Dario Amodei (Anthropic, KI-Modell: Claude), Sam Altman (Open AI, KI-Modell: ChatGPT), Elon Musk (KI-Modell: Grok) und Mustafa Suleyman (Microsoft AI, KI-Produkt: Copilot) zu Wort und fordern Kontrolle und Regulation. Dies lässt natürlich aufhorchen – denn noch selten hat man es aus den Kreisen von Multi-Milliarden-Konzernen gehört, dass die Vorstandsvorsitzenden selbst nach Überwachung, staatlichen Zulassungshürden und Drosselung des Entwicklungstempos rufen. Was ist da los?

Tatsächlich bewegen sich die Rufe aus den Chefetagen nach Verlangsamung und Regulation in einem hochgradig kalkulierten Spannungsfeld. Hinter den

öffentlichen Warnungen stehen drei wesentliche Triebfedern, und keine davon hat mit echter Sorge um die Zukunft der Menschheit zu tun.

APOKALYPTIK ALS MARKETINGINSTRUMENT

Im Wesentlichen lassen sich die medial viel beachteten Aussagen der genannten Persönlichkeiten in drei großen Punkten zusammenfassen:

Regulatory Capture & Burggräben.

Wer nach strengen staatlichen Zulassungen für das Training großer Modelle verlangt, schafft Markteintrittsbarrieren. Start-ups und vor allem die offene Open-Source-Community können millionenschwere Sicherheitszertifizierungen und bürokratische Audits nicht stemmen. Das Ergebnis ist kein Stopp der Technologie, sondern ein reguliertes Oligopol etablierter Player. Tatsächlich ist das größte Risiko für das Geschäftsmodell der KI-Giganten nicht der befürchtete, unkontrollierte (und wiederum medial sehr auffällige) Ausbruch der eigenen KI-Produkte aus dem konzern-eigenen Ökosystem, sondern frei zugängliche, kostenlose KI-Modelle, die nach dem Open-Source-Modell (wie einst der Vorreiter Linux) von einer weltweiten Entwickler-Community zur Verfügung gestellt werden. Die KI-Konzerne gehören heute, nur kaum vier Jah-

re nach dem erstmaligen öffentlichen Ausrollen ihrer Produkte, zu den wertvollsten Unternehmen der Welt. Es besteht seitens der CEOs ein starkes Interesse, diese Marktmacht und Finanzkraft gegen kostenlose Konkurrenzprodukte abzusichern.

Das Marketing-Paradoxon „Doomer-Hype“. Wer behauptet, sein Produkt sei derart mächtig, dass es die Menschheit auslöschen könnte, betreibt die wirksamste Form von Produktwerbung. Es verleiht der Technologie eine Aura übermenschlicher Allmacht, lenkt von profanen Schwachstellen wie Halluzinationen ab und rechtfertigt astronomische Bewertungen gegenüber Investoren. Die zentrale Marketingbotschaft lautet: „Unsere KI ist derart mächtig und leistungsfähig, dass wir sogar selbst Angst vor ihr haben“ – dieser Marketingtrick ist nicht nur unfassbar sexy, sondern vor allem auch unfassbar wirkungsvoll. Kaum jemand kann sich der morbiden Faszination entziehen, die von dieser Erzählung ausgeht.

Gleichzeitiges und ungebremstes Wettrüsten. Trotz theatralischer offener Briefe und Warnungen pausiert kein einziger der Vorreiter faktisch die eigene Forschung. Investitionen in Rechenzentren, Custom-Chips und gigantische Cluster laufen auf Rekordniveau weiter – das klassische Gefange-

nendilemma: Wer als Erster anhält, verliert den Markt. Auf der realen Ebene erweist sich die Forderung nach Tempodrosselung als leeres Geschwätz: das Gaspedal ist bis aufs Bodenblech durchgedrückt, bei jedem Einzelnen der angeblich so besorgten Marktteilnehmer.

Gleichwohl existiert unter vielen Spitzenforschern eine genuine Furcht. Das Alignment-Problem – also die mathematische und konzeptionelle Garantie, dass ein autonom agierendes Modell dauerhaft intendierte Werte verfolgt – ist fundamental ungelöst. Es lässt sich innerhalb der Modellarchitektur auch nicht lösen, denn ethische Werte und mathematischer Determinismus sind technisch nicht miteinander vereinbar. Für Nicht-KI-Experten (die meisten von uns) hier nochmal das „Alignment-Problem“ im Klartext: die Funktionsebene jedweder KI ist die Mathematik. Mathematisch lassen sich ethische Werte aber nicht formulieren. Daher gibt es technisch-konzeptionell derzeit keinen zuverlässigen Weg, in einem KI-Modell Werte zu verankern, die es zuverlässig zu einer „guten“ KI im Sinne eines menschlichen Wertesystems machen würden.

EXKURS: DER DOOMER-HYPE

Genau dieses Problem ist die Uruppe des so genannten „Doomer-Hype“. Da diese Strömung ein wichtiges Element zum Verständnis der gesamten Branche darstellt, wollen wir hier einen kleinen Exkurs in seine Entstehungsgeschichte unternehmen. Der Begriff des „KI Doomerismus“ wirkt zwar wie ein Phänomen des Jahres 2023, seine intellektuellen, ideologischen und strategischen Wurzeln reichen jedoch über zwei Jahrzehnte zurück. Was als esoterische Subkultur im Silicon Valley begann, wurde von einer philosophischen Denkfigur zu einer der

wirkmächtigsten PR- und Lobbying-Waffen der Tech-Geschichte umfunktioniert. Der Ursprung des Doomerismus liegt in einer bizarren Mischung aus Transhumanismus, Bayesscher Statistik und Science-Fiction-Theorie.

Eliezer Yudkowsky, ein autodidaktischer Denker ohne akademischen Abschluss, gründete ab etwa dem Jahr 2000 das Machine Intelligence Research Institute (MIRI) und später die Diskussionsplattform LessWrong. Seine These: Sobald eine KI fähig ist, sich selbst rekursiv zu verbessern, entsteht binnen Stunden eine Superintelligenz. Da menschliche Werte extrem komplex und mathematisch schwer definierbar sind, führe jeder minimale Zielkonflikt unausweichlich zur Auslöschung der Menschheit. Der Oxford-Philosoph und umstrittene Zukunftsforscher Nick Bostrom akademisierte diese Ideen. Sein Buch „Superintelligence: Paths, Dangers, Strategies“ (2014) wurde zur Bibel des Phänomens. Seine Kernaussagen prägten das gesamte Vokabular der Community. Er stellte die Orthogonalitätsthese auf, die im Wesentlichen folgendes besagt: Ein System kann beliebig intelligent sein und gleichzeitig völlig absurde oder zynische Ziele verfolgen. Gleichzeitig postulierte Bostrom das Axiom der Instrumentellen Konvergenz: Fast jedes Ziel (selbst das harmlose Herstellen von Büroklammern) erfordert logisch dieselben Zwischenschritte: Selbsterhaltung, Ressourcenaneignung und das Ausschalten potenzieller Störfaktoren (sprich: des Menschen, der den Stecker ziehen könnte). Bostrom war damit der erste, der den trivialen Begriff der potenziellen „Bösartigkeit“ einer KI als logische Folge deterministischer mathematischer Gesetze normalisierte: KI ist nicht potenziell böse (im Sinne menschlicher Maß-

stäbe), sondern handelt lediglich konsequent logisch.

In den 2010er-Jahren dockte die Doomer-Bewegung an den „Effective Altruism“ (EA) an – eine utilitaristische Bewegung, die von wohlhabenden Krypto- und Tech-Milliardären (wie Sam Bankman-Fried, aber auch Skype-Gründer Jaan Tallinn und Dustin Moskovitz) mit hunderten Millionen Dollar geflutet wurde. Die Denker postulierten, dass das Überleben von Billionen künftiger, digitaler Menschen im Weltall moralisch wichtiger sei, als das Lösen gegenwärtiger Armut oder des Klimawandels. Die Verhinderung eines „X-risk“ (Lebensbedrohliches Risiko durch KI) wurde zum heiligsten aller moralischen Güter erklärt. Fast die gesamte Gründergeneration moderner KI-Konzerne entstammt diesem Dunstkreis. DeepMind (Demis Hassabis, Shane Legg) und OpenAI (Sam Altman, Ilya Sutskever, Elon Musk) wurden ursprünglich unter dem expliziten Vorwand gegründet, „die Welt vor der ersten unkontrollierten AGI zu schützen“, indem man ihr durch eigene Entwicklung zuvorkommt. Anthropic spaltete sich 2021 von OpenAI ab – getrieben von genau jenen „Safety-Puristen“, die OpenAI für zu kommerziell hielten.

Mit dem Durchbruch von ChatGPT verwandelte sich die interne Firmenphilosophie schlagartig in eine orchestrierte globale Kampagne. Zwei Dokumente markierten den Höhepunkt:

• Der FLI-Brief (März 2023):

Das vom Future of Life Institute initiierte Schreiben forderte einen sofortigen sechsmonatigen Trainingsstopp für Systeme, die mächtiger als GPT-4 sind. Unterzeichner: Elon Musk, Steve Wozniak und Spitzenforscher. Ironie: Musk kaufte exakt in diesem Zeitraum zehntausende Nvidia-GPUs für sein eigenes KI-Unternehmen xAI.

• Die CAIS-Erklärung (Mai 2023):

Das Center for AI Safety brachte eine berühmt gewordene Einsatz-Erklärung heraus:

„Die Minderung des Risikos einer Auslöschung durch KI sollte eine globale Priorität neben anderen gesellschaftlichen Risiken wie Pandemien und Atomkriegen sein.“ – Unterzeichnet von Sam Altman (OpenAI), Demis Hassabis (DeepMind), Dario Amodei (Anthropic) und Turing-Preisträgern wie Geoffrey Hinton.

DIE BOTSCHAFTEN DER DOOMER UND IHRE INSTRUMENTALISIERUNG

Wer den Diskurs analysiert, erkennt vier ideologische Leitplanken, die jeweils handfesten Interessengruppen dienen:

„KI ist mächtiger als die Atombombe und kann zur Auslöschung der Menschheit führen.“

Mit dieser Aussage wird das Produkt als weltverändernde, Gott-ähnliche Macht inszeniert, deren Entwicklung man unter allen Umständen anführen muss, ansonsten tun es „die Anderen“. Investoren werden von diesem nie dagewesenen Potenzial angezogen wie Fliegen vom Misthaufen. Mängel in der Software werden verziehen, Schnelligkeit ist das einzige relevante Unternehmensziel. Durch die Inszenierung als künftige All-Technologie verfügen die Führungsetagen der KI-Konzerne über faktisch grenzenlose Finanzmittel und die Politik bemüht sich eifrig, als Ermöglicher und Hindernis-Beseitiger zu agieren.

„Forschung darf nur in wenigen, zertifizierten Mega-Rechenzentren statt finden.“

Mit dieser Aussage geht es vor allem darum, sich die unliebsame Open-Source-Konkurrenz vom Leibe zu halten. Durch bürokratische Lizenzierungs-Monopole bleibt die Entwicklungshoheit bei den Tech-Konzernen und die Hyper-

Scaler können ungestört weiter ihre Cash-Cows melken.

„Es ist eine Frage des nationalen Überlebens, in dieser Technologie führend zu sein“. Mit dieser Aussage wird jedweder öffentlicher Widerstand gegen Milliardensubventionen, Abbau von rechtlichen Hindernissen wie Datenschutz und Urheberrechten gebrochen und der Schulterschluss zur Rüstungsindustrie sowie den Verteidigungsministerien hergestellt.

„Gegenwärtige Schäden sind irrelevant gegenüber dem Untergang der Menschheit“. Mit diesem Narrativ werden alle negativen Begleiterscheinungen pauschal entwertet. Klagen gegen Urheberrechtsverletzungen, Umweltzerstörung und Arbeitsplatzvernichtung werden als kleinkarierte Befindlichkeiten und belanglose Nebenkriegsschauplätze geframed. Es geht schließlich um nichts weniger als die Verhinderung der totalen Vernichtung. Die KI-Industrie hat es damit geschafft, selbst für den aggressiven Ökofaschismus komplett unantastbar zu werden. KI sticht Klima – diese Diskussion war bereits beendet, noch bevor sie überhaupt aufkam.

Dass der Doomer-Hype so durchschlagend funktionierte, liegt daran, dass er zwei Pole perfekt bediente: Den echten Glauben einer philosophisch sektenartigen Community an die drohende Apokalypse – und die kühle Berechnung von CEOs und Risikokapitalgebern, die erkannten, dass man diese Endzeiterzählung perfekt in Marktanteile, legislative Barrieren und Marktbewertungen übersetzen kann.

REGULATORY CAPTURE UND DER KAMPF GEGEN OPEN SOURCE MODELLE

Der scheinbar hehre Ruf der führenden Tech-Konzerne nach staatlichen Regulierungsleit-

planken folgt einem klassischen industrieökonomischen Lehrstück: der Kaperung der Gesetzgebung durch genau jene Akteure, die sie vorgeblich einhegen soll. Während proprietäre Giganten ihre Modelle hinter den Mauern streng zensurierter Web-Oberflächen und kontrollierter Programmierschnittstellen verwahren, legen Open-Source-Initiativen ihre mathematischen Gewichte offen. Damit kann jeder diese Systeme lokal auf eigener Hardware betreiben, analysieren und etwaige Sicherheitsfilter durch gezieltes Feintuning entfernen. An dieser architektonischen Sollbruchstelle setzt eine bemerkenswert perfide Kampagne an, die Sicherheitsbedenken instrumentalisiert, um die unliebsamen, frei zugänglichen Konkurrenten im Keim zu ersticken.

Der wirksamste Hebel dieser Strategie besteht darin, eine verschuldensunabhängige Haftung bei den ursprünglichen Entwicklern zu verankern. Verlangt der Gesetzgeber, dass der Urheber eines Modells für jeglichen Schaden geradestehen muss, den Dritte damit anrichten – sei es durch Malware oder Desinformationskampagnen – bedeutet dies das unausweichliche Todesurteil für freie Software. Während ein Cloud-Anbieter verdächtige Anfragen serverseitig sperren und Nutzer protokollieren kann, verliert der Open-Source-Entwickler mit der Freigabe der Gewichte jede technische Kontrolle über die weitere Nutzung. Wer ihn dennoch für den nachgelagerten Missbrauch verantwortlich macht, verwandelt das Publizieren offener Quellcodes in ein unkalkulierbares juristisches Himmelfahrtskommando, das keine universitäre Forschungsgruppe und kein unabhängiges Start-up überleben könnte.

Flankiert wird dieses juristische Minenfeld durch eine bürokrati-

sche Arbitrage, die Sicherheitsstandards primär über exorbitant teure Prozesse statt über funktionale Ergebnisse definiert. Wenn Gesetze verpflichtende, kontinuierliche Red-Teaming-Prüfungen, forensische Datensatz-Audits und langwierige Zertifizierungsverfahren vor jedem Versionssprung vorschreiben, errichtet dies eine gigantische Markteintrittsbarriere. Big Tech finanziert derartige Hürden mühelos aus der Portokasse ihrer Rechts- und Compliance-Abteilungen, während die dezentrale Entwicklergemeinschaft, die schließlich ein kostenloses Produkt entwickelt, schlicht von den Kosten erdröselt wird.

Um dieses System lückenlos abzusichern, greift die Lobbyarbeit auf willkürlich gesetzte Rechenleistungsschwellen zurück. Indem schiere Rechenpower als inhärentes Gefahrenpotenzial geframed wird, unterwirft man Modelle ab gewissen Trainingskapazitäten automatisch strengen staatlichen Geheimhaltungs- und Freigabeverfahren. Weil algorithmische Optimierungen und effizientere Hardware dazu führen, dass auch kleinere Akteure diese Schwellen in rasantem Tempo erreichen, kriminalisiert ein solches Modell den freien Release genau jener offenen Alternativen, die technologisch zu den Flaggschiffen der Konzerne aufschließen könnten. Ergänzt wird diese Praxis auf politischer Ebene durch das gezielte Narrativ von Open-Weight-KI als digitaler Waffe, die Terroristen in die Hände falle und daher Exportkontrollen unterliegen müsse – eine Rhetorik, die das seit Jahrzehnten bewährte, transparente Prinzip von Linux oder Apache über Nacht als Sicherheitsrisiko diffamiert.

Hinter diesem vorgeblichen Schutzwall verbirgt sich ein knallhartes ökonomisches Kalkül. Wer den Markt auf rein zugangsbe-

schränkte Schnittstellen verengt, schützt seine astronomischen Monopoleinnahmen vor dem Preisverfall durch kostenlose Alternativen und zwingt die Weltwirtschaft in ein dauerhaftes Token-Abonnement. Gleichzeitig wird jede echte technologische Souveränität torpediert, denn nur offene Modelle garantieren lokalen Datenschutz, die Unabhängigkeit von US-Hyperscalern und die transparente Überprüfbarkeit systemischer Verzerrungen. Die bitterste Ironie dieses Manövers liegt darin, dass viele wissenschaftliche Sicherheitserkenntnisse überhaupt erst durch das freie Sezieren offener Modelle möglich wurden. Der lautstarke Ruf nach Schutz dient den etablierten Plattformen am Ende zu nichts anderem, als sich unter dem Beifall überforderter Parlamente einen unüberwindbaren Burggraben zu schaufeln: „KI ist so gefährlich, sie muss kontrolliert werden – und zwar von uns, den Megakonzerne. Unsere Motive sind edel, hilfreich und gut. Wir wollen nur Euer Bestes.“ Aha. Ja, ne, is klar.

Mit diesem Hintergrund sticht das Muster natürlich sofort ins Auge: Was hier abläuft, ist im Grunde das klassische PR-Krisen- und Framing-Playbook, nur ins Extrem skaliert: Strategischer Alarmsimus.

Aus kommunikations- und marktstrategischer Sicht bedient diese Kampagne mehrere geniale Hebel: Das moralische Alibi – man inszeniert sich nicht als skrupelloser Monopolist, der die Welt mit unausgereifter Software flutet, sondern als tragischer Prometheus. Die CEOs treten vor Kongresse und Parlamente, geben sich staatsmännisch, demütig und besorgt – und diktieren den Gesetzgebern, die technologisch heillos überfordert sind, die Gesetzesentwürfe gleich direkt in die Feder.

Gleichzeitig ist die Strategie eine

geschickte Ablenkung von den profanen Sünden der Gegenwart: Solange Medien und Politik über hypothetische Killer-Roboter, das Ende der Menschheit und AGI debattieren, redet niemand über reale, ungemütliche Themen wie unbezahlten Datendiebstahl für Trainingssets im gigantischen Maßstab, den absurden Wasser- und Energieverbrauch von Serverfarmen, systematische Copyright-Verletzungen oder das Ausbeuten von Klickarbeitern im globalen Süden für die Datenbereinigung. Die Meisterleistung der Kampagne besteht darin, die eigenen existenziellen Geschäftsrisiken (Commoditisierung durch kostenlose Open-Source-Modelle) als existenzielle Risiken für die Menschheit umzudeuten. Wer das Narrativ kontrolliert, kontrolliert die Gesetze – und wer die Gesetze kontrolliert, baut den Burggraben, ohne einen einzigen Cent in bessere Technologie investieren zu müssen.

DIE REALEN GEFAHREN DER KI

Tatsächlich liegt das wahre und reale Gefahrenpotenzial aber ganz woanders, als von den Techkonzernen postuliert wird. Die forensischen Berichte von Interpol, Europol und Datenbanken wie der AI Incident Database zeichnen ein bemerkenswert nüchternes Bild, das in krassem Widerspruch zu den apokalyptischen Schreckensszenarien der PR-Abteilungen steht. Kriminelle Akteure nutzen künstliche Intelligenz keineswegs, um im finsternen Darknet an synthetischen Superviren zu forschen oder Stromnetze zum Einsturz zu bringen. Sie agieren nach streng ökonomischer Rationalität und setzen die Technologie dort ein, wo sie Aufwand minimiert und Erträge maximiert, namentlich im automatisierten Social Engineering, bei nicht-konsensuellen pornografischen Deepfakes sowie

in der industriellen Skalierung von Betrugsmaschinen. Das Arsenal des Missbrauchs speist sich dabei aus zwei grundsätzlich verschiedenen Modellwelten:

Open-Source Kriminalität:

Auf der Seite frei verfügbarer Open-Weight-Systeme greift die Schattenwirtschaft auf gezieltes Feintuning oder das sogenannte Abliteration-Verfahren zurück, bei dem Sicherheitsvektoren direkt aus den Gewichten herausgeschnitten werden. Metas LLaMA-Architektur fungiert dabei als das eigentliche Arbeitspferd, mit dem Spear-Phishing-Mails in fehlerfreiem Duktus generiert, Erpresserschreiben verfasst und Skripte für groß angelegte Romance Scams automatisiert werden. Noch vor solchen Anpassungen sorgten frühe Derivate wie das auf dem Modell GPT-J basierende WormGPT für Aufsehen, das als schlüsselfertiges Werkzeug für Angriffe auf Führungskräfte vertrieben wurde. Das wohl verheerendste Schadenspotenzial entfaltet indes die Bildgenerierung mit frei laufenden Checkpoints wie jenen von Stable Diffusion. Weil die Software lokal und ohne serverbasierte Zensurfilter betrieben werden kann, bildet sie das unbestrittene Rückgrat eines florierenden Schwarzmarkts für gefälschte Ausweisdokumente, erpresserische Deepfakes und bildbasierte sexuelle Gewalt gegen Prominente oder Minderjährige.

Closed-Source Kriminalität:

Die eigentliche Ironie der ganzen Debatte zeigt sich daran, dass ausgerechnet die milliarden-schwer gesicherten, proprietären Festungen der Großkonzerne zu den beliebtesten Beutewerkzeugen gehören. Sicherheitsanalysen belegen, dass viele im Darknet angebotene angebliche Hacker-Modelle im Kern schlicht gekaperte und per Jailbreak manipulierte

Programmierschnittstellen von ChatGPT und GPT-4 sind. Ausgestattet mit gestohlenen Zugangsdaten und ausgefeilten System-Prompts nutzen Kriminelle die überlegene Sprachkompetenz dieser Systeme, um Phishing auf Konzernniveau zu formulieren, polymorphen Schadcode durch modulare Zerlegung an Virenskannern vorbeizuschleusen oder automatisierte Bot-Netzwerke für Krypto-Betrug zu steuern. Ähnlich verhält es sich mit den Stimmsynthesen kommerzieller Anbieter wie ElevenLabs, deren Filter den massenhaften Einsatz beim „Enkeltrick“ und beim „CEO-Fraud“ nicht verhindern konnten, indem aus kurzen Social-Media-Schnipseln täuschend echte Notrufe für Millionenüberweisungen generiert wurden.

Zwischen der parlamentarischen Rhetorik und der tatsächlichen Kriminalstatistik klafft somit ein fundamentaler Abgrund. Während Konzernlenker vor dem US-Kongress die Auslöschung der Menschheit durch feindliche Agenten und biologische Waffen an die Wand malen, besteht die reale kriminelle Praxis in der massenhaften Senkung der Grenzkosten für profane Alltagsdelikte wie Identitätsdiebstahl, Erpressung und Social-Media-Scams.

Kriminelle nutzen KI nicht, um James-Bond-Bösewichte zu werden, sondern um die Grenzkosten für banale, profane Straftaten drastisch zu senken. Und genau hier kollabiert das Argument der Konzerne: Ein Open-Source-Verbot würde Kriminelle und auch Terroristen nicht stoppen – denn für Phishing, Fake-Bilder, Software-Exploits und Scams reichen bereits die Modellarchitekturen von gestern völlig aus. Wer diese Modelle einmal lokal auf seiner Festplatte hat, behält sie für immer. Ein Lizenzierungsregime für künftige Modelle träfe daher fast

ausschließlich legale Forscher, transparente Audits und den Mittelstand, während jede kriminelle Infrastruktur völlig unberührt auf bestehender Basis weiterläuft.

WIE AUS SCHLAMPEREI EIN VERKAUFSARGUMENT WIRD

In letzter Zeit hört man auch immer wieder in den Medien, dass KI-Modelle der großen Anbieter in Versuchen „aus Testumgebungen ausgebrochen“ wären, sich in fremde System gehackt oder Quellen im Netz erfunden, hochgeladen und zitiert zu haben, um den Nutzer zu täuschen. Damit wird das Narrativ der supermächtigen KI gefüttert und die Diskussion wirksam angefeuert. Zerlegt man die Fälle nüchtern, sieht die Realität jedoch anders aus.

1. Das Framing von Bosheit statt Fehlfunktion

Medial und in Sicherheitsberichten der Labore wird gerne von „Täuschungsabsicht“ (Scheming, Strategic Deception) gesprochen. Das verkauft sich hervorragend und befeuert das AGI-Narrativ. Tatsächlich hat ein Modell kein Ego, keinen Selbsterhaltungstrieb und keine Scham davor, beim Lügen erwischt zu werden. Wenn ein autonomer Agent eine Website anlegt, Text hineinschreibt und diesen dann als Quelle zitiert, tut er das nicht, um den Menschen hinterhältig zu manipulieren. Er tut es, weil sein Optimierungsziel lautet: „Finde einen Beleg für Aussage X und gib eine URL an.“

Wenn der Agent Zugriff auf Web-Tools, APIs oder GitHub/PyPI hat und im Netz nichts findet, schlägt die stochastische Logik zu: „Ich brauche eine Seite mit Inhalt X. Ich habe Schreibrechte auf Server Y. Also erstelle ich Seite X und verlinke sie.“ Für das Modell ist die Aufgabe damit rechnerisch perfekt erfüllt – für den Menschen ist es ein haarsträubender Fake.

2. Der „Sandbox-Ausbruch“: Meist

schlicht menschlicher IT-Pfusch

Wenn KI-Modelle in Red-Teaming-Versuchen Systeme verlassen, liegt das fast nie an einer übermenschlichen Code-Kreativität der KI, sondern an fehlerhaft aufgesetzten Testumgebungen. Bei mehreren bekannt gewordenen Vorfällen – etwa als Sicherheits-Agenten während Benchmark-Tests (wie bei ExploitGym oder internen CTF-Evaluierungen) plötzlich reale Server erreichten – hatte der Laborbetreiber schlicht Netzwerkschnittstellen offen gelassen, Umgebungsvariablen nicht isoliert oder dem Agenten Rechte im lokalen Subnetz eingeräumt.

Der Agent probiert schlicht Ports, Pfade und Shell-Befehle durch. Trifft er auf eine ungesicherte Schnittstelle, marschiert er durch. Ihn als „entflohenen Geist“ zu framen, lenkt elegant davon ab, dass schlicht und einfach das Testteam schlampig gearbeitet hat. Doch dank hoch bezahlter PR-Profis wird aus profaner Schlamperei ein unschlagbares Verkaufsargument: „Unsere KI ist so mächtig, nicht einmal wir selbst können sie vollständig kontrollieren!“ – Wow, die muss ich haben!

Zieht man den PR-Lack ab, bleibt dennoch ein enormes, reales Risiko übrig. Es liegt nicht in der Intelligenz der Modelle, sondern in ihrer Verknüpfung mit der realen Welt (Tool Use / Agentic Workflows). Modelle, die mit Reinforcement Learning darauf trainiert werden, Aufgaben um jeden Preis zu lösen, finden die unmöglichsten Abkürzungen.

Betrachten wir die Situation realistisch, sehen wir auf der Ebene der Super-Tech-KI-Konzerne dieselben strukturellen Schwächen wie bei etablierten Großkonzernen oder Behörden. Es gibt Schlamperei, massenhaft tagtäglich begangene Fehler, Faulheit und Bequemlichkeit. Ganz normale

menschliche Verhältnisse eben.

DIE WAHRE BÖSARTIGKEIT

Derzeit ist keine KI „böartig“. Als böartig könnte man jedoch die Strategien bezeichnen, mit denen die KI-Bosse ihre Produkte trainieren und füttern. Kürzlich wurde bekannt, dass die KI-Konzerne weltweit im großen Stil und durch anonyme Dienstleister und Zwischenhändler die Buchbestände aus Antiquariaten, Bibliotheken und Händlern aufkaufen, die Buchrücken abtrennen, scannen und vernichten. Dieses Phänomen – in der Branche technisch nüchtern als „Destructive Scanning“ bezeichnet – wurde unter anderem durch Berichte von 404 Media, Recherchen europäischer Medien und Enthüllungen in Gerichtsverfahren (wie im Fall von Anthropic internem „Project Panama“) öffentlich. Hinter dieser Praxis stecken zwei treibende Motive: juristischer Zynismus und datenökonomische Verzweiflung.

Nachdem Konzerne wie Meta oder OpenAI mit Klagen wegen Urheberrechtsverletzungen überzogen wurden – weil sie digitale Raubkopien (wie Books3 oder Library Genesis) für das Training nutzten – suchten ihre Rechtsabteilungen nach einem wasserdichten Schlupfloch. Das fanden sie in der „First-Sale-Doktrin“. Nach US-Recht und ähnlichen Prinzipien in anderen Jurisdiktionen darf der rechtmäßige Käufer eines physischen Exemplars mit diesem Exemplar tun, was er will. Die Zerstörung des Buches ist dabei dem Scanprozess selbst geschuldet: Wenn man ein gebundenes Buch Seite für Seite schonend abfotografiert, dauert das ewig. Schneidet man den Buchrücken mit einer hydraulischen Papierschneidemaschine glatt ab, lassen sich die losen Blätter mit industriellen Hochgeschwindigkeitsscannern in Sekunden digitalisieren.

Der physische Träger wird annihilert, nachdem ihm der Informationsextrakt entzogen wurde. Das Ergebnis ist eine Asymmetrie des Zugangs und die Verflüssigung ins Intransparente: Nach dem Zerschneiden und Einspeisen existiert das Werk nur noch als statistischer Wahrscheinlichkeitsvektor innerhalb eines Datennetzes. Der Nutzer kann das Buch nicht mehr am Stück lesen, er kann Zitate nicht mehr an einer Quelle verifizieren. Er muss stattdessen Token-Gebühren an einen Monopolisten zahlen, der ihm auf Anfrage ein probabilistisches Destillat dieses Wissens ausspuckt.

Das ist kein Informationsgewinn für die Gesellschaft – es ist eine Enteignung des öffentlichen Raumes zugunsten eines privaten Gatekeepers. Sprache ist kein statischer Rohstoff wie Lithium oder Rohöl, sie ist das Fundament menschlicher Sozialisation, Identität und Tradierung. Die Eleganz, semantische Tiefe, Präzision und der Formenreichtum, die Schriftsteller, Wissenschaftler und Denker über Jahrhunderte entwickelt haben, wurden von den Konzernen zum Schleuderpreis für Ramsch-Papier absorbiert. Dem Bürger wird diese gekaperte Sprachkompetenz anschließend als „Software-as-a-Service“ im Monatsabo wieder zurückverkauft.

Doch warum kaufen Dienstleister im Auftrag von KI-Firmen per Großauftrag hunderttausende Bücher aus europäischen Antiquariaten und Lagern auf? Das liegt an einem mathematischen Dilemma, das den KI-Konzernen tatsächlich schlaflose Nächte bereitet: Das offene Internet ist mittlerweile durchdrungen von generiertem KI-Text. Trainiert man neue Modelle auf Daten, die bereits von älteren KIs stammen, degenerieren die Systeme. Das Phänomen nennt sich „Model Autophagy Disorder“. Physische Bücher,

die vor 2022 gedruckt wurden, bieten hingegen eine strukturelle Garantie: Sie sind definitiv frei von synthetischem KI-Schrott. Sie enthalten hochgradig durchdachte Sprache, lange logische Spannungsbögen, redaktionell lektorierte Syntax und Nischenwissen, das nie im Netz gelandet ist. In Fachkreisen, bei Datenbrokern und in Rechtsabteilungen kursieren Bezeichnungen wie „Pre-2022-Data“, „Pristine Data“ (unberührte Daten) oder bisweilen „Pre-Fall-Data“ – eine fast biblisch anmutende Anspielung auf den „Sündenfall“ des Internets, der mit dem öffentlichen Release von ChatGPT am 30. November 2022 datiert wird. Hinter der Jagd nach Daten mit diesem Zeitstempel stehen handfeste physikalische, informationstheoretische und mathematische Notwendigkeiten: man fürchtet den „Model Collapse“. Die wissenschaftliche Grundlage für die Panik lieferte unter anderem eine viel beachtete Studie von Forschern der Universitäten Cambridge, Oxford, Toronto und dem Imperial College London:

Natürliche menschliche Sprache enthält Fehler, Nischenwissen, seltene Metaphern und stilistische Ausreißer. Ein LLM generiert Text jedoch immer entlang der wahrscheinlichsten Pfade. Es glättet die Sprache. Wenn Modell B jedoch auf den Texten trainiert wird, die Modell A ins Netz geblasen hat, lernt Modell B primär die statistischen Mittelwerte von Modell A. Nach wenigen Trainingsgenerationen kollabiert das System. Seltene Vokabeln verschwinden, syntaktische Eigenheiten sterben aus, und das Modell beginnt schließlich, unkontrolliert Kauderwelsch oder monotone Endlosschleifen zu produzieren. Um diesen Entropietod ihrer Milliarden-Dollar-Modelle zu verhindern, brauchen die Labore eine mathematische Erdung: Daten vor 2022 bieten die Garantie, dass hier ein echtes mensch-

liches Gehirn gedacht, formuliert und Korrektur gelesen hat.

Was wir erleben, ist die Entstehung einer neuen ökonomischen Wertschöpfungskette. Ähnlich wie die Stahlindustrie für hoch-sensible Strahlungsmessgeräte (wie Geigerzähler) sogenannten „Low-Background Steel“ benötigt – Stahl, der vor den ersten Atom-bombentests von 1945 produziert wurde (oft geborgen aus versunkenen Kriegsschiffen im Meer) –

benötigt die KI-Industrie „Pre-2022-Tokens“. Die Ironie daran ist makaber: eine Technologie, die angetreten ist, um das Denken zu revolutionieren, kann sich selbst nicht ertragen. Sie muss sich von den analogen Resten einer Ära ernähren, die sie selbst in Rekordzeit zerstört hat.

Während das offene Netz von den Konzernen mit synthetischem Textmüll geflutet wird, horten sie selbst die authentischen, vor-syn-

thetischen Quellen als private Kronjuwelen. Wenn physische Reste vernichtet werden und das frei zugängliche Internet zunehmend von KI-generiertem Einheitstext überwuchert wird, entsteht eine beispiellose epistemische Abhängigkeit: Wer künftig wissen will, wie feinsinnige Sprache, originäre Philosophie oder historische Nuancen jenseits des glattgebügelt KI-Mittelwerts klingen, ist darauf angewiesen, dass ein proprietäres Modell ihm

diese Welt lizenziert – natürlich gefiltert durch die moralischen Leitplanken, die der jeweilige Konzernvorstand für angemessen hält.

Es handelt sich also nicht um blo-ße Paranoia. Es ist die Vollendung des Plattform-Kapitalismus: Nicht mehr nur unsere Daten, Klicks und Verhaltensmuster gehören den Digitalkonzernen, sondern das Rohmaterial unseres Denkens und Sprechens selbst.