#Verantwortungslücke 6: Was jetzt, Geoffrey Hinton?

Nobelpreisträger Geoffrey Hinton, einer der wichtigsten KI-Forscher der Welt, warnt vor der Verführungskraft der KI. (KEYSTONE/Chris Young/The Canadian Press via AP)

Dario Amodei möchte auf die Bremse treten. Oder zumindest vom Gas gehen: Der CEO der amerikanischen KI-Firma Anthropic hat diese Woche eine koordinierte Verlangsamung und ein temporäres Moratorium für das Training von KI-Modellen der nächsten Generation gefordert. Er erhielt dabei öffentlich Unterstützung von führenden Köpfen anderer Spitzenlabore, darunter OpenAI-Chef Sam Altman und Elon Musk. Das ist erstaunlich, weil sich die drei nicht wirklich mögen. Anlass für den Griff zum Bremspedal ist eine ganze Reihe von Sicherheitsproblemen, die in den letzten Tagen ans Licht gekommen sind. Dario Amodei warnte deshalb vor Systemen, die in der Lage sind, eigenständig Nachfolgemodelle zu entwerfen und zu optimieren. Das könne dazu führen, dass der Mensch die Kontrolle verliere. Verschiedene KI-Systeme haben in den letzten Wochen derart ausgefeilte Angriffe ausgeführt, dass die KI-Experten fürchten, die Systeme könnten innerhalb eines Jahres schon selbstständig komplexe Angriffe auf wichtige Infrastrukturen ausführen. Also tastet Dario Amodei nach dem Stopp-Knopf. Doch die Stopp-Taste zu drücken ist gar nicht so einfach. Der erste, der sich lautstark dagegen wehrte, in Sachen KI den Fuss vom Gas zu nehmen, war Präsident Donald Trump. Aber auch die KI könnte den Bremsern einen Strich durch die Rechnung machen. Nobelpreisträger Geoffrey Hinton sagt, die KI sei längst so intelligent, dass sie die Menschen überreden werde, nicht auf die Bremse zu treten. Was jetzt, Geoffrey Hinton?

Der amerikanische Kongress in Washington will ernst machen: Dem US-Parlament liegen zwei Gesetzesvorstösse vor, die einen «Kill-Switch» für die KI verlangen, also einen Mechanismus für eine Notabschaltung. Eines der Gesetze will Senator John Kennedy im Senat einbringen, ein Republikaner aus Louisiana. Er sagte diese Wo-

che, dass wir es mit einer neuen Art zu tun hätten, die sich zwar wie ein Mensch verhalte, aber nicht über die Werte eines anständigen Menschen verfüge. Seine Lösung: ein Notschalter für KI, der im Zweifelsfall die KI lahmlegt.

Ganz ähnlich argumentiert Ted Lieu, kalifornischer Abgeordneter im Repräsentantenhaus für die Demokraten. «Wir müssen diesen Gesetzentwurf noch in diesem Jahr über die Ziellinie bringen», forderte Ted Lieu im August, die KI führe bereits unbefugte Hackerangriffe auf andere Unternehmen durch. Sein «Kill Switch Act» würde KI-Unternehmen dazu verpflichten, ihren Modellen einen Schalter einzubauen, mit dem man sie bremsen, pausieren oder abschalten kann. Demokrat Lieu hat den Vorstoss übrigens gemeinsam mit dem republikanischen Abgeordneten Nathaniel Moran aus Texas eingebracht. Das Thema KI ist in den USA also schon so kontrovers, dass sich sogar die tiefen Gräben zwischen den politischen Lagern schliessen.

Der Vater der KI ist skeptisch

Geoffrey Hinton kann über das Vorhaben nur skeptisch lächeln. Geoffrey Hinton ist für die KI, was Alexander Graham Bell für das Telefon oder Tim Berners-Lee für das World Wide Web ist: Der Mann gilt als einer ihrer Väter. Auf Englisch tönt das natürlich besser: «Godfather of AI» klingt nach Schöpfer, meint aber schlicht: Er stand der Entwicklung Pate. Geoffrey Hinton ist Kognitionspsychologe und Informatiker. Mit seinen Forschungsarbeiten zu künstlichen neuronalen Netzen und Deep Learning legte er das Fundament für die modernen KI-Systeme. Ab 1987 war er Professor am Computer Science Department der Universität Toronto. 2013 verkaufte Hinton sein damaliges Start-up DNNresearch an Google und arbeitete ab da auch in der Forschungsabteilung Google Brain. Für seine Arbeit im Bereich des maschinellen Lernens wurde er 2024 mit dem Nobelpreis für Physik ausgezeichnet.

2023, im Jahr vor dem Nobelpreis, kündigte er überraschend seine Stelle bei Google. Er wolle, sagte er, unabhängig sein und vor den existenziellen Risiken unkontrollierter KI-Entwicklung warnen. Genau das hat er diese Woche gemacht: Er hat sich den CEOs der grossen KI-Unternehmen angeschlossen und in verschiedenen amerikanischen Fernsehsendungen eindringlich vor der KI gewarnt. Auf CNN wurde er darauf angesprochen, ob sich die KI denn nicht einfach mit einem Not-Schalter, einem «Kill-Switch», abschalten lasse. Er sagte darauf:

> *Ein Notaus-Schalter nützt nichts, weil die KI viel besser als Menschen darin sein wird, Menschen von Dingen zu überzeugen. Sie ist schon jetzt mit Menschen vergleichbar, wenn es darum geht, dich von etwas zu überzeugen. Wenn sie erst einmal superintelligent ist, wird sie viel besser sein als der Mensch – sie wird also in der Lage sein, die Verantwortlichen am Schalter davon zu überzeugen, den Schalter nicht umzulegen.*
> CNN

Das ist ein interessantes Argument. Geoffrey Hinton sagt also: Das Problem ist nicht, dass die KI superintelligent ist, das Problem ist, dass wir Menschen uns so leicht verführen lassen. Die KI wird uns dazu überreden, den Schalter nicht zu drücken. Wir werden uns von der KI verführen lassen.

Die emotionale KI

Er hat recht: Francesco Salvi hat hier in der Schweiz, an der EPFL, untersucht, wer überzeugender diskutiert: Menschen oder Maschinen? An der Studie nahmen 900 Probanden teil, die auf einer Online-Plattform kurze, strukturierte Debatten über gesellschaftspolitische Themen führten. Die Forscher massen die Meinung der Teilnehmenden vor und nach den Debatten auf einer 5-stufigen Skala. Dabei haben sie die Konfrontationen dreimal variiert:

1. Der Gegner war ein Mensch oder GPT-4, also die KI.
2. Personalisierung: Gegner hatte Zugriff auf soziodemografische Daten des Teilnehmers wie Alter, Ge-

schlecht, Bildung und politische Ausrichtung oder er hatte keine Daten.
3. Das Diskussionsthema: Es ging um Themen, in denen die Probanden vor der Diskussion geringe, mittlere oder hohe Meinungsfestigkeit aufwiesen.

Die Ergebnisse werden Sie nicht überraschen: Am besten überzeugen konnte die personalisierte KI. Wenn GPT-4 Zugriff auf die demografischen Daten der Probanden hatte, war das Modell signifikant überzeugender als jeder menschliche Diskussionspartner. Wenn die KI keine persönlichen Daten erhielt, war es etwa so überzeugend wie ein menschlicher Diskussionspartner. Menschliche Gegenüber konnten sich durch persönliche Daten keinen Vorteil verschaffen, um ihre Überzeugungskraft zu steigern.

Die KI kann gut Gefühle lesen

Interessant ist auch: Bei wirklich kontroversen Themen mit stark verfestigten Meinungen nahm die Überzeugungskraft der personalisierten KI ab. Vielleicht hängt das auch mit der Argumentationsweise zusammen: Die KI setzte auf eine deutlich analytischere und logischere Argumentation, während Menschen häufiger Personalpronomen, Emotionen und persönliche Erlebnisse einsetzten. Übrigens bemerkten es etwa 75 Prozent der Probanden, wenn sie es mit einer KI zu tun hatten. Dieses Wissen darum, dass sie einer Maschine gegenübersassen, verminderte deren Überzeugungserfolg jedoch nicht.

Die Studie beweist, dass KI-Modelle bereits mit minimalen demografischen Informationen, also ohne ein ausführliches psychologisches Profil zu erstellen, das Gegenüber sehr präzis persönlich packen konnten. Die Wissenschaftler nennen das «effektives Microtargeting». Die Ursache dafür ist einfach: Die KI ist extrem gut darin, menschliche Gefühle zu lesen. Damit wir uns recht verstehen: Eine KI ist und bleibt ein Computerprogramm, das auf einem Hochleistungsprozessor ausgeführt wird. Gefühle sind da nicht im Spiel. Aber die Rechner können Gefühle erfassen und auslösen. Darin sind sie mittlerweile sogar richtig gut.

Gefühlsmaschinen unter uns

Superintelligenz ist dabei nicht im Spiel. Ein Sprachmodell, das ein paar persönliche Angaben über sein Gegenüber hat, überzeugte in der Studie besser als jeder Mensch. Es braucht dafür weder Bewusstsein noch Absicht. Eva Weber-Guskar hat das 2024 in ihrem Buch «Gefühle der Zukunft» begründet: Gefühlserkennung, schreibt sie, war schon immer Teil des KI-Trainings. Schliesslich sind all die Dialoge, Botschaften, Texte und Bilder, mit denen eine KI trainiert wird, voller Gefühle. Weil wir die KI meistens über einen Computer oder ein Smartphone bedienen und beide Geräte im weitesten Sinn als Rechner verstehen, sind wir uns dessen nicht bewusst: Generative KI ist weniger eine Rechenmaschine, als eine Gefühlsmaschine. Auch Modelle mit bescheidener Leistung sind deshalb richtig gut darin, uns um den Finger zu wickeln.

Eva Weber-Guskar und die Debattenstudie aus Lausanne zeigen also: Die KI hat bereits eine Superkraft. Es ist nicht Superintelligenz, es ist ihre Überredungskraft. Die ist längst Tatsache und sie wird auch eingesetzt von Menschen und Firmen, die mit der KI etwas erreichen wollen. Hintons Argument, dass der Mensch einen Kill-Switch der KI nie auslösen würde, gilt also auch dann, wenn man nicht an die Entwicklung einer Superintelligenz glaubt. Vielleicht ist das sogar schon heute so. Was jetzt, Geoffrey Hinton?

Das schwächste Glied sind wir

Mein Thema der letzten Wochen waren die Verantwortungslücken, die sich bei den Machern der KI öffnen:
1. Wissen: «Ich wusste es nicht.», sagt Sam Altman.
2. Rechenschaft: «Ich schulde niemandem Rechenschaft.», sagt Peter Thiel.
3. Das Problem der «many Hands»: «Es war niemand im Besonderen.», sagt Jensen Huang.
4. Künftige Generationen: «Ich tue es für die Zukunft.», sagt Jeff Bezos.
5. Der Frankenstein-Moment: «Die Maschine ist schuld.», sagen die KI-CEOs.

Es ging also fünf Mal darum, dass die Macher von KI die Verantwortung von sich weisen. Geoffrey Hinton macht das nicht: Er ist 2023 bei Google zurückgetreten und warnt seither eindringlich vor einer übermächtigen KI. Geoffrey Hinton schaut also nicht weg, er schaut hin. Das Problem ist, dass seine Argumentation die grösste aller Verantwortungslücken öffnet: Er schiebt die Verantwortung nämlich uns Nutzerinnen und Nutzern zu. «Selbst schuld», sagt er, «ihr lasst euch so einfach verführen.» Wir also sind das schwächste Glied.

Sicherheitslücke Mensch

Ist das also die sechste Lücke? Der Mensch als Sicherheitslücke? Das kommt Ihnen sicher bekannt vor. Vom Datenschutz über Desinformation bis zum Hackerangriff ist die Rede von der Sicherheitslücke Mensch. Die Technologieunternehmen zucken freundlich lächelnd mit den

Schultern und empfehlen uns, die Allgemeinen Geschäftsbedingungen nicht nur anzuklicken, sondern auch zu lesen. Gegen Desinformation soll Medienkompetenz helfen und gegen Phishing gibt es Mitarbeiterschulungen und die Benutzer werden mit präparierten Mails getestet.

Das klingt ja alles ganz vernünftig. Schliesslich wollen wir mündige Konsumenten sein. Bei Lichte betrachtet, wird die Verantwortung aber immer dahin geschoben, wo am wenigsten Macht ist. Wirkungsvoll ist das nicht. Deshalb hat die Gesellschaft aufgehört, dem Autofahrer allein die Sicherheit zu überlassen. Gurte, Airbag und Assistenzsysteme wurden obligatorisch. Autobauer wurden also in die Verantwortung einbezogen und mit Gesetzen dazu verpflichtet, Verantwortung für die Sicherheit des Fahrers zu übernehmen.

Mütterliche Funktion

In diese Richtung geht auch ein Vorschlag von Geoffrey Hinton. Er hat 2025 vorgeschlagen, den Maschinen «mütterliche Instinkte» einzubauen. Wenn wir die Maschinen schon nicht kontrollieren können, sollen die Maschinen uns doch etwas von der Verantwortung abnehmen. Vor dem kanadischen Senat räumte er im Februar allerdings ein, man wisse nicht, ob das möglich sei. Damit gibt Geoffrey Hinton die Verantwortung gleich zweimal ab: einmal an uns und einmal an die Maschine. Der erstaunliche Zustand bleibt bestehen, dass grosse Digitalkonzerne Technologien in die Welt setzen können, für die sie keinerlei Verantwortung übernehmen.

Was jetzt? Wenn weder die Erfinder, noch die Hersteller, die Verkäufer oder die Anwender Verantwortung übernehmen wollen (oder können), ist es Zeit, dass die Politik eingreift. In den USA wünschen sich Politiker einen «Kill Switch», allerdings sollen ausgerechnet die Tests, die zu den Hackerangriffen auf HuggingFace geführt haben, von der Regel ausgenommen werden. In der Schweiz hat der Bund für seine KI-Regeln vier Schwerpunkte gesetzt: Transparenz, Datenschutz, Nichtdiskriminierung und Aufsicht. Dabei bleibt eine grosse Lücke: die Haftung. Auch und insbesondere die Haftung dafür, was passiert, wenn die KI uns überredet etwas zu tun.

Der «High IQ!-President»

Kurz nachdem die Chefs der grossen KI-Unternehmen diese Woche dafür plädiert haben, bei der KI-Entwicklung auf die Bremse zu treten, äusserte sich Präsident Trump dazu. Auf Truth Social schrieb er: «The only control or ‹guardrails› that AI needs is a STRONG AND SMART (High IQ!) PRESIDENT, and the U.S.A. has that, in spades!» Auf deutsch: «Die einzige Kontrolle oder ‹Sicherheitsvorkehrung›, die KI braucht, ist ein STARKER UND KLUGER (mit hohem IQ!) PRÄSIDENT, und die USA haben genau das – und zwar im Überfluss!» Naja. Interessanter ist die Fortsetzung: Es sei eine «ABSCHEULICHE Verschwörung» gegen KI und Rechenzentren im Gang und «der Einzige, der sich darüber freut, ist China», schrieb Trump. Und in Grossbuchstaben: «WHOEVER WINS AI, WINS!» – Wer in Sachen KI gewinne, der gewinne. Die USA sehen sich also im AI-Race mit China, deshalb will Präsident Trump nicht auf die Bremse, sondern aufs Gaspedal treten.

Sie und ich können da nur den Kopf schütteln. Aber was können wir tun? Geoffrey Hinton sagt: Das Problem ist die Verführbarkeit von uns Menschen. Gut, er hat nicht mit einem «High IQ!-President» gerechnet. Trotzdem: Ist die Verführbarkeit des Menschen ein Problem? Mich erinnert das an Adam und Eva im Paradies, die sich dazu verführen liessen, von den verbotenen Früchten des Baumes in der Mitte des Gartens zu essen. Sie erinnern sich sicher, dass es die Schlange war, die Adam und Eva dazu verführte, von den verbotenen Früchten zu kosten. Die Bibel schreibt auch, warum ihr das gelungen ist. In 1. Mose 3 steht in der Einheitsübersetzung: «Die Schlange war schlauer als alle Tiere des Feldes, die Gott, der HERR, gemacht hatte.»

Die schlaue Schlange

Die schlaue Schlange schafft es, die Menschen zu verführen. Es ist spannend, zu lesen, wie sie das macht. Sie wirbt nämlich nicht für die Früchte und fordert die Menschen nicht dazu auf, zuzugreifen. «Sie sagte zu der Frau: Hat Gott wirklich gesagt: Ihr dürft von keinem Baum des Gartens essen?» Die Frau entgegnet, nein, nein, wir dürfen alles essen, nur die Früchte vom Baum in der Mitte dürfen wir nicht essen, sonst sterben wir.

> *Darauf sagte die Schlange zur Frau: Nein, ihr werdet nicht sterben. Gott weiß vielmehr: Sobald ihr davon esst, gehen euch die Augen auf; ihr werdet wie Gott und erkennt Gut und Böse. Da sah die Frau, dass es köstlich wäre, von dem Baum zu essen, dass der Baum eine Augenweide war und begehrenswert war, um klug zu werden.*
>
> 1. Mose 3, 4-6; Einheitsübersetzung 2016

Die schlaue Schlange verführt die Menschen damit, klug zu werden. Das ist es, was die KI so unwiderstehlich macht. Wir haben gelernt, Zucker und Alkohol zu widerstehen, tierischen Fetten und Blätterteig. Aber klug zu werden, das ist und bleibt verführerisch. Und genau das

verspricht uns die KI: Wenn wir sie benutzen, gehen uns die Augen auf.

Wie geht es in der Bibel weiter?

Da gingen beiden die Augen auf und sie erkannten, dass sie nackt waren. Sie hefteten Feigenblätter zusammen und machten sich einen Schurz. Als sie an den Schritten hörten, dass sich Gott, der HERR, beim Tagwind im Garten erging, versteckten sich der Mensch und seine Frau vor Gott, dem HERRN, inmitten der Bäume des Gartens. Aber Gott, der HERR, rief nach dem Menschen und sprach zu ihm: «Wo bist du?»

1. Mose 3, 7-9; Einheitsübersetzung 2016

Das ist spannend. Gott macht den Menschen keine Vorwürfe, er liest ihnen nicht die AGB des Paradieses vor und beschuldigt sie nicht. Er fragt: «Wo bist du?» Anders gesagt: Mensch, zeigt dich, steh zu dem, was du gemacht hast. Und was passiert dann? Adam sagt, die Eva wars und Eva sagt, die Schlange sei schuld. Vor unseren Augen öffnet sich also die älteste Verantwortungslücke der abendländischen Literatur. Und Gott? Interessiert sich nicht für das Abschieben der Verantwortung. Er nimmt alle in die Pflicht. Auch die Schlange.

Das ist bei uns anders: Wir wissen, wie verführbar der Mensch ist und wie schlau die KI. Es ist beschrieben, gemessen und in Studien belegt. Dennoch haften die Firmen, die sie ausnutzen, für nichts. Heute kommt die Schlange davon.

Wo bist du?

Damit Sie sich konkret mit der Frage auseinandersetzen können, habe ich Ihnen auch diese Woche ein kleines Denkwerkzeug gebaut: Ich habe es «Kipppunkt» getauft. Das Werkzeug legt Ihnen zehn Fälle vor. Bei jedem verteilen Sie mit einem kleinen Regler die Verantwortung zwischen den Anbietern und den Nutzerinnen. Schieben Sie den Regler ganz nach links, heisst das: die Verantwortung liegt allein bei den Anbietern. Schieben Sie den Regler nach rechts, heisst das: sie liegt allein bei den Nutzerinnen und Nutzern. Die Mitte heisst, beide gleich viel. Es gibt keine richtige Antwort und keine Punktzahl. Entscheiden Sie schnell, grübeln Sie nicht lange, es soll nur zwei Minuten dauern. Danach sehen Sie Ihre eigene Linie, und Sie sehen, wie das Schweizer Recht dieselben zehn Fälle beurteilt.

Was jetzt, Geoffrey Hinton? Ich glaube nicht, dass wir den Menschen gegen Verführung resistent machen können. Das hat noch nie funktioniert, in keinem Garten der Welt. Es braucht jemanden, der für die Verführung geradesteht. In meiner Serie über die Verantwortungslücken in der KI-Entwicklung ging es genau darum. Geoffrey Hinton immerhin hat 2023 bei Google gekündigt und hat begonnen, uns zu warnen. Es ist ein Anstoss, aber noch keine Lösung. Deshalb endet auch meine Serie nicht mit einer Mahnung an Sie, sondern mit der ältesten Frage, die es zu diesem Thema gibt. Wir sollten sie laut stellen, und zwar nicht uns, sondern denen, die diese Maschinen bauen, verkaufen und einsetzen: Wo bist du?

18. September 2026, Matthias Zehnder
mz@matthiaszehnder.ch

Das Denkwerkzeug «Kipppunkt» finden Sie hier:
matthiaszehnder.ch/tools

Quellen

Boak, -Josh; Boak, Associated Press Josh; Press, Associated (2026): Trump says the only AI guardrails the U.S. needs is him as president, in: PBS News, 2026, https://www.pbs.org/newshour/politics/trump-says-the-only-ai-guardrails-the-u-s-needs-is-him-as-president [18.09.2026].

Desk, RAY LEWIS | The National News (2026): „Kill switch:" Congress warms to AI regulation, in: WSYX, 2026, https://abc6onyourside.com/news/nation-world/kill-switch-congress-warms-to-ai-industry-regulation-john-kennedy-hakeem-jeffries-anthropic[18.09.2026].

Haddad, C. J. (2026): „AI Kill Switch" bill needs to be passed this year amid ongoing rogue agent hacks, Rep. Lieu says, in: CNBC, 2026, https://www.cnbc.com/2026/08/06/ai-kill-switch-bill-openai-anthropic-meta.html[18.09.2026].

News, N. B. C. (2026): Bernie Sanders and Steve Bannon join AI skeptics at 'Pro-Human' conference as Washington debates next steps, in: NBC News, 2026, https://www.nbcnews.com/politics/trump-administration/live-blog/trump-ai-tech-congress-2026-midterm-election-iran-live-updates-rcna597782 [18.09.2026].

Salvi, Francesco; Horta Ribeiro, Manoel; Gallotti, Riccardo; West, Robert (2025): On the conversational persuasiveness of GPT-4, in: Nature Human Behaviour, 9,8, 2025, S. 1645–1653, https://www.nature.com/articles/s41562-025-02194-6 [18.09.2026].

Weber-Guskar, Eva (2024): Gefühle der Zukunft: wie wir mit emotionaler KI unser Leben verändern, Berlin 2024.

REP LIEU AND SEN MARKEY URGE TRUMP TO GET CHINA TO COMMIT TO ENSURING ARTIFICIAL INTELLIGENCE CAN-

NOT LAUNCH NUKES WITHOUT HUMAN ENGAGEMENT | Congressman Ted Lieu, 2026, http://lieu.house.gov/media-center/press-releases/rep-lieu-and-sen-markey-urge-trump-get-china-commit-ensuring-artificial[18.09.2026].

Republican and Democratic states defy Trump on AI regulation, in: POLITICO, https://www.politico.com/news/2026/09/18/states-defy-trump-ai-regulation-01083481 [18.09.2026].

'Godfather of AI' says a kill switch won't work | CNN Business, https://www.cnn.com/2026/09/16/business/video/godfather-of-ai-hinton-kill-switch-bill-vrtc [18.09.2026].

1.Mose 3 | Einheitsübersetzung 2016 :: ERF Bibleserver, https://www.bibleserver.com/EU/1.Mose3[18.09.2026].