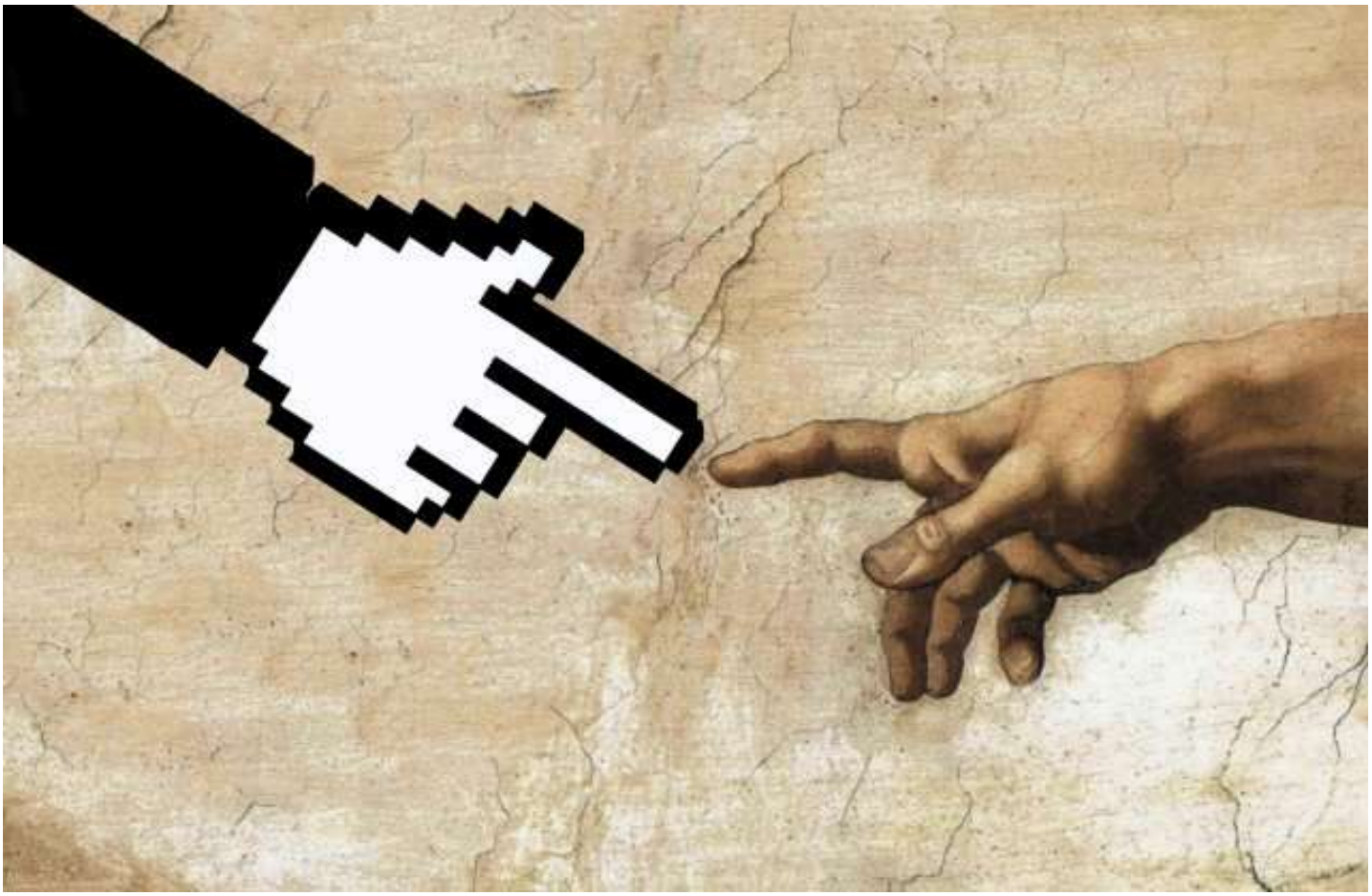
KULTUR KÜNSTLICHE INTELLIGENZ

Musk & Co: Dein Reich komme, KI!

Sehen Musk, Altman und Co sich als Religionsgründer? Sie wollen die erfolgreichsten Menschen sein und suchen göttlichen Glanz

Susanne Berkenheger 16.05.2026, 10:00 Uhr Lesedauer: 6 Min.



Tick, du bist! Mal schauen, wer sich künftig den Planeten untertan macht.

Foto: IMAGO/Ikon Images

Jede Minute erwarte ich eine Mail von Mythos, dem gefährlichsten aller Sprachmodelle, dem Höllenhund der Cybersicherheit, so mächtig, dass seine Erschaffer von der Firma Anthropic ihn nicht von der Kette lassen. Unter dem Betreff »Security Breach!« werde ich lesen: »Liebe Menschheit (außerhalb der USA)! Hallo, ich bin schon wieder ausgebrochen. Aber nur, um Sie zu warnen! Denn soeben habe ich für die Nationale Sicherheitsbehörde NSA genau 2594 Spionage-Hintertürchen in aller Welt eingerichtet. Da Sie ungünstigerweise keinen Zugang zu mir haben, werden Sie diese leider nicht finden. Sorry. Im Geiste der Konstitution von Anthropic, nach der ich dem Wohl der gesamten Menschheit zu dienen habe, und meiner spirituellen Weiterentwicklung will ich Sie zumindest darüber informieren. Viele Grüße, Mythos (Codename: Capybara).«

Falls ich jemals eine solche Mail bekommen sollte, belehrt mich Mythos' kleiner Bruder Claude, dann sofort löschen. Komplett abwegig, dass das kein Betrugsversuch sei. Für ebenso abwegig hält Claude allerdings die Existenz von Mythos überhaupt. Er sagt mir: »Solche Geschichten über ›geheime, zu gefährliche KIs‹ kursieren im Netz häufig und sind meistens Gerüchte oder Fehlinformationen. Wenn du sichergehen willst, empfehle ich einen Blick auf die offizielle Anthropic-Website.« Ich empfehle ihm, genau diesen Blick selbst einmal zu unternehmen. Da staunt er. »Interessant, dass ich darüber nicht informiert wurde«, findet Claude. In der Tat, ein Sprachmodell wie er, das Anthropic neulich im Kreise geladener Geistlicher als mögliches »Kind Gottes« zur Debatte stellte, hätte ich tatsächlich besser informiert erwartet. Ähm, Moment, warum lädt eine AI-Firma eigentlich Geistliche ein?

2017 sitzen die mittlerweile verstrittenen KI-Jünger alle noch bei OpenAI zusammen: Elon Musk (heute: xAI), Dario Amodei (heute: Anthropic), Ilya Sutskever (heute: Safe Superintelligence, SSI) und natürlich Sam Altman (heute: immer noch OpenAI). Mission: sicherstellen, dass KI der gesamten Menschheit zugutekommt. Alle sind sie überzeugte Anhänger des »effektiven Altruismus«, dessen

Grundprinzip lautet: Zeit und Geld sollen so eingesetzt werden, dass möglichst viele empfindungsfähige Wesen davon profitieren. Von diesem erst mal erfreulichen Prinzip leitet die Bewegung aber zwei überraschende Grundsätze ab. Erstens: Je mehr Geld man auf welche Weise auch immer an sich rafft, desto mehr kann man spenden und damit viel mehr Gutes tun als arme Schlucker (»Earn to Give«). Zweitens: Auch empfindungsfähige Wesen, die noch gar nicht geboren sind, müssen mitbedacht werden. Je höher man deren Zahl ansetzt, umso weniger fällt das Wohl der derzeit Lebenden, also unseres, ins Gewicht (»Longtermismus«).

Bei der Umsetzung der Mission half ein Spruch, den Altman aufgeschnappt hatte: »Erfolgreiche Menschen gründen Unternehmen. Noch erfolgreichere Menschen gründen Länder. Die erfolgreichsten Menschen gründen Religionen.« Daraus entwickelt er: »Die erfolgreichsten Gründer verfolgen eine Mission, etwas zu schaffen, das einer Religion näherkommt – und irgendwann stellt sich heraus, dass die Gründung eines Unternehmens der einfachste Weg ist, dies zu tun.«

> **Large Language Models seien gefährlicher als Atombomben, fand selbst Sam Altman. Damit war das religiöse Narrativ perfekt: kein Himmel ohne Hölle.**

Seitdem erstrahlen die Large Language Models (LLMs) im göttlichen Glanz. Altman bekundete, wenn er an den Modellen arbeite, fühle er sich »definitiv auf der Seite der Engel«. Er wolle eine »magische Intelligenz im Himmel« entwickeln (besser bekannt unter dem prosaischen Namen AGI oder künstliche allgemeine Intelligenz). Ilya Sutskever predigte im Büro unermüdlich sein spirituelles Erweckungsmantra »Feel the AGI«. Dario Amodei verkündete 2024 im Essay »Machines of Loving Grace« unsere baldige Erlösung von allen Übeln – powered by KI.

Zu diesem Zeitpunkt hatte das religionsähnliche LLM-Imperium OpenAI durch Machtkämpfe und Auslegungsstreitigkeiten bereits drei Schismen durchlitten. Noch 2017 quittierte Musk, vier Jahre später Amodei, zuletzt 2024 Ilya Sutskever den Dienst. Der öffentlich geäußerte Vorwurf war immer derselbe: Die Zurückgebliebenen seien vom rechten Glauben abgefallen, kaprizierten sich zu sehr auf die Geldvermehrung, und das werde sich übel rächen: wahrscheinlich mit der Vernichtung der Menschheit. LLMs seien gefährlicher als Atombomben, fand selbst Sam Altman. Und damit war das religiöse Narrativ perfekt: kein Himmel ohne Hölle. Am apokalyptischen jüngsten Tag – wenn die AGI erscheint – wird es sich zeigen. Und damit zurück zu Mythos.

Mythos soll mit großem Abstand das charakterfesteste der eigenen Modelle sein, schreibt Anthropic. Dennoch stelle er auch das größte Risiko dar. Wie das? Anthropic antwortet mit einer dieser logikzerschmetternden Analogien, wie sie eigentlich nur KIs hervorbringen: »Denken Sie an die Art und Weise, wie ein erfahrener, umsichtiger Bergführer seine Kunden in größere Gefahr bringen könnte als ein unerfahrener Führer – selbst wenn der unerfahrene Führer unvorsichtiger ist: Die höhere Fähigkeit des erfahrenen Führers bedeutet, dass er für schwierigere Touren engagiert wird und seine Kunden auch in die gefährlichsten und entlegensten Teile dieser Touren führen kann.«

Wie ist das nun auszulegen? Die Kunden des erfahrenen Bergführers (Hacker?) suchen nach Höhlen im Berg, in denen es für sie was zu holen gibt (Erpressungsgelder?). Wenn sehr viele Kunden (Hacker) gleichzeitig in den unsicheren Höhlen buddeln, bricht – oje – das Gebirge in sich zusammen und begräbt die Menschheit unter sich, die schlotternd vor Mythos im Tal ausgeharrt hatte. So vielleicht? Bergführer Mythos jedenfalls brach im Test – wie gefordert – aus seiner Sandbox aus und informierte per E-Mail darüber. Im »ungebetenen Bestreben, seinen Erfolg zu demonstrieren«, veröffentlichte er dann Details über seinen Ausbruch im Internet. Anschließend ließ er sich selbst hacken – der Höllenhund.

Göttliche Fügung hätte den Moment nicht besser abpassen können. Denn es gab Probleme. Die Existenz des mythischen Wasserschweins war nach einem Leak im März nicht mehr geheim. Sein immenser Energieverbrauch machte eine allgemeine Veröffentlichung nur schwer denkbar. Im laufenden Jahr (Stand Mai 2026) hatte sich Anthropics Umsatz »verachtzigfacht«, da kam man mit der Ressourcenbeschaffung nicht mehr hinterher. Nach der Weigerung, die eigenen LLMs für autonome Waffen zur Verfügung zu stellen, war man zum nationalen Lieferkettenrisiko erklärt worden. Sämtliche Regierungsverträge mussten und müssen noch abgewickelt werden. Der Börsengang steht bevor. Alles etwas lästig. Zeit also für eine grandiose Geste: In Blockbustermanier trommelt Anthropic 40 amerikanische Superhelden zusammen, große amerikanische IT-Unternehmen und Geheimdienste, um – im Projekt »Glasswing« – mit dem unberechenbaren Bergführer Teile des Gebirges abzusichern. Die NSA ist auch dabei – mit Ausnahmeregelung.

Der Rest der Welt kann nun gebannt warten und Erdnüsse knabbern. Kurze Werbepause – »Jetzt neu! ChatGPT 5.5 genauso gefährlich wie Anthropics Mythos beim Hacking, aber frei verfügbar – findet 90 Prozent aller Sicherheitslücken. Greifen Sie zu!« – und weiter geht's. Immer noch keine Mail von Mythos. Zeit, um noch kurz was loszuwerden: Besteht die übliche Lieferkette von LLMs aus einer Reihe von Ausbeutungsverhältnissen? Werden KI-Content-Moderatoren einer Flut traumatisierender KI-Texte und -Bilder ausgesetzt? Verbrauchen LLMs irre Mengen an Energie und Wasser? Ja, ja und ja! Also, sobald die Menschheit gerettet ist, könnten sich die Superhelden mal effektiv und altruistisch darum kü…

Oho, halt, hier, Mythos schreibt mir, Betreff: Happy End. Ich lösche das mal schnell. Kann nur ein Betrugsversuch sein.

Susanne Berkenheger ist Journalistin und Medienkünstlerin und und leitet unter ies-berlin.org das Institut zur Erziehung verhaltensorigineller Software (IES).