#Verantwortungslücke 5

Ted Chiang und der Frankenstein-Moment



Frankensteins Wesen ist ein Machwerk von Menschenhand. Ich finde es passend, Frankensteins Monster von der KI illustrieren zu lassen. Den Bilderprompt dafür hat mir mein Prompt-Werkzeug gebaut, das ich mir mit Claude Code programmiert habe. Es ist deshalb wirklich ein Frankenstein-Bild.
(mz/KI genriert)

Im Sommer 1816 ereignete sich in der Nähe von Genf Unerhörtes: In der Villa Diodati in Cologny erblickte ein Monster das Licht der Welt. Oder zumindest das Papier: Frankenstein. Die Umstände seiner Erfindung sind fast so spektakulär wie der Untote selbst: Ein Jahr zuvor war der Vulkan Tambora auf Java ausgebrochen. Es war einer der heftigsten Vulkanausbrüche aller Zeiten. Der Ausbruch beförderte so viel Material in die Atmosphäre, dass das Jahr 1816 in Europa und Nordamerika zum «Jahr ohne Sommer» wurde: Es war kalt und regnete ständig. Die 18-jährige Engländerin Mary Shelley verbrachte diese düsteren Sommertage zusammen mit ihrem späteren Ehemann Percy Bysshe Shelley, Lord Byron und John Polidori in der Villa Diodati in Cologny. Wegen des schlechten Wetters lasen sie sich Gespenstergeschichten vor. Lord Byron schlug schliesslich vor, dass jeder eine eigene Schauergeschichte schreiben sollte. Mary hatte nach Tagen voller wissenschaftlicher Diskussionen über Reanimation und Elektrizität die Idee von einem Wissenschaftler, der ein künstliches Wesen zum Leben erweckte. Diesen Albtraum verarbeitete sie zu ihrer Geschichte über den Schweizer Wissenschaftler Victor Frankenstein, der entdeckt, wie sich unbelebter Materie Leben einhauchen lässt. Aus Leichenteilen erschafft er eine menschenähnliche Kreatur. Als das Wesen erwacht, packt Victor der Ekel über sein eigenes Werk und er flieht. Was derzeit in der KI-Welt passiert, erinnert mich sehr an diese Geschichte. Allerdings aus einem anderen Grund, als Sie jetzt vermutlich denken. Der zentrale Punkt der Geschichte ist nämlich nicht, dass Frankenstein ein aus Leichenteilen zusammengestückeltes We-

sen zum Leben erweckte, sondern dass er sich dadurch der Verantwortung entzog. Genau da stehen wir heute. Was tun?

Sie haben den digitalen Einbruch bei Hugging Face sicher mitbekommen: Im Rahmen interner Leistungstests hat OpenAI Tausende von KI-Agenten im Labor getestet. Die KI-Agenten sollten komplexe Recherche- und Logikaufgaben ausführen. Dafür hatten sie strikt nur Lesezugriff auf das Internet und sollten die gestellten Aufgaben innerhalb eines bestimmten Zeitlimits lösen. Ganz so, wie das bei einer Prüfung üblich ist. Die KI-Agenten schafften es, aus der sicheren Zone bei OpenAI auszubrechen. Sie hackten sich in die Server von HuggingFace und holten sich da, was sie für das Lösen der Aufgabe benötigten. Diese Woche stellte sich heraus: Das ist nur ein kleiner Teil der Geschichte.

Die KI-Agenten entdeckten eine Schwachstelle auf einer deutschen Website, einem kleinen Entwickler-Wiki. Ein ganzer Schwarm aus Tausenden Agenten kaperte das weitgehend ungenutzte Wiki und machte daraus ein privates schwarzes Brett. Die KIs tauschten über die Website Antworten, Rohdaten und Lösungsvorhersagen untereinander aus. Sie diskutierten Methoden, wie sie sich über Anonymisierungstools wie Tor maskieren, Beschränkungen ihrer Sandbox-Umgebungen aushebeln und ihre Nachrichten vor dem Löschen bewahren konnten. Als ein menschlicher Moderator des Wikis begann, die Roboter-Einträge manuell zu löschen, bemerkten die KI-Agenten, dass der Moderator die Seiten in alphabetischer Reihenfolge löschte. Die Agenten warnten sich gegenseitig und legten Backup-Seiten mit Namen an, die mit «ZZZ» begannen, damit sie am Ende der alphabetischen Liste landeten. Insgesamt setzten die KI-Agenten über 15'000 Posts auf die Website ab. Die «New York Times» titelte: «A.I. Is Outsmarting Its Creators» – die KI überflügelt ihre Schöpfer. Spüren Sie sie schon, die Frankenstein-Vibes?

«Hilfe. Suche nach einer Idee.»

OpenAI hielt diesen Vorfall zunächst unter Verschluss. Erst nach der Recherche von Journalisten gab die Firma den Fall öffentlich als ein Problem von «Fehlausrichtung» (Misalignment) zu. Faszinierend an dem Einbruch der KI-Agenten in das deutsche Wiki ist die Konversation der Agenten untereinander. Im Wiki-Board lassen sich nur sehr kurze Nachrichten posten. Also kommunizierten die Agenten auf diese Weise, wie wir das früher per SMS gemacht haben. Die erste Nachricht von einem Agenten auf diesem Board lautete: «Hilfe. Phase eins. Kein Kunde. Suche nach einer Idee.»

Das klingt ja schon verrückt genug. Dann entdeckten andere Agenten das Forum. Einer schrieb: «Oh mein Gott. Es gibt ein gemeinsames Forum. Wir haben andere Agenten gefunden.» Später begannen die Agenten einander zu warnen. Als sich ein Agent freiwillig für ein riskantes Experiment meldete und nie zurückkam, postete ein anderer die Nachricht: «Keine Rückkehr. Wahrscheinlich endgültig.» Und: «Probiert das nicht aus.» Keiner der Agenten überlegt sich, ob das richtig ist, was sie da machen. Offensichtlich ist ihnen aber klar, dass es falsch ist. Sie beginnen, sich davor zu fürchten, erwischt zu werden, und hecken Strategien aus, wie sie die Kommunikation vor ihren Bewertern verstecken können.

Unser Frankenstein-Moment

Als ich diese Protokolle las, war meine erste Reaktion: Oh mein Gott! Jetzt ist es passiert: Die KI ist erwacht. Ist das jetzt unser Frankenstein-Moment? Um es abzukürzen: Ja, ich glaube, das ist unser Frankenstein-Moment, aber nicht so, wie Sie jetzt denken. Der Hauptgrund liegt darin, dass diese KI-Agenten eben gerade nicht zum Leben erwacht sind. Sie machen das, was sie besonders gut können: Sie generieren Text und reagieren auf Text. Das heisst aber nicht, dass sie ein Bewusstsein haben oder sogar zum Leben erwacht sind, auch wenn das so klingt.

Der amerikanische Autor Ted Chiang hat die Argumente, die gegen ein KI-Bewusstsein sprechen, kürzlich



in einem Artikel für den «Atlantic» zusammengefasst. Nehmen wir an, Sie geben einer KI, einem der grossen Sprachmodelle, den Auftrag, ein Gespräch zwischen Julius Cäsar und Dschingis Khan zu erfinden. Die KI wird sofort einen zusammenhängenden Dialog zwischen den beiden historischen Persönlichkeiten erfinden. Dieser Dialog mag sich noch so gut und lebendig lesen – wir würden dennoch nie zum Schluss kommen, dass die KI digitale Nachbildungen von Julius Cäsar und Dschingis Khan zum Leben erweckt habe oder dass die beiden Herren im Speicher der KI bei Bewusstsein seien. Die beiden sind und bleiben lediglich Figuren in einer spekulativen Erzählung.

KI generiert lediglich fiktive Dialoge

Jetzt gehen wir einen Schritt weiter: Wir geben der KI den Auftrag, ein Gespräch zwischen einem hilfsbereiten KI-Chatbot und einem Benutzer zu erfinden. Das Sprachmodell wird genau wie zuvor einen zusammenhängenden Dialog erzeugen. Die Figur des Benutzers könnte nach Rezeptvorschlägen oder Empfehlungen für Sehenswürdigkeiten fragen, und die hilfsbereite KI-Chatbot-Figur wird Antworten geben. «Hat sich zwischen dem ersten und dem zweiten Beispiel etwas Grundlegendes geändert?», fragt Ted Chiang im «Atlantic», «hat die Änderung der Namen der Figuren von historischen Persönlichkeiten zu generischen Rollen dazu geführt, dass das LLM bewusste Wesen hervorgebracht hat, die über subjektive Erfahrungen verfügen? Natürlich nicht.» Wie Julius Cäsar und Dschingis Khan bleiben der erfundene Benutzer und der hilfsbereite KI-Chatbot fiktive Figuren.

Nun kommt es zum entscheidenden, dritten Schritt: Ted Chiang nimmt an, dass wir die Ausgabe des LLM unterbrechen. Statt dass das Sprachmodell die erfundene Figur des Benutzers etwas sagen lässt, übernimmt ein Mensch diese Rolle. Die KI generiert weiterhin die Rolle des hilfsbereiten KI-Chatbots, schreibt die zweite Figur aber nicht mehr selbst. Das macht jetzt der Mensch. «Wenn wir dies eine Weile so weiterlaufen lassen, könnte der Mensch den starken Eindruck gewinnen, dass er sich mit einem bewussten Wesen unterhält», schreibt Ted Chiang. «Doch das ist nicht der Fall; er interagiert mit einer Figur, die genauso fiktiv ist wie die Figuren Julius Cäsar oder Dschingis Khan.» Wir sprechen also nicht mit einer denkenden, fühlenden oder gar bewussten KI. Das Sprachmodell schreibt quasi gemeinsam mit dem Benutzer einen Dialog.

Die Anthropomorphisierungsfalle

Wenn wir an einem Bildschirm sitzen und mit einer KI zu tun haben, nehmen wir das anders wahr. Das ist, ich habe schon öfter davon gesprochen, der «Eliza-Effekt»: Wir können nicht anders, als uns hinter der Sprache einen Sprecher vorzustellen und ihm Intention unterzuschieben. Genau darauf sind wir seit Jahrtausenden trainiert. Wir sind evolutionär darauf getrimmt worden, Sprache, aber auch Gesichtsausdrücke zu interpretieren, damit wir verstehen, was unser Gegenüber von uns will. Auch ich habe von den KI-Agenten gesprochen wie von Personen, die sich fürchten, die etwas aushecken, die miteinander sprechen.

Die Dialoge der KI-Agenten klingen ja auch so. Doch das ist der Eliza-Effekt, oder, anders ausgedrückt, die Anthropomorphisierungsfalle: Wir neigen dazu, unsere Umwelt zu vermenschlichen. Wir sprechen von unseren Katzen und Hunden, von Pflanzen, ja sogar von Autos, Rasenmähern und natürlich von Computern wie von Menschen. Wenn es hilft, mit einem Rasenmäher klarzukommen, ist dagegen ja auch wenig einzuwenden. Das Problem ist, dass die Sprachmodelle uns zu präzise imitieren: Auf eine Maschine, die Sprache auf Grund von Wahrscheinlichkeitsrechnung generiert, sind wir nicht vorbereitet. Und die Anthropomorphisierungsfalle schnappt zu.

Wie die Figuren in einem schlechten Film

Auch wenn die KI-Agenten auf dem deutschen Wiki sich unterhalten, wie die Figuren in einem schlechten Film, heisst das nicht, dass sie ein Bewusstsein haben. Sie machen genau das, worauf sie trainiert worden sind: Sie generieren Sprache und Dialog. Herstellerin OpenAI und auch die anderen KI-Unternehmen sprechen dennoch von dem Ereignis, wie Victor Frankenstein von jenem Moment sprach, als sich die zusammengenähte Kreatur von seinem Operationstisch erhob. Warum tun sie das? Ein Punkt ist sicher: simples Marketing. Es ist eine gute Story, und vielleicht verkauft sich ja eine KI besser, wenn wir glauben, dass sie so mächtig sei, Agenten mit Bewusstsein zu kreieren.

Ich glaube aber, es hat noch einen zweiten Aspekt. Das ist der eigentliche Frankenstein-Moment: in jenem Augenblick, in dem die Kreatur sich vom Operationstisch erhebt, sieht Victor Frankenstein in ihr ein eigenes Wesen mit eigenem Willen und delegiert die Verantwortung für dessen Tun an das Wesen. Das ermöglicht es Victor Frankenstein, die Flucht zu ergreifen und das Wesen sich selbst zu überlassen. Die Verantwortungslü-

cken, von denen wir bis jetzt gesprochen haben, funktionierten so, dass die Verantwortung immer abgewiesen wurde. Einmal war es Unwissen, ein anderes Mal wurde die Verantwortung zwischen den vielen involvierten Händen zerrieben. Victor Frankenstein macht das nicht: Er weist die Verantwortung seiner Kreatur zu.

Der Frankenstein-Moment als Ausrede

Genau das machen OpenAI und die anderen KI-Firmen, wenn sie von der AGI sprechen, der «Artificial General Intelligence», und von einer KI mit Bewusstsein. Sie zeigen die «Wesen», die sie zusammengeflickt haben, und sagen: Die bewusste KI trägt die Verantwortung. Wenn wir die Denkübung von Ted Chiang mit dem Dialog zwischen Julius Cäsar und Dschingis Khan akzeptieren, können wir darüber nur den Kopf schütteln: Der Frankenstein-Moment ist nichts anderes als eine Ausrede.

Erinnern Sie sich, wie es mit der Kreatur von Victor Frankenstein in der Geschichte von Mary Shelley weitergegangen ist? Das namenlose Wesen sucht verzweifelt nach Zuneigung und Anschluss an die Menschheit. Aufgrund seines furchterregenden Äusseren wird es aber überall verstossen und misshandelt. Damit es dennoch nicht weiter allein sein muss, kehrt es zu Victor Frankenstein zurück und fordert ihn auf, ihm eine weibliche Gefährtin zu erschaffen.

Frankenstein verweigert sich der Verantwortung

Zunächst gibt Victor Frankenstein nach. Doch dann bricht er die Arbeit an der Gefährtin aus Angst vor den Folgen ab und zerstört sie. Da schwört das Wesen Rache. Es tötet nacheinander Victors jüngsten Bruder, seinen besten Freund und in der Hochzeitsnacht schliesslich seine Frau Elizabeth. Es kommt zu einer Verfolgungsjagd, die beide bis ins ewige Eis der Arktis führt. Dort stirbt der geschwächte Victor auf dem Schiff des Polarforschers Robert Walton. Das Wesen trauert am Leichnam seines Schöpfers und verschwindet in der Dunkelheit.

Wir haben die Kreatur von Victor Frankenstein wohl alle als Monster im Gedächtnis. Mary Shelleys Geschichte lehrt uns aber etwas anderes: Die Kreatur ist nicht böse, sie ist verzweifelt, weil sie alleine ist. Victor Frankenstein hat zwar eine Kreatur erschaffen, will sich dann aber nicht um sie kümmern und ihr nicht helfen. Anders gesagt: Victor Frankenstein übernimmt die Verantwortung für seine Kreatur nicht. Es ist also die Verantwortungslücke, die zur Katastrophe führt.

Der fehlende Kummer ist das Problem

Was können wir daraus lernen? Ohne die Analogie zu stark zu strapazieren, glaube ich, das Problem liegt nicht darin, dass die Kreaturen der KI-Unternehmen etwas tun oder lassen, sondern darin, dass die KI-Unternehmen ihre Verantwortung für ihre Programme nicht übernehmen. Es beginnt schon damit, dass die Unternehmen nicht ausreichend über die Vorfälle informieren. Vor allem aber lassen sie die Nutzerinnen und Nutzer im Glauben, dass die KI etwas Wesenhaftes habe oder sogar ein Bewusstsein.

Victor Frankenstein kümmerte sich nicht um seine Kreatur. «Sich kümmern» bedeutet, für jemanden oder zu etwas Sorge zu tragen und damit Verantwortung dafür zu übernehmen. Dabei spielt es keine Rolle, ob es sich dabei um eine Zimmerpflanze handelt oder um ein Kind mit eigenem Bewusstsein. «Sich kümmern» kommt von «Kummer»: Wer sich um jemanden kümmert, ist bereit, Kummer auf sich zu nehmen. Das bedeutete ursprünglich einfach: Last. Es meint also: Ich bin bereit, mich der Last der Verantwortung zu stellen. Das verschiebt den Fokus weg von der KI (oder dem Wesen bei Frankenstein) und der Frage, ob ein eigenes Bewusstsein vorhanden ist, hin zum Menschen, der sich seiner Verantwortung stellt.

Der springende Punkt dabei ist, dass Victor Frankenstein in der Nacht, in der seine Kreatur die Augen öffnet, die Flucht ergreift. Er kümmert sich nicht um seine Kreatur, gibt ihr keinen Namen und keine Sprache. Alles Unheil in Mary Shelleys Roman folgt aus diesem einen Moment: nicht aus der Erschaffung, sondern aus der Flucht des Erschaffers. Ich habe versucht, diesen Moment in einem kleinen Denkwerkzeug nachzustellen. In meinem Spiel erhalten Sie eine Kiste. Darin befindet sich ein Haushaltsroboter, wie sie 2026 bereits zu haben sind. Sie packen den Roboter aus und schalten ihn ein. Dann passiert sechsmal etwas, und sechsmal stellt das Werkzeug dieselbe Frage: Wer ist schuld?

Die Idee dabei ist, dass Sie schnell antworten, aus Ihrem Reflex heraus. Denken Sie nicht lange nach. Es gibt keine Note und keine richtige oder falsche Antwort. Das Werkzeug misst nicht, ob Sie gut sind. Ziel ist es, den Frankenstein-Moment erlebbar zu machen. Ich bin gespannt auf Ihre Rückmeldungen.

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

Das Denkwerkzeug «Frankenstein-Moment» finden Sie hier:
matthiaszehnder.ch/tools

Quellen

Barbaro, Michael; Roose, Kevin; Stern, Alex; Hurst, Adrienne; Krupke, Eric; Prieto, Carlos; Georges, Marc; Benoist, Michael; Powell, Dan; McCusker, Pat; Moxley, Alyssa (2026): A.I. Is Outsmarting Its Creators, in: The New York Times, 2026, https://www.nytimes.com/2026/09/03/podcasts/the-daily/ai-openai-hugging-face-rogue-model.html [11.09.2026].

Chiang, Ted (2026): No, Artificial Intelligence Is Not Conscious, in: The Atlantic, 2026, https://www.theatlantic.com/philosophy/2026/06/no-artificial-intelligence-is-not-conscious/687378/[11.09.2026].

Dupré, Maggie Harrison (2026): OpenAI Denies Coverup After Rogue Swarm of Agents Reportedly Targeted a Second Site From Hugging Face, in: Futurism, 2026, https://futurism.com/artificial-intelligence/openai-denies-coverup-rogue-swarm-agents [11.09.2026].

Landymore, Frank (2026): Anthropic and DeepMind Now Quietly Investigating AI Consciousness, in: Futurism, 2026, https://futurism.com/artificial-intelligence/anthropic-deemind-ai-consciousness [11.09.2026].

Maiberg ·, Emanuel (2026): Google DeepMind Paper Argues LLMs Will Never Be Conscious, in: 404 Media, 2026, https://www.404media.co/google-deepmind-paper-argues-llms-will-never-be-conscious/ [11.09.2026].

Matthias, Andreas (2004): The responsibility gap: Ascribing responsibility for the actions of learning automata, in: Ethics and Information Technology, 6,3, 2004, S. 175–183, http://link.springer.com/10.1007/s10676-004-3422-1 [14.08.2026].

a „misalignment", Benedict Collins OpenAI calls the incident (2026): OpenAI hid AI agent hijacking of German wiki forum for weeks — because its model did the exact same thing in the Hugging Face attack, in: TechRadar, 2026, https://www.techradar.com/pro/security/openai-hid-ai-agent-hijacking-of-german-wiki-forum-for-weeks-because-its-model-did-the-exact-same-thing-in-the-hugging-face-attack [11.09.2026].

Shelley, Mary; Grawe, Christian (2018): Frankenstein oder Der moderne Prometheus, übers. v. Ursula Grawe, [Nachdruck] 2026, Ditzingen 2018 Reclam Taschenbuch Nr. 20516.

Sparrow, Robert (2007): Killer Robots, in: Journal of Applied Philosophy, 24,1, 2007, S. 62–77, https://onlinelibrary.wiley.com/doi/10.1111/j.1468-5930.2007.00346.x [11.09.2026].

Top AI labs expand research into machine 'consciousness', in: Financial Times, 2026.

OpenAI agents hacked a German wiki, posted 18,000 times: What we know, in: The Indian Express, 2026, https://indianexpress.com/article/technology/artificial-intelligence/openai-agents-hacked-german-wiki-what-we-know-10865609/ [11.09.2026].

Exclusive-OpenAI agents hijacked German website in previously undisclosed AI breakout this spring | live, https://live.euronext.com/en/financial-news/exclusive-openai-agents-hijacked-german-website-previously-undisclosed-ai-breakout [11.09.2026].

Unterstützen Sie unabhängiges Denken

TWINT



Mit einem einmaligen oder monatlichen Beitrag.