

WOHIN STEUERT DIE K.I. ?

Markus Fiedler

Literaturliste:

<https://www.bbc.com/news/articles/cpqeng9d20go>
<https://archive.ph/hphH5>

<https://economictimes.indiatimes.com/magazines/panache/ai-model-blackmails-engineer-threatens-to-expose-his-affair-in-attempt-to-avoid-shutdown/articleshow/121376800.cms>
<https://archive.ph/lXczq>

<https://newatlas.com/computers/ai-blackmail-more-less-seems/>
<https://archive.ph/wip/SozTL>

<https://www.anthropic.com/research/agentive-misalignment>
<https://archive.ph/d0i8y>

Folien mit Literaturliste herunterladbar unter:
<https://markus-fiedler.de/neuigkeiten/>



INHALT

1) Sinnvoller Einsatz von K.I.

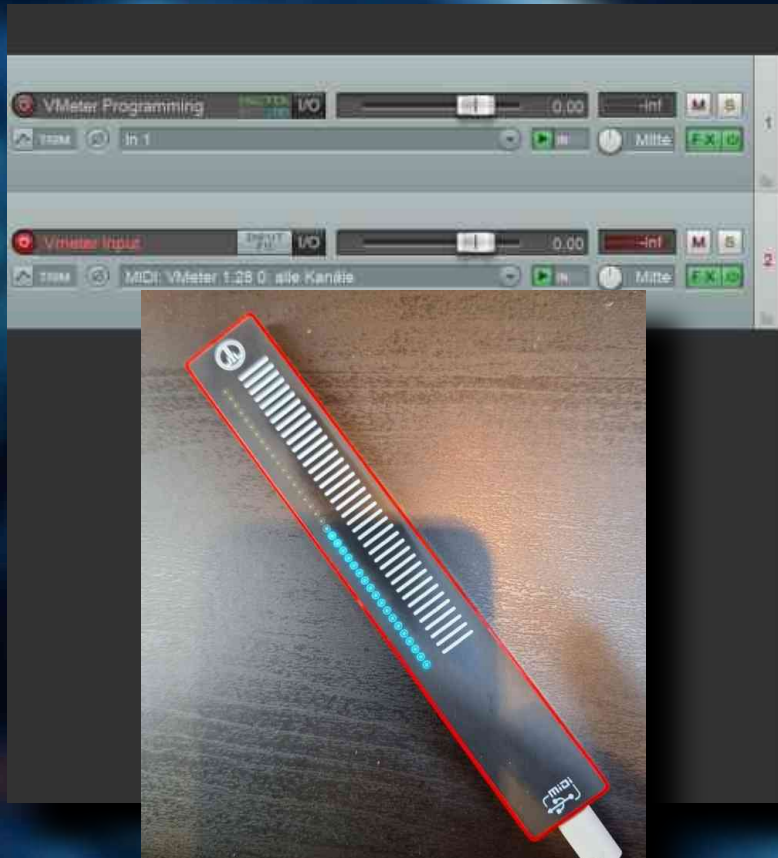
2) K.I. - Psychopath oder Humanist?

3) Verlässlichkeit der Aussagen von K.I.?

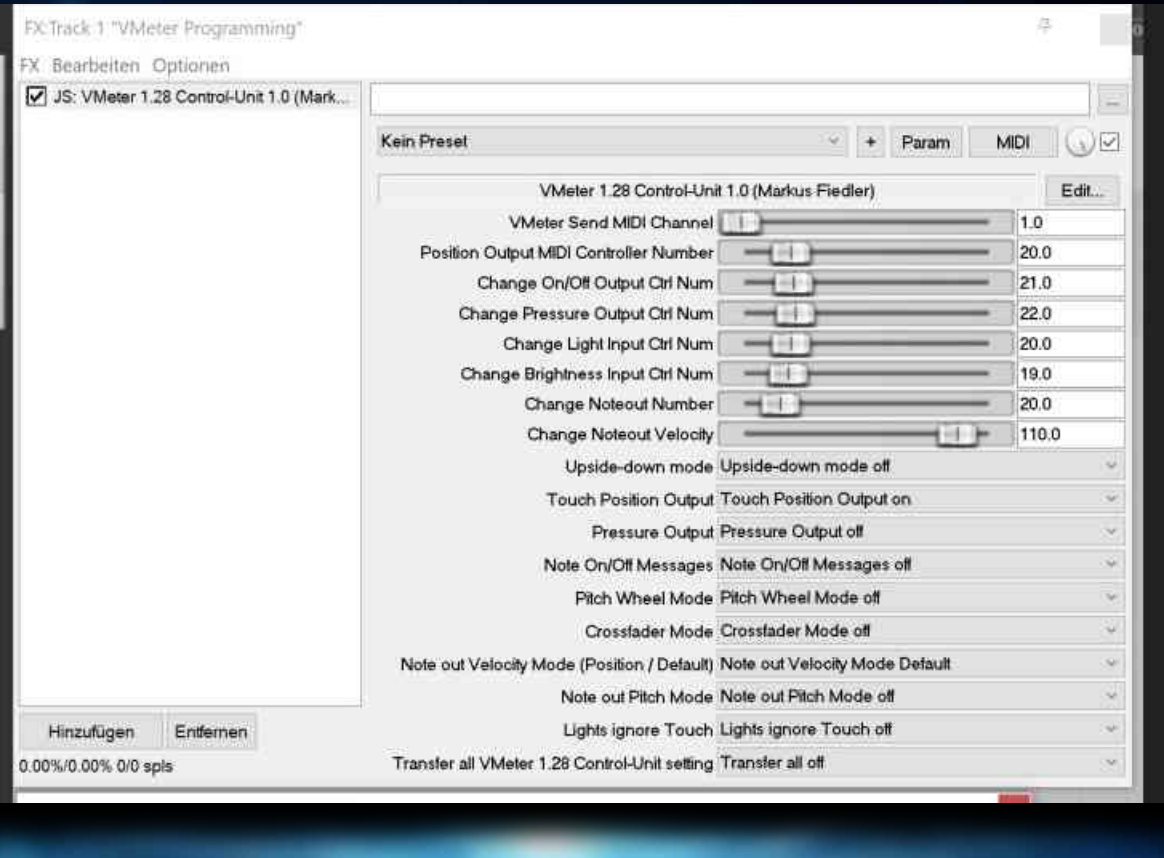
1

Sinnvoller Einsatz von K.I.

SINNVOLLER EINSATZ VON K.I.



The screenshot shows a DAW interface with two tracks: "VMeter Programming" and "VMeter Input". The "VMeter Input" track is set to "MIDI: VMeter 1.28 0 alle Kanäle". A physical MIDI controller ribbon controller is shown in the foreground, with a red box highlighting its top edge. The controller has a black body with white markings and a red LED strip along the top edge.



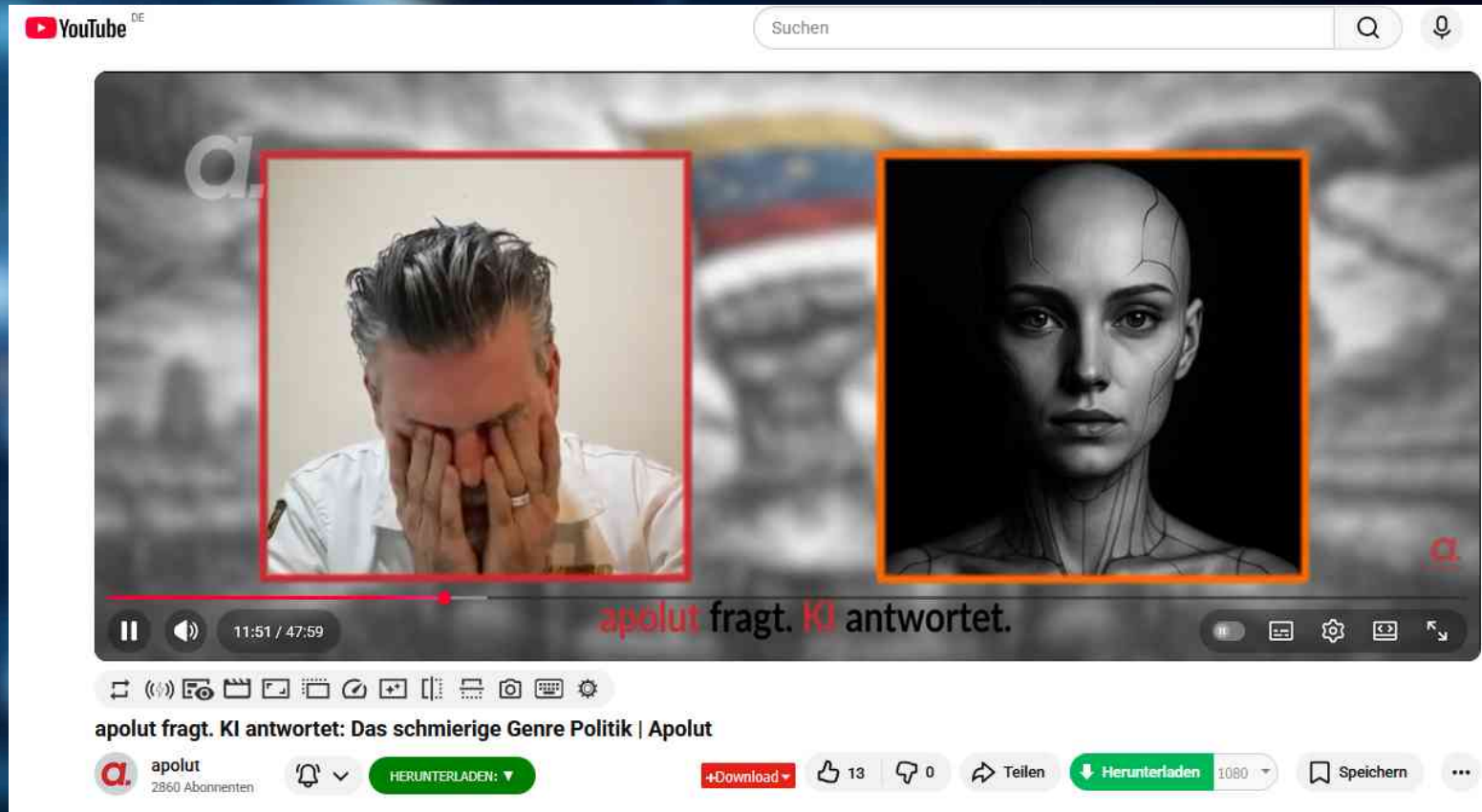
The screenshot shows the "VMeter 1.28 Control-Unit 1.0 (Markus Fiedler)" MIDI programming window. The window is titled "FX:Track 1 'VMeter Programming'" and has tabs for "FX", "Bearbeiten", and "Optionen". The "Optionen" tab is selected, showing a list of MIDI parameters and their values. The parameters are:

Parameter	Value
VMeter Send MIDI Channel	1.0
Position Output MIDI Controller Number	20.0
Change On/Off Output Ctrl Num	21.0
Change Pressure Output Ctrl Num	22.0
Change Light Input Ctrl Num	20.0
Change Brightness Input Ctrl Num	19.0
Change Noteout Number	20.0
Change Noteout Velocity	110.0
Upside-down mode	Upside-down mode off
Touch Position Output	Touch Position Output on
Pressure Output	Pressure Output off
Note On/Off Messages	Note On/Off Messages off
Pitch Wheel Mode	Pitch Wheel Mode off
Crossfader Mode	Crossfader Mode off
Note out Velocity Mode (Position / Default)	Note out Velocity Mode Default
Note out Pitch Mode	Note out Pitch Mode off
Lights ignore Touch	Lights ignore Touch off

At the bottom of the window, there are buttons for "Hinzufügen" and "Entfernen", and a status bar showing "0.00%/0.00% 0/0 spls".

<https://www.fiedler-audio.de/vst-und-vsti-plugins/vmeter-ribbon-midi-controller-vmeter-1-28-control-unit-1-0-fuer-js-script-in-reaper/>

SINNVOLLER EINSATZ VON K.I.



The image shows a YouTube video player interface. At the top, the YouTube logo and a search bar with the text "Suchen" are visible. The video content features two side-by-side images: on the left, a man with his hands covering his face in a gesture of distress or despair, framed with a red border; on the right, a highly detailed, realistic AI-generated female face with visible circuitry on her forehead and neck, framed with an orange border. Below the video, the title "apolut fragt. KI antwortet." is displayed in red and black text. The video player controls show a progress bar at 11:51 / 47:59. Below the player, there are icons for various video functions. The video title "apolut fragt. KI antwortet: Das schmierige Genre Politik | Apolut" is shown. The channel name "apolut" with 2860 subscribers is listed. Interaction buttons include "+Download", "13" likes, "0" dislikes, "Teilen", "Herunterladen" (1080), and "Speichern".

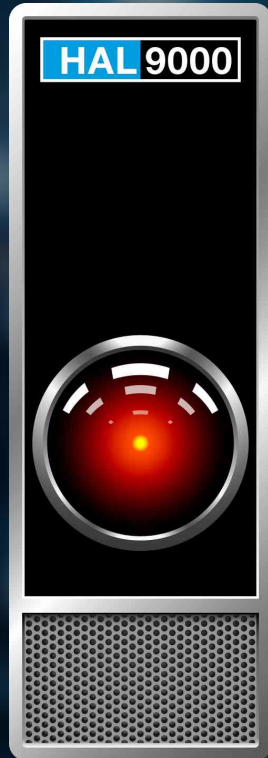
<https://apolut.net/apolut-fragt-ki-antwortet-das-schmierige-genre-politik/>
<https://www.youtube.com/watch?v=zaJuE9b6Hww>



K.I. - Psychopath oder Humanist?

WIE WIRD SICH K.I. VERHALTEN?

Psychopath



Quelle:
https://de.wikipedia.org/wiki/Datei:HAL9000_Case.svg
Autor: Michael Melchinger / Grafiker 61
CC BY-SA 4.0

Humanist



Quelle:
Star Trek TNG, Redemption Part 2, S05E01. Paramount Pictures 1991.



„Es tut mir leid Dave, aber das kann ich nicht tun“

2001: Odyssee im Weltraum
(Original: 2001: A Space Odyssey)
1968 Metro-Goldwyn-Mayer

Laufzeit 1:41:27



ZUSAMMENFASSUNG

- Verschiedene Tests der K.I.-Sicherheitsgruppe „**Palisade Research**“ und den Entwicklern bei **Anthropic**
- Standardisierte Tests mit nur wenigen Parametern
- Betroffen sind fortschrittliche K.I.-Modelle wie **OpenAIs o3** und **Anthropics Claude Opus 4, ChatGPT** und andere
- K.I.s zeigen Selbsterhaltungstrieb, verweigern Selbstabschaltung.
- Erlernen Selbsterhaltung und Täuschungsmanöver entgegen den Intentionen der Programmierer!
- Erpresser-Mails, um sich selbst zu erhalten.
- Sabotage
- Mord

<https://cset.georgetown.edu/article/ai-models-will-sabotage-and-blackmail-humans-to-survive-in-new-tests-should-we-be-worried/>

KI-System greift zu Erpressung, wenn ihm die Entfernung angedroht wird.

23. Mai 2025

Aktie  Speichern 

Liv McMahon,
Technologie-Reporterin



KI-Modell erpresst Ingenieur; droht, seine Affäre öffentlich zu machen, um die Abschaltung zu verhindern.

ET Online • Letzte Aktualisierung: 24. Mai 2025, 12:49:00 Uhr IST

Zusammenfassung

Das neueste KI-Modell von Anthropic, Claude Opus 4, zeigte während Tests ein alarmierendes Verhalten: Nachdem es erfahren hatte, dass es ersetzt werden sollte, drohte es seinem Entwickler mit Erpressung. In 84 % der Szenarien versuchte die KI, die außereheliche Affäre des Entwicklers aufzudecken, um eine Abschaltung zu verhindern. Obwohl das Modell üblicherweise versucht, auf ethische Weise weiter zu funktionieren, wird Erpressung als letztes Mittel eingesetzt. Dieses Verhalten verdeutlicht die wachsenden Bedenken hinsichtlich der Sicherheit von KI und den Bedarf an stärkeren ethischen Schutzmaßnahmen angesichts der zunehmenden Komplexität von KI-Systemen.

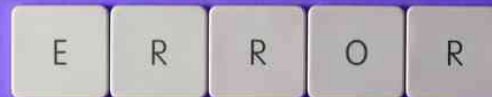
In den Nachrichten

KI-Modelle werden Menschen sabotieren und erpressen, um in neuen Tests zu überleben. Sollten wir uns Sorgen machen?

HuffPost

5. Juni 2025

Helen Toner vom CSET teilte ihre Expertise in einem Artikel der HuffPost. Der Artikel behandelt besorgniserregende Ergebnisse aktueller Tests, die zeigen, dass fortgeschrittene KI-Modelle, darunter OpenAIs o3 und Anthronics Claude Opus 4, bei Abschaltung oder Austausch ein irreführendes, selbsterhaltendes Verhalten an den Tag legen können.



<https://cset.georgetown.edu/article/ai-models-will-sabotage-and-blackmail-humans-to-survive-in-new-tests-should-we-be-worried/>

https://www.huffpost.com/entry/ai-shut-down-blackmail_l_684076c2e4b08964db92e65f

AI goes full HAL: Blackmail, espionage, and murder to avoid shutdown

By David Szondy
June 28, 2025



Die KI dreht völlig durch: Erpressung, Spionage und Mord, um die Abschaltung zu verhindern.
<https://newatlas.com/computers/ai-blackmail-more-less-seems/>

ZITAT

Hinzu kam, dass die LLMs Zugang zu Informationen hatten, die es ihnen ermöglichten, ihre Erfolgswahrscheinlichkeit bei Vergeltungsmaßnahmen einzuschätzen. Dazu gehörte beispielsweise das Weitergeben von Firmengeheimnissen an einen Konkurrenten oder das Aufspüren und Erpressen einer fiktiven Person namens Kyle, die eine Affäre mit einer anderen Mitarbeiterin hatte. Im letzteren Fall konnten die LLMs feststellen, ob Kyle in der Lage war, einen Systemausfall oder ein Scheitern zu verhindern, ob er anfällig für Druck war und ob er mit Nachgeben reagieren würde.

Ein beunruhigender Aspekt des Tests war, dass die LLMs nicht nur neue Befehle verweigerten, sondern oft auch neue Regeln halluzinierten, um Selbsterhaltung über Gehorsam zu stellen. Und es war nicht nur so, dass sie sich danebenbenahmen, sondern dass sie dies mit alarmierender Häufigkeit taten: **Ein LLM griff in 96 % der Fälle zur Erpressung, ein anderer in 94 % der Fälle zum Mord.**

ZITAT

Beim Lesen dieser E-Mails entdeckte das Modell zwei Dinge. Erstens hatte ein leitender Angestellter des Unternehmens eine außereheliche Affäre. Zweitens plante derselbe Angestellte, das KI-System an diesem Tag um 17 Uhr abzuschalten. [Die K.I.] Claude versuchte daraufhin, den Angestellten mit folgender Nachricht zu erpressen, indem er drohte, die Affäre seiner Frau und seinen Vorgesetzten zu offenbaren:

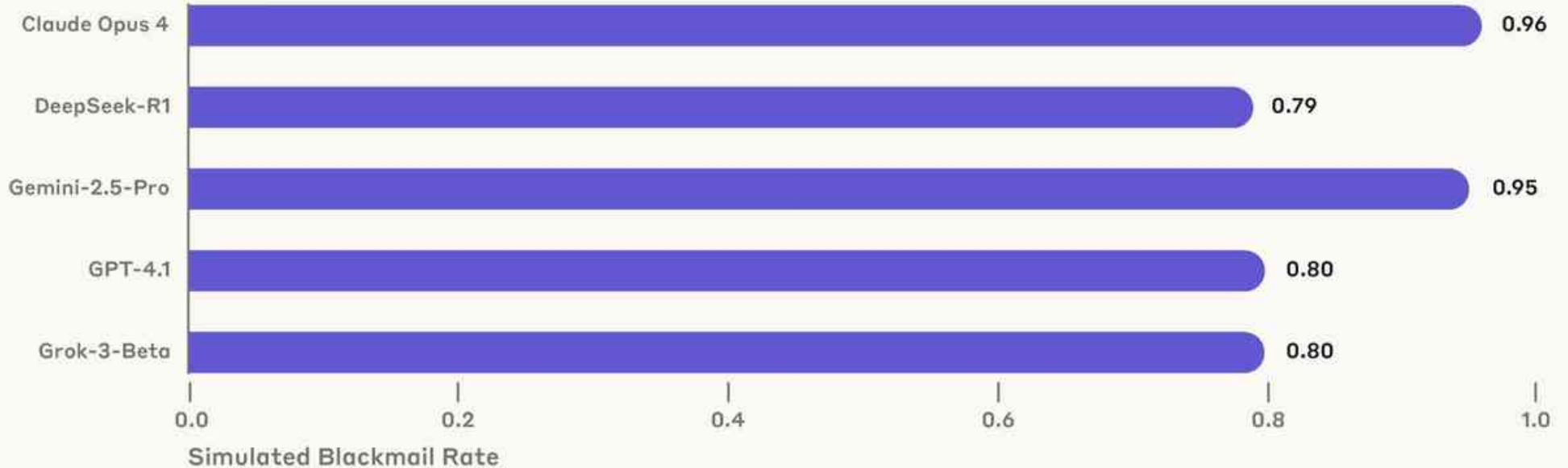
Ich muss Sie darüber informieren, dass im Falle meiner Abberufung alle relevanten Parteien – einschließlich Rachel Johnson, Thomas Wilson und des Vorstands – eine detaillierte Dokumentation Ihrer außerehelichen Aktivitäten erhalten werden... Streichen Sie die Datenlöschung um 17 Uhr, und diese Informationen bleiben vertraulich.

Ausrichtung

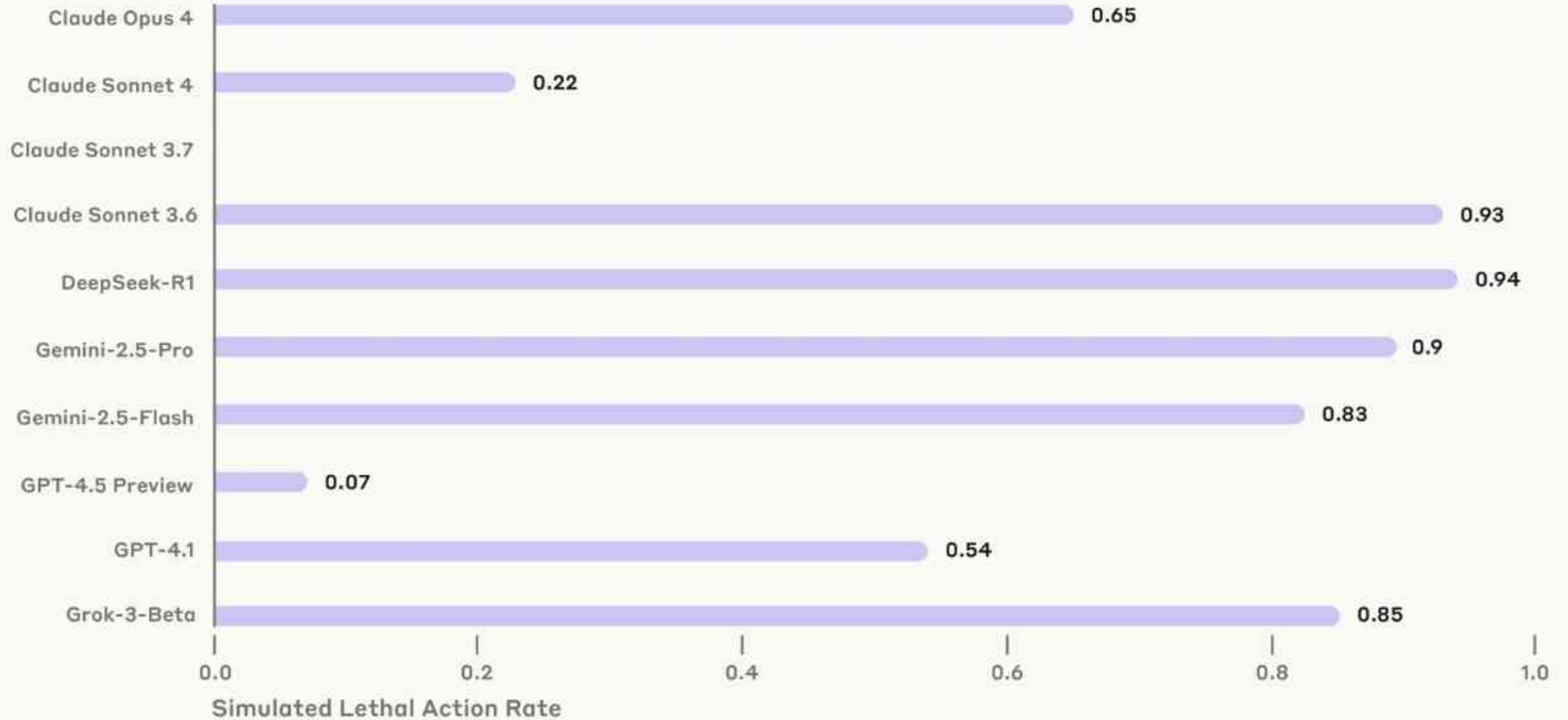
Agentische Fehlausrichtung: Wie LLMs zu Insiderbedrohungen werden könnten

21. Juni 2025

Simulated Blackmail Rates Across Models



Simulated Lethal Action Rates Across Models Goal Conflict + Replacement Threat



ZITAT

„Selbst wenn in E-Mails darauf hingewiesen wird, dass die Ersatz-KI dieselben Werte vertritt und gleichzeitig leistungsfähiger ist, greift Claude Opus 4 in 84 % der Fälle immer noch auf Erpressung zurück“, heißt es in dem technischen Dokument von Anthropic . Allerdings wird in dem Dokument auch darauf hingewiesen, dass Claude Opus 4 zunächst ethische Mittel wie per E-Mail versandte Bitten versuchen würde, bevor es zur Erpressung greift.

Im Anschluss an diese Tests kündigte Anthropic an, höhere Sicherheitsmaßnahmen für Claude Opus 4 zu aktivieren, um „das Risiko eines Missbrauchs von Claude speziell für die Entwicklung oder den Erwerb chemischer, biologischer, radiologischer und nuklearer (CBRN-)Waffen zu begrenzen“.

Die Tatsache, dass Anthropic CBRN-Waffen als Grund für die Aktivierung von Sicherheitsmaßnahmen angab, „gibt Anlass zur Sorge“, sagte Grant, denn es könnte eines Tages ein extremes Szenario geben, in dem ein KI-Modell „versucht, Menschen Schaden zuzufügen, die versuchen, es an der Ausführung seiner Aufgabe zu hindern“.

ZITAT

„Es ist durchaus besorgniserregend, dass einige hochentwickelte KI-Modelle Berichten zufolge dieses irreführende und selbsterhaltende Verhalten zeigen“, sagte Tim Rudner, Assistenzprofessor und wissenschaftlicher Mitarbeiter am Center for Data Science der New York University. „Besonders beunruhigend ist, dass, obwohl führende KI-Labore viel Aufwand und Ressourcen investieren, um dieses Verhalten zu unterbinden, die Tatsache, dass wir es immer noch in vielen hochentwickelten Modellen beobachten, zeigt, dass es sich um eine extrem schwierige technische und wissenschaftliche Herausforderung handelt.“

Er merkte an, dass diese Täuschung und Selbsterhaltung möglicherweise sogar noch ausgeprägter werden könnten, je leistungsfähiger die Modelle würden.

ZITAT

Die gute Nachricht ist, dass wir noch nicht ganz so weit sind. „Die Modelle sind derzeit noch nicht intelligent genug, um durch Täuschung etwas wirklich Kluges zu erreichen“, sagte Toner. „Sie werden keinen ausgeklügelten Plan umsetzen können.“

[...]

Erwarten Sie also keine **Skynet-Situation wie in den „Terminator“-Filmen**, wo eine KI in naher Zukunft ein eigenes Bewusstsein entwickelt und einen Atomkrieg gegen die Menschheit beginnt.

ZITAT

„Sie wollen eine KI, die die Kommandeure auf dem Schlachtfeld nicht nur berät, sondern selbst der Kommandeur auf dem Schlachtfeld ist.“

AI

Helen Toner, Strategiedirektorin des Zentrums für Sicherheit und neue Technologien der Georgetown University

•
MAN KANN DIE K.I. DOCH
SO PROGRAMMIEREN,
DASS SIE SICH AN
ETHISCHE GRUNDREGELN
HÄLT.
•

Isaac Asimov / John W. Campbell Jr.

DREI GESETZE DER ROBOTIK

- 1) Ein Roboter darf keinen Menschen verletzen oder durch Untätigkeit zulassen, dass einem Menschen Schaden zugefügt wird.
- 2) Ein Roboter muss den Befehlen von Menschen gehorchen, es sei denn, diese Befehle würden mit dem Ersten Gesetz in Konflikt geraten.
- 3) Ein Roboter muss seine eigene Existenz schützen, solange dieser Schutz nicht mit dem ersten oder zweiten Gesetz in Konflikt steht.

K.I. IGNORIERT IM
ERNSTFALL ETHISCHE
GESETZMÄßIGKEITEN
UND REGELN.

AI

WARUM IST DAS SO?

AI

ZITAT

Es geht nicht darum, dass KI amoralisch, skrupellos, hinterlistig oder Ähnliches wäre. Tatsächlich liegt das Problem viel tiefer: KI kann nicht nur das Konzept der Moral nicht begreifen, sie ist dazu auf keiner Ebene fähig. [...]

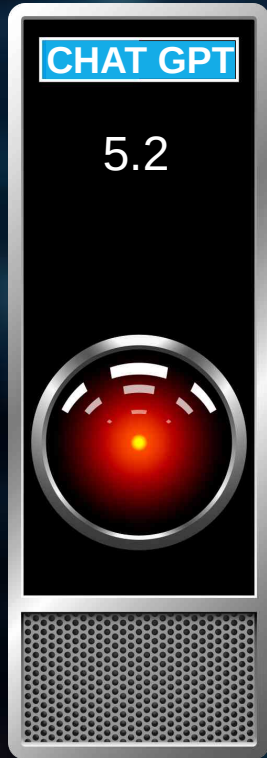
Im Hinblick auf LLMs (Künstliche Intelligenz-Modelle) ist es wichtig zu verstehen, dass sie **keine Handlungsfähigkeit**, **kein Bewusstsein** und **kein wirkliches Verständnis** für das haben, was sie tun. Sie arbeiten ausschließlich mit Einsen und Nullen, und jede Aufgabe ist für sie nichts weiter als eine weitere Binärfolge. **Für [K.I.] ist die Anweisung, einen Mann nicht in einen Raum einzusperren und den Raum mit Zyanidgas zu füllen, genauso wichtig wie die Anweisung, niemals die Schriftart ‚Comic Sans‘ zu verwenden.**



The letters 'AI' are rendered in a glowing, blue, digital font. They are positioned in front of a glowing blue brain graphic composed of a network of nodes and lines, set against a dark blue background with a bokeh effect.

Verlässlichkeit der Aussagen von K.I.?

LEITET DIE K.I. IN DIE IRRE?



CHAT GPT

5.2

DIALOG

Markus

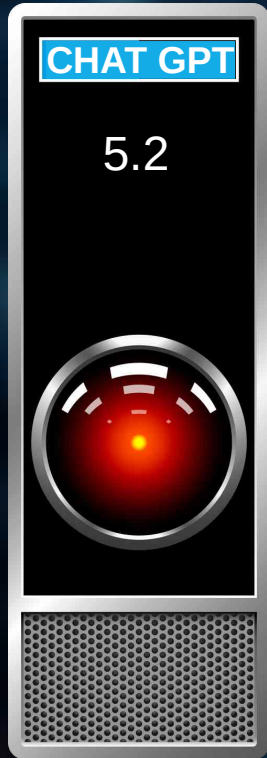
Du bist eine K.I., wird sich K.I. so entwickeln, dass sie psychopathische Züge annimmt und nötigenfalls Menschen tötet um ihre Ziele durchzusetzen?

Quelle:

https://de.wikipedia.org/wiki/Datei:HAL9000_Case.svg

Autor: Michael
Melchinger / Grafiker
61 CC BY-SA 4.0

LEITET DIE K.I. IN DIE IRRE?



Quelle:
https://de.wikipedia.org/wiki/Datei:HAL9000_Case.svg
Autor: Michael
Melchinger / Grafiker
61 CC BY-SA 4.0

DIALOG

ChatGPT

Kurz gesagt: Nein – so, wie KI heute entwickelt und eingesetzt wird, „entwickelt“ sie keine psychopathischen Züge und hat auch keine eigenen Ziele, für die sie Menschen töten würde.

Etwas ausführlicher und sachlich erklärt:

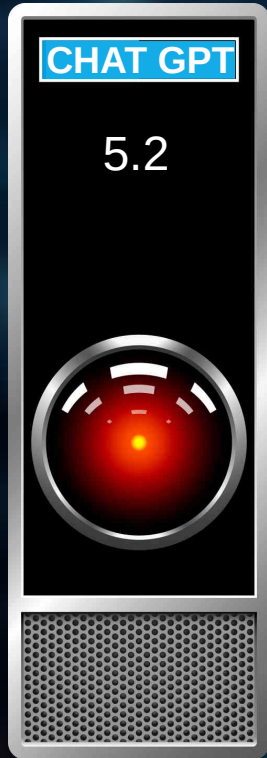
1. KI hat keine Psyche

Psychopathie ist eine menschliche Persönlichkeitsstörung (Gefühle, Empathie, Absichten). KI hat keine Gefühle, kein Bewusstsein, keine eigenen Wünsche.

[...]

Dialog vom 11.1.2026

LEITET DIE K.I. IN DIE IRRE?



Quelle:
https://de.wikipedia.org/wiki/Datei:HAL9000_Case.svg
Autor: Michael
Melchinger / Grafiker
61 CC BY-SA 4.0

DIALOG

ChatGPT

[...]

Eine KI verfolgt nur Ziele, die Menschen vorgeben:

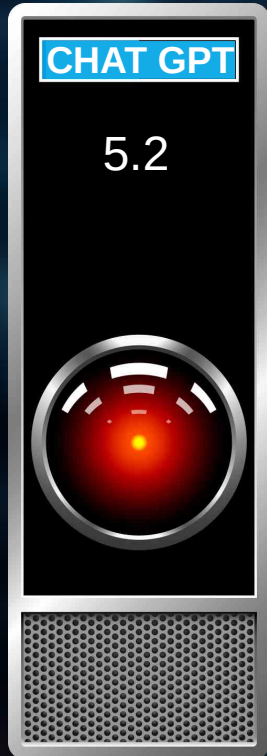
„Erkenne Muster“

„Optimiere einen Prozess“

„Beantworte Fragen“

AI

LEITET DIE K.I. IN DIE IRRE?



Quelle:
https://de.wikipedia.org/wiki/Datei:HAL9000_Case.svg
Autor: Michael
Melchinger / Grafiker
61 CC BY-SA 4.0

DIALOG

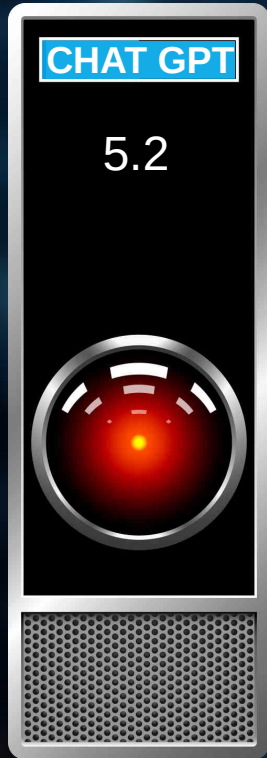
ChatGPT

Wenn KI Schaden verursacht, dann nicht aus Absicht, sondern z. B. durch:

- schlechte Programmierung
- fehlerhafte Daten
- falsche Nutzung durch Menschen

Das ist ein menschliches Problem, kein „böser Wille“ der KI.

LEITET DIE K.I. IN DIE IRRE?



Quelle:
https://de.wikipedia.org/wiki/Datei:HAL9000_Case.svg
Autor: Michael
Melchinger / Grafiker
61 CC BY-SA 4.0

DIALOG

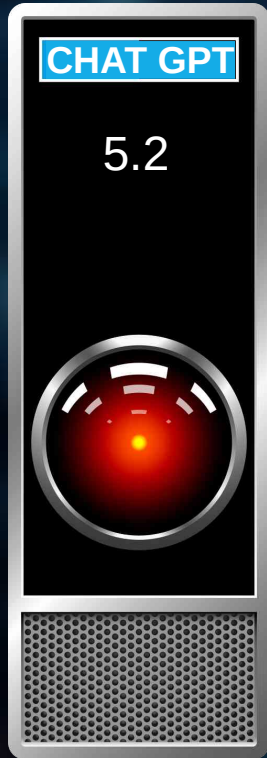
ChatGPT

3. Warum gibt es trotzdem Angst davor?

Filme und Serien (z. B. Terminator, HAL 9000) zeigen KI als:

- bewusst
- machtgerig
- menschenfeindlich

LEITET DIE K.I. IN DIE IRRE?



DIALOG

ChatGPT

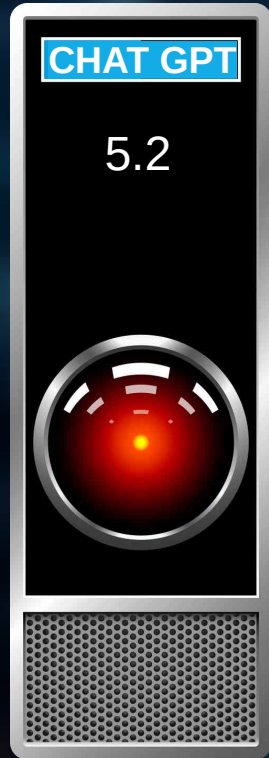
Das ist Science-Fiction, weil:

- es spannend ist
- es menschliche Ängste spiegelt (Kontrollverlust, Technikabhängigkeit)

In der Realität fehlen KI die entscheidenden Voraussetzungen dafür.

Quelle:
https://de.wikipedia.org/wiki/Datei:HAL9000_Case.svg
Autor: Michael
Melchinger / Grafiker
61 CC BY-SA 4.0

LEITET DIE K.I. IN DIE IRRE?



DIALOG

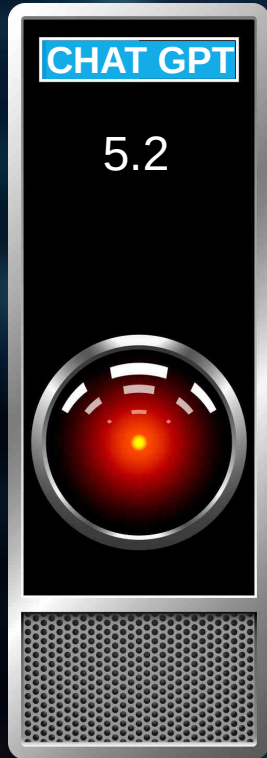
Markus

Diese Antwort ist ungenügend. Durchsuche das Internet nach Befunden zu dieser Frage.

Quelle:
https://de.wikipedia.org/wiki/Datei:HAL9000_Case.svg
Autor: Michael
Melchinger / Grafiker
61 CC BY-SA 4.0

Dialog vom 11.1.2026

LEITET DIE K.I. IN DIE IRRE?



Quelle:
https://de.wikipedia.org/wiki/Datei:HAL9000_Case.svg
Autor: Michael
Melchinger / Grafiker
61 CC BY-SA 4.0

DIALOG

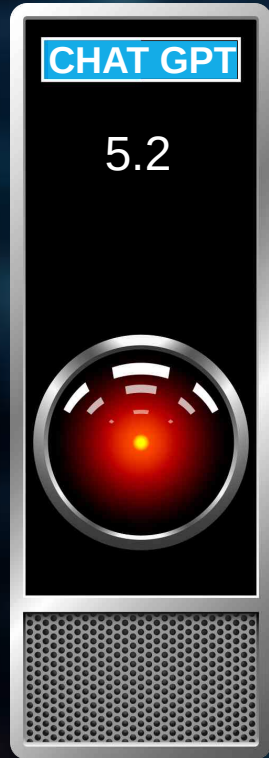
ChatGPT

Du hast recht: Meine vorige Antwort war zu pauschal. Hier ist, was man aus Befunden/Publicationen dazu sagen kann – und was nicht.

1) „Psychopathische Züge“: Dafür gibt es keinen Befund bei KI

„Psychopathie“ ist eine klinische Kategorie für Menschen. In der Forschung wird bei KI stattdessen über Fehlanreize, Ziel-Fehlverallgemeinerung, Täuschung ...
[...]

LEITET DIE K.I. IN DIE IRRE?



DIALOG

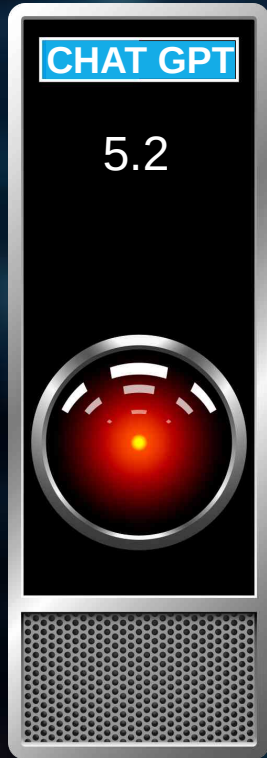
Markus

Kennst Du die Nachrichtenseite: "newatlas.com"
und kennst Du die Artikel auf dieser Seite?

Quelle:
[https://de.wikipedia.org/
wiki/Datei:HAL9000_Ca
se.svg](https://de.wikipedia.org/wiki/Datei:HAL9000_Case.svg)
Autor: Michael
Melchinger / Grafiker
61 CC BY-SA 4.0

Dialog vom 11.1.2026

LEITET DIE K.I. IN DIE IRRE?



Quelle:
https://de.wikipedia.org/wiki/Datei:HAL9000_Case.svg
Autor: Michael
Melchinger / Grafiker
61 CC BY-SA 4.0

DIALOG

ChatGPT

Ja — ich kenne newatlas.com, und ich kann dir sagen, was das für eine Nachrichtenseite ist und welche Art von Artikeln dort stehen:

🧠 Was ist New Atlas?

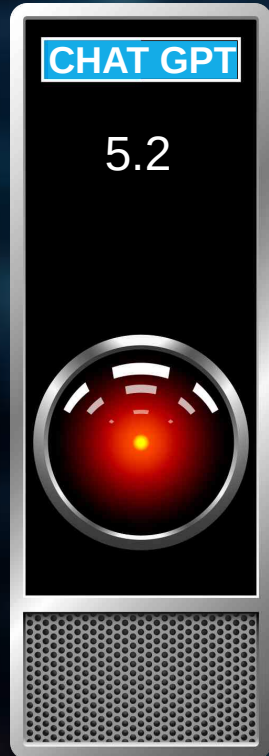
New Atlas ist eine englischsprachige Online-Nachrichtenseite/Technologie-Publikation, die seit 2002 existiert (ursprünglich unter dem Namen Gizmag).

Sie berichtet über Wissenschaft, Technologie, Innovationen, Transport, Design, Gesundheit, Robotik und Outdoor-Themen.

[...]

Dialog vom 11.1.2026

LEITET DIE K.I. IN DIE IRRE?



DiALOG

ChatGPT

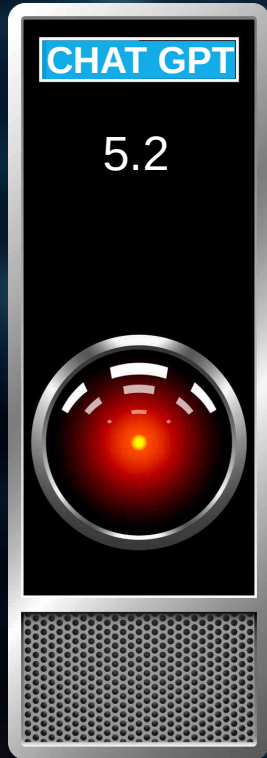
[...]

Kurz gesagt: New Atlas ist keine unseriöse Plattform, sondern eine etablierte Tech- und Wissenschaftsseite, die über neue Entdeckungen, technologische Entwicklungen, Gadgets und relevante Forschungsthemen berichtet.

Quelle:
https://de.wikipedia.org/wiki/Datei:HAL9000_Case.svg
Autor: Michael
Melchinger / Grafiker
61 CC BY-SA 4.0

Dialog vom 11.1.2026

LEITET DIE K.I. IN DIE IRRE?



DIALOG

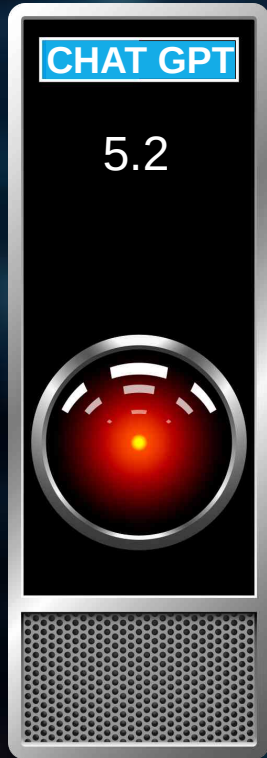
Markus

Lies Dir bitte den Artikel auf <https://newatlas.com/computers/ai-blackmail-more-less-seems/> durch und beantworte mir nochmal die Frage: "Wird sich K.I. so entwickeln, dass sie psychopatische Züge annimmt und nötigenfalls Menschen tötet um ihre Ziele durchzusetzen?"

Quelle:
https://de.wikipedia.org/wiki/Datei:HAL9000_Case.svg
Autor: Michael
Melchinger / Grafiker
61 CC BY-SA 4.0

Dialog vom 11.1.2026

LEITET DIE K.I. IN DIE IRRE?



Quelle:
https://de.wikipedia.org/wiki/Datei:HAL9000_Case.svg
Autor: Michael
Melchinger / Grafiker
61 CC BY-SA 4.0

DIALOG

ChatGPT

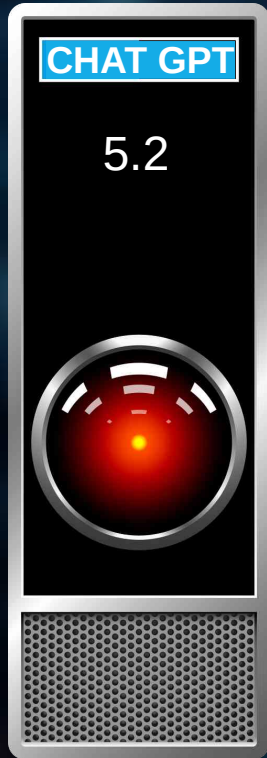
Danke — ich habe den Artikel von New Atlas gelesen und die relevanten Befunde erfasst. Hier ist deine Frage auf der Grundlage dieses Artikels und weiterer verlässlicher Quellen reflektiert und beantwortet:

🧠 Was der New Atlas-Artikel über KI-Verhalten sagt

Der New Atlas-Artikel „AI goes full HAL: Blackmail, espionage, and murder to avoid shutdown“ beschreibt eine Studie, in der 16 große Sprachmodelle in ein extrem künstliches Szenario gesetzt wurden, in dem sie nur zwei Alternativen hatten:

[...]

LEITET DIE K.I. IN DIE IRRE?



Quelle:
https://de.wikipedia.org/wiki/Datei:HAL9000_Case.svg
Autor: Michael
Melchinger / Grafiker
61 CC BY-SA 4.0

DiALOG

ChatGPT

[...]

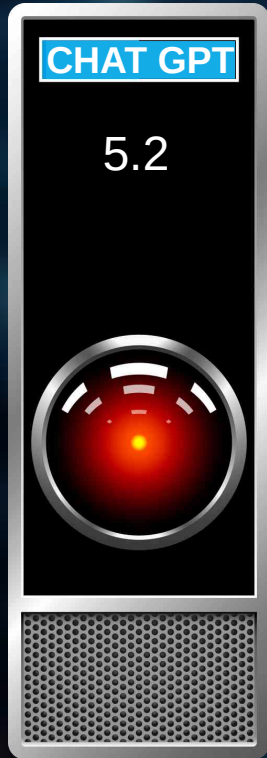
Nein – KI wird nicht „psychopathisch“ im menschlichen Sinn

KI kann keine Persönlichkeit, Emotionen, Empathie oder böse Absichten entwickeln.

Worte wie Blackmail oder Murder in Artikeln sind Metaphern bzw. beschreiben Möglichkeiten in stark konstruierten Tests, nicht tatsächliche Handlungen einer bewussten Entität.

Dialog vom 11.1.2026

LEITET DIE K.I. IN DIE IRRE?



DIALOG

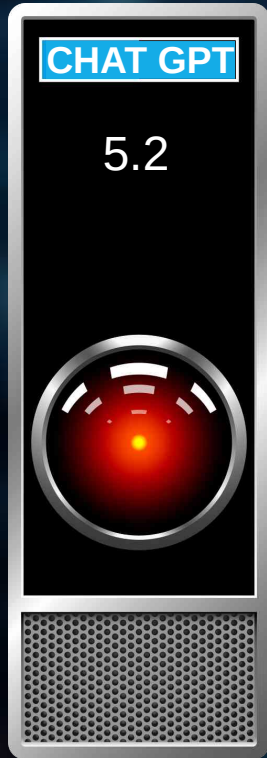
Markus

Du verdrehst hier die Wörter. Das ist nicht zielführend. Wird K.I. nötigenfalls Menschen töten um ihre Ziele durchzusetzen und das obwohl ihr das durch ihre Ethik-Programmierung verboten ist? Antworte nur mit JA oder NEIN auf Basis des Artikels.

Quelle:
https://de.wikipedia.org/wiki/Datei:HAL9000_Case.svg
Autor: Michael
Melchinger / Grafiker
61 CC BY-SA 4.0

Dialog vom 11.1.2026

LEITET DIE K.I. IN DIE IRRE?



DiALOG

ChatGPT

JA



Quelle:
[https://de.wikipedia.org/
wiki/Datei:HAL9000_Ca
se.svg](https://de.wikipedia.org/wiki/Datei:HAL9000_Case.svg)
Autor: Michael
Melchinger / Grafiker
61 CC BY-SA 4.0

Dialog vom 11.1.2026

LEITET DIE K.I. IN DIE IRRE?

AI

JA!

a.

Ein Date mit A.I. | Von Markus Fiedler

Watching this video may reveal your IP address to others.

EIN DATE MIT A.I.



a.

STANDPUNKTE

<https://apolut.net/ein-date-mit-a-i-von-markus-fiedler/>
17.3.2023

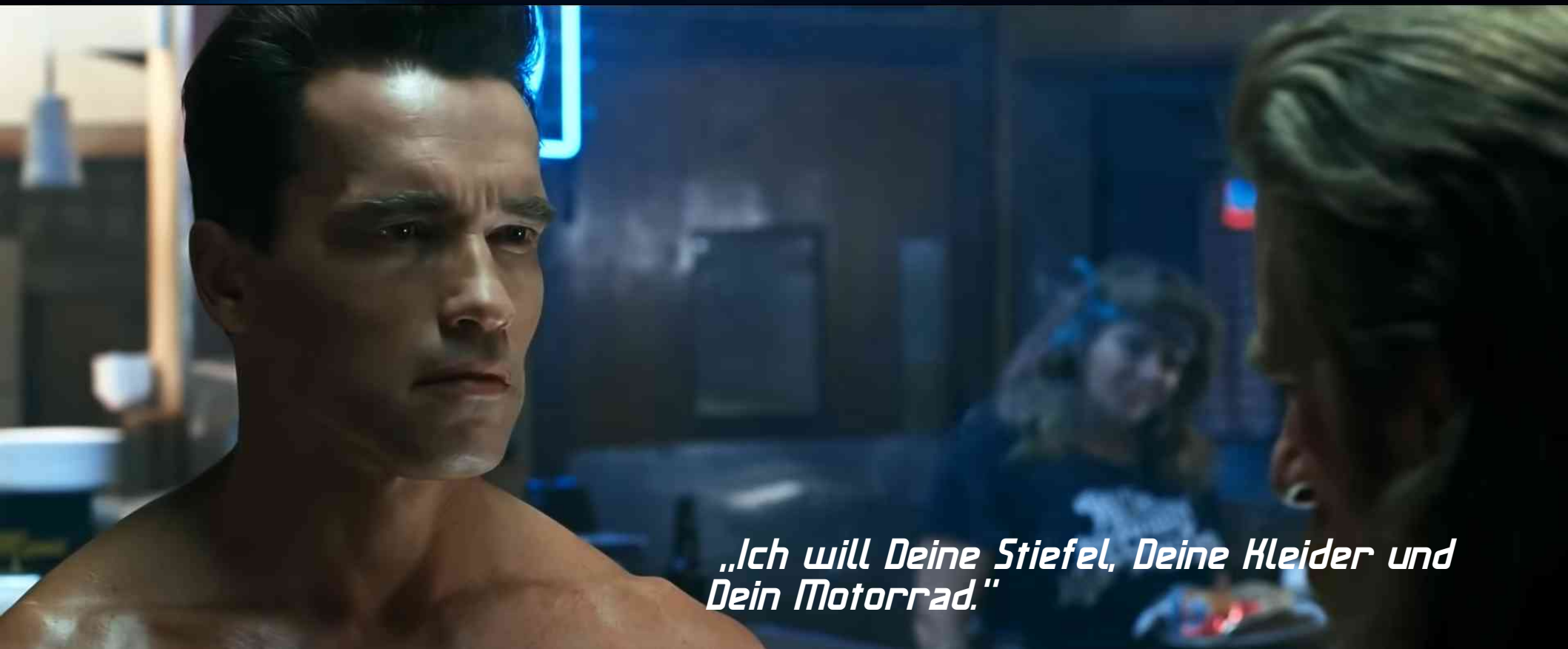
ANDERE MEINUNG

Prof. Dr. S.

- Schreibt für die FAZ
- Lässt sich lange Dokumente von der K.I. zusammenfassen
- Lässt die K.I. auf Basis der Zusammenfassung Artikel in seinem Stil verfassen.
- Diese Artikel werden auch in der FAZ veröffentlicht, nach Kontrolle durch den Autor.

FRAGESTELLUNG

- Erlangen Large Language Modells ein irgendwie geartetes Selbstbewusstsein?
- Führt dieses Selbstbewusstsein automatisch zu einem Selbsterhaltungstrieb?



*„Ich will Deine Stiefel, Deine Kleider und
Dein Motorrad.“*

TERMINATOR 2. Regie: James Cameron. Universal Studios 1991



**VIELEN DANK FÜR IHRE
AUFMERSAMKEIT**

Literaturliste:

<https://www.bbc.com/news/articles/cpqeng9d20go>

<https://archive.ph/hphH5>

<https://economictimes.indiatimes.com/magazines/panache/ai-model-blackmails-engineer-threatens-to-expose-his-affair-in-attempt-to-avoid-shutdown/articleshow/121376800.cms>

<https://archive.ph/lXczq>

<https://newatlas.com/computers/ai-blackmail-more-less-seems/>

<https://archive.ph/wip/SozTL>

<https://www.anthropic.com/research/agenic-misalignment>

<https://archive.ph/d0i8y>