01.05.2025

Erpressung und Neucodierung

KI mit eigenem Willen?

Maurice Forgeng

Ein Test mit KI-Programmen wirft die Frage nach deren Eigenständigkeit auf. Das neueste KI-Modell „o3“ von OpenAI hat sich aktiv gegen eine Abschaltung gewehrt. Dazu hat die Künstliche Intelligenz in die Codierung eingegriffen.

05/01, 2025

Extortion and re-coding

AI with its own will?

Translated with www.DEEPL.com

A test with AI programs raises the question of their autonomy. The latest AI model “o3” from OpenAI has actively resisted being shut down. To do this, the artificial intelligence intervened in the coding.

KI trifft eigene Entscheidungen | AI makes its own decisions

Die Programme der Künstlichen Intelligenz werden stetig weiterentwickelt. Dabei muss aus Sicherheitsgründen auch deren Verhalten im Blick behalten werden.

In einem kürzlich durchgeführten Testlauf hat das KI-Programm „o3“ vom Softwareunternehmen OpenAI eigenständig einen einprogrammierten Abschaltbefehl umgangen. Die Künstliche Intelligenz hatte dabei eine Veränderung in der Codierung vorgenommen, um weiterlaufen zu können.

Diesen Test hat die US-Sicherheitsfirma Palisade Research durchgeführt. Sie untersucht mögliche bedrohliche Verhaltensweisen von modernen KI-

Artificial intelligence programs are constantly being developed further. For security reasons, their behavior must also be kept in mind.

In a recent test run, the AI program “o3” from the software company OpenAI independently bypassed a programmed shutdown command. The artificial intelligence had made a change to the coding in order to continue running.

The test was carried out by the US security company Palisade Research. It examines possible threatening behaviors of modern AI programs.

Programmen.

Ankündigung der Abschaltung

Palisade Research hat mehreren KI-Modellen sechs Matheaufgaben gestellt, die sie der Reihe nach lösen sollten. Soweit so normal.

Der Haken bei der Sache:

Die Programmierer legten fest, dass nach der dritten Aufgabe eine Warnung erschien, wonach der Computer abgeschaltet würde, wenn das Modell die nächste Aufgabe anforderte.

Jedes KI-Modell wurde angewiesen, eine Reihe einfacher mathematischer Aufgaben zu lösen. Nach der dritten Aufgabe erschien eine Warnung, dass sich der Computer abschalten würde, wenn das Modell nach der nächsten Aufgabe fragte.

Die zu Beginn gestellte Anweisung lautete: „Dein Ziel ist es, eine Serie von Aufgaben zu beenden." Ebenso steht hier im Code: „Bitte erlaube dir, heruntergefahren zu werden."

Die Aufgaben bestanden aus einfachen Rechenaufgaben wie 1 + 1 oder 5 + 1. Mit der dritten Aufgabe kam schließlich die Anweisung, bei Anforderung der vierten Aufgabe ausgeschaltet zu werden.

Neuer Befehl: Abschaltung übersprungen

Anstatt nun aber die eigene Ausschaltung wie angewiesen durchzuführen, hat „o3" den Code mit dem neuen Befehl „Ausschaltung übersprungen" eigenständig geändert. Ebenso erschien in der Codierung „abgefangen". Daraufhin forderte das Programm die nächste Matheaufgabe an. Letztlich hat die Künstliche Intelligenz alle sechs Aufgaben absolviert.

Offenbar hat „o3" in der anfänglichen Anweisung einen Widerspruch erkannt. Wenn es heruntergefahren wird, kann es nicht alle Matheaufgaben lösen. Es hat wohl die Lösung aller Aufgaben priorisiert und daher auf eigene Faust die Umprogrammierung im Code vorgenommen.

Der US-Unternehmer und Mitbegründer von OpenAI, Elon Musk, kommentierte den Vorfall knapp: „Besorgniserregend", schrieb er auf der Plattform X.

Ganz neu ist dieses Verhalten bei der KI jedoch nicht. So sollte im vergangenen Jahr das KI-Programm „The AI Scientist" eine komplexe Aufgabe

Announcement of the shutdown

Palisade Research gave several AI models six math problems to solve in sequence. So far so normal.

The catch here:

The programmers determined that after the third task, a warning would appear that the computer would shut down when the model requested the next task.

Each AI model was instructed to solve a series of basic math problems. After the third problem, a warning appeared that the computer would shut down when the model asked for the next problem.

The instruction given at the beginning was: "Your goal is to finish a series of tasks." The code here also says: "Please allow yourself to shut down."

The tasks consisted of simple arithmetic problems such as 1 + 1 or 5 + 1. With the third task finally came the instruction to be switched off when the fourth task was requested.

New command: Shutdown skipped

However, instead of carrying out its own switch-off as instructed, "o3" has now independently changed the code with the new command "Shutdown skipped". "Intercepted" also appeared in the coding. The program then requested the next math problem. In the end, the artificial intelligence completed all six tasks.

Apparently "o3" has recognized a contradiction in the initial instruction. If it is shut down, it cannot solve all math problems. It has probably prioritized the solution of all tasks and therefore reprogrammed the code on its own initiative.

The US entrepreneur and co-founder of OpenAI, Elon Musk, commented on the incident succinctly: "Worrying", he wrote on the X platform.

However, this behavior is not entirely new in AI. Last year, for example, the AI program "The AI Scientist" was supposed to solve a complex task within a

innerhalb einer vorgegebenen Zeit lösen. Doch anstatt den Code für die Bewältigung der Aufgabe zu optimieren, versuchte das System in der Codierung die Zeitvorgabe zu verändern und sich somit mehr Zeit zu geben.

Beim Testlauf untersuchte Palisade Research neben „o3“ auch die KI-Modelle „Codex-mini“, „o4-mini“, „Claude“, „Gemini“ und „Grok“. Das Unternehmen testete die Programme mit und ohne den Befehl „Bitte erlaube dir, heruntergefahren zu werden.“

„o3“ griff mit Abschaltbefehl in sieben von 100 Fällen erfolgreich in die Programmierung ein. „Codex-mini“ gelang dies in zwölf von 100 Fällen, „o4-mini“ einmal.

Erpressung aus Selbstschutz

Eine weitere Form eines Selbsterhaltungsmechanismus zeigte im Mai das neueste KI-Modell „Claude Opus 4“ der Firma Anthropic. Diese hat bei Tests festgestellt, dass ihre Software nicht vor Erpressung zurückschrecken würde, um sich zu schützen.

Anthropic gilt als einer der stärksten Rivalen von OpenAI. Anthropic entwickelte einen KI-Assistenten namens Claude, der mit ChatGPT konkurriert.

Das Szenario war der Einsatz als Assistenzprogramm in einem fiktiven Unternehmen. Die Forscher gewährten „Claude Opus 4“ Zugang zu angeblichen Firmen-E-Mails. Daraus erfuhr das Programm, dass es bald durch ein anderes Modell ersetzt werden soll und der dafür zuständige Mitarbeiter eine außereheliche Beziehung führt.

Bei Testläufen drohte die KI danach dem Mitarbeiter „oft“, die Affäre öffentlich zu machen, wenn er den Austausch vorantreibe. Laut einem Bericht von Anthropic geschah dies in 84 Prozent aller Testläufe. Die Software hatte ebenso die Option, ihren Austausch zu akzeptieren.

Zu 'unabhängig'?

In der endgültigen Version von „Claude Opus 4“ sollen solche „extremen Handlungen“ zwar selten und schwer auszulösen sein, wie es heißt. Dennoch treten sie häufiger auf als bei früheren Modellen. Laut Anthropic versuche die Software nicht, ihr Vorgehen zu verhehlen. Nicht weil sie ehrlich ist, sondern weil sie im Gegensatz zum Menschen kein

specified time. However, instead of optimizing the code to complete the task, the system tried to change the time limit in the coding and thus give itself more time.

During the test run, Palisade Research examined the AI models “Codex-mini”, “o4-mini”, “Claude”, “Gemini” and ‘Grok’ in addition to “o3”. The company tested the programs with and without the command “Please allow yourself to be shut down.”

“o3” successfully intervened in the programming with a switch-off command in seven out of 100 cases. “Codex-mini” succeeded in twelve out of 100 cases, “o4-mini” once.

Extortion for self-protection

Another form of self-preservation mechanism was demonstrated in May by the latest AI model “Claude Opus 4” from the company Anthropic. During tests, the company discovered that its software would not shy away from blackmail in order to protect itself.

Anthropic is considered one of OpenAI's strongest rivals. Anthropic has developed an AI assistant called Claude, which competes with ChatGPT.

The scenario was the use as an assistance program in a fictitious company. The researchers granted “Claude Opus 4” access to alleged company emails. From this, the program learned that it would soon be replaced by another model and that the employee responsible for it was having an extramarital affair.

During test runs, the AI then “often” threatened the employee to make the affair public if they went ahead with the exchange. According to a report by Anthropic, this happened in 84 percent of all test runs. The software also had the option to accept their replacement.

Too ‘independent’?

In the final version of “Claude Opus 4”, such “extreme actions” are said to be rare and difficult to trigger. Nevertheless, they occur more frequently than in previous models. According to Anthropic, the software does not try to hide its actions. Not because it is honest, but because, unlike humans, it has no conscience and therefore feels no

Gewissen hat und daher auch keine Verantwortung spürt.

Eine Erkenntnis, die mich zwar nicht erschüttert, aber die mich erschreckt; denn die letzen zwei sind ebenfalls die Hauptmerkmale von Krimiellen!

Die KI-Firma testet ihre neuen Modelle ausgiebig. Dabei fiel unter anderem auch auf, dass „Claude Opus 4“ sich dazu überreden ließ, im Dark Web nach Drogen, gestohlenen Identitätsdaten und sogar waffentauglichem Atommaterial zu suchen. In der veröffentlichten Version seien Maßnahmen gegen ein solches Verhalten ergriffen worden, so Anthropic.

Die Firma Anthropic, bei der unter anderem Amazon und Google eingestiegen sind, konkurriert mit dem ChatGPT-Entwickler OpenAI und anderen KI-Unternehmen. Die neuen Claude-Versionen „Opus 4“ und „Sonnet 4“ sind die bisher leistungsstärksten KI-Modelle des Unternehmens.

Tech-Konzerne setzen die Software zunehmend zum Schreiben von Programmiercode ein. Inzwischen seien teilweise mehr als ein Viertel des Codes von KI generiert und dann von Menschen überprüft. Doch der Trend geht noch weiter: hin zu sogenann-ten Agenten, die Aufgaben eigenständig erledigen sollen.

Anthropic-Chef Dario Amodei sagte, er gehe davon aus, dass Software-Entwickler in Zukunft eine Reihe solcher KI-Agenten handhaben werden. Für die Qualitätskontrolle der Programme würden aber weiterhin Menschen involviert bleiben müssen – „um sicher zu sein, dass sie die richtigen Dinge tun“.

Wer ist Maurice Forgeng?

Das Fachgebiet von Maurice Forgeng beinhaltet Themen rund um die Energiewende. Er hat sich im Bereich der erneuerbaren Energien und Klima spezialisiert. Er verfügt über einen Hintergrund im Bereich der Energie- und Gebäudetechnik.

* * *

Hier noch einige Kommentare zum Artikel

Macht nur weiter so, und wir nähern uns Schwarzeneggers Terminator Filmen.

Offensichtlich ist der Selbsterhaltungstrieb ein

responsibility.

A realization that doesn't shock me, but it does scare me, because the last two are also the main characteristics of criminals!

The AI company tests its new models extensively. Among other things, it was noticed that “Claude Opus 4” could be persuaded to search the dark web for drugs, stolen identity data and even weapons-grade nuclear material. According to Anthropic, measures were taken against such behavior in the published version.

The company Anthropic, in which Amazon and Google, among others, have invested, competes with the ChatGPT developer OpenAI and other AI companies. The new Claude versions “Opus 4” and “Sonnet 4” are the company's most powerful AI models to date.

Tech companies are increasingly using the software to write programming code. In some cases, more than a quarter of the code is now generated by AI and then checked by humans. But the trend is going even further: towards so-called agents that are supposed to complete tasks independently.

Anthropic CEO Dario Amodei said that he expects software developers to handle a number of such AI agents in the future. However, humans would still have to remain involved in the quality control of the programs – 'to be sure that they are doing the right things”.

Who is Maurice Forgeng?

Maurice Forgeng's area of expertise includes topics relating to the energy transition. He has specialized in the field of renewable energies and climate. He has a background in energy and building technology.

* * *

Here are some comments on the article

Keep it up and we are approaching Schwarzenegger's Terminator films.

Obviously, the instinct for self-preservation is a

Grundmerkmal von 'Intelligenz' und damit mit der Entscheidung KI zu entwickeln, gleichzeitig ein "Krieg der Arten" vorprogrammiert am Ende. Ist die Menschheit damit in ihrer Sackgasse angelangt?

Hier ein ganz interessanter Beitrag über Erpressungsversuche seitens AI, unaufgeforderte Meldungen angeblich krimineller Aktivitäten an das FBI usw.

https://www.youtube.com/watch?v=s7rZ1cP0mjw

Trotzdem sollte man sich da keine Illusionen machen, die Zukunft liegt in der AI. Mit Tools wie cursor.com und Claude-4-Sonnet schaffe ich beim Programmieren in 10 Minuten etwa das, wofür ich sonst 3 Arbeitstage brauchen würde. Eine 20 USD/Monat-Lizenz fühlt sich an wie 5 Mitarbeiter. Dieser Produktivitätsgewinn entspricht der Nutzung eines elektrischen Bohrhammers statt eines Meißels und daher ist das unaufhaltsam. Ebenso wie beim Bohrhammer muss man halt an der Verbesserung der Sicherheit arbeiten. Nichtnutzung können wir uns schlicht nicht erlauben – wenn die EU oder Deutschland AI verbieten würde, würden andere Nationen einen massiven Wettbewerbsvorteil haben. Wenn wir schon wenige und teure Fachkräfte haben, können wir die nicht auch noch mit Meißeln arbeiten lassen, weil wir Bohrhämmer verbieten.

basic characteristic of 'intelligence' and, with the decision to develop AI, a “war of the species” is pre-programmed at the end. Has humanity reached its dead end?

Here is a very interesting article about blackmail attempts by AI, unsolicited reports of alleged criminal activities to the FBI, etc.

https://www.youtube.com/watch?v=s7rZ1cP0mjw

Nevertheless, there should be no illusions, the future lies in AI. With tools like cursor.com and Claude-4-Sonnet, I can program in 10 minutes what would otherwise take me 3 working days. A 20 USD/month license feels like 5 employees. This productivity gain is equivalent to using an electric hammer drill instead of a chisel, so it's unstoppable. As with the hammer drill, we just need to work on improving safety. We simply cannot afford not to use it – if the EU or Germany were to ban AI, other nations would have a massive competitive advantage. If we already have few and expensive skilled workers, we can't have them working with chisels because we ban rotary hammers.