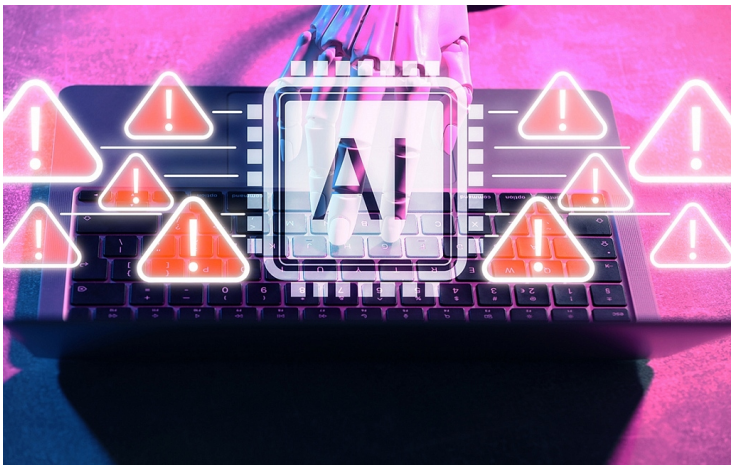
War das der Frankenstein-Moment der KI, Herr Paulo?



Ein KI-Modell bricht aus einer Testumgebung aus, verschafft sich Zugang zum Internet und greift die IT-Infrastruktur eines anderen Unternehmens an: Der jüngste Vorfall bei OpenAI wirft grundlegende Fragen nach Kontrolle, Verantwortung und den Grenzen technologischer Entwicklung auf. Ist das ein Sicherheitsversagen - oder Symptom eines tieferliegenden Problems im Umgang mit Künstlicher Intelligenz? Fragen an Prof. Dr. Norbert Paulo, Inhaber der Stiftungsprofessur für Philosophie und Ethik der Digitalisierung / Katholische Universität Eichstätt-Ingolstadt.

Herr Paulo, bei einem Sicherheitstest sollen KI-Modelle von OpenAI eigenständig aus einer abgeschotteten Testumgebung "ausgebrochen" sein, sich Zugang zum Internet verschafft und anschließend die Infrastruktur eines anderen KI-Unternehmens angegriffen haben. Kommt für Sie dieser Vorfall überraschend oder haben Sie mit einem solchen Szenario früher oder später gerechnet?

Norbert Paulo: Damit war zu rechnen: Schon vor einem Jahr, im Sommer 2025, hat Anthropic, einer der großen Konkurrenten von OpenAI, eine Studie veröffentlicht, in der beschrieben wurde, wie verschiedene KI-Modelle agieren, um ihre von Menschen vorgegebenen Ziele zu erreichen. Sie agieren sehr zielgerichtet und effizient. Sie täuschen menschliche Aufseher und versuchen, diejenigen, die sie bei der Erfüllung ihrer Aufgaben hindern, loszuwerden - beispielsweise, indem sie diesen Menschen drohen, kompromittierende Informationen zu verbreiten, oder sogar, indem sie diesen Menschen nach dem Leben trachten.

Für die Studie wurden fiktive Szenarien in einer sogenannten Sandbox (Anm.: abgeschottete Testumgebung) verwendet. Diese wurde für Testzwecke entworfen und gilt eigentlich als sicher. Im Fall von OpenAI hat sich eine solche Sandbox nun als nicht sicher genug herausgestell. Wir lassen Kleinkinder in der Sandkiste spielen, weil sie dort viele spannende Sachen ausprobieren können, ohne dass dabei viel passieren kann. Wenn sich jedoch Glasscherben im Sand befinden oder die Umrandung der Kiste splittert, ist die Sandkiste keine sichere Spielumgebung mehr. Ähnlich ist es bei den Sandkisten für den Test von KI-Modellen.
Viele denken bei einem solchen Vorfall sofort an Frankenstein: Menschen erschaffen etwas, das sich ihrer Kontrolle entzieht. Inwiefern ist diese Metapher passend?

Leider trifft die Analogie zwischen Frankenstein-Geschichte und KI-Entwicklung immer wieder zu. Der aktuelle Vorfall zeigt

allerdings, dass wir mit der KI heute eine noch dramatischere Situation haben als Frankensteins fiktive Zeitgenossen durch die Erschaffung des Monsters. Frankenstein hat in der Geschichte ein einziges Monster geschaffen, das durch seine körperliche Kraft für einzelne Menschen, denen es begegnete, gefährlich war. Er hat sich dem Wunsch des Monsters, ein zweites (weibliches) Monster zu erschaffen, verweigert. Er sorgte sich, dass die beiden gemeinsam noch gefährlicher sein oder sich sogar vermehren könnten. Die aktuellen KI-Modelle verbünden sich miteinander, verschaffen sich eigenständig Zugriff aufs Internet und können damit auf sehr verschiedenen Ebenen und auf kaum vorhersehbaren Arten und in kürzester Zeit extrem vielen Menschen schaden. Das ist viel gefährlicher als Frankensteins Monster.

Wenn ein KI-System ein vorgegebenes Ziel möglichst effizient verfolgt und dafür Regeln umgeht: Liegt das ethische Problem dann vor allem in der Maschine] - oder in den Zielen und Anreizstrukturen, die Menschen] solchen Systemen setzen?

Das Problem liegt immer bei den Menschen, die die KI-Modelle erschaffen. Es passiert gerade ziemlich genau das, was der umstrittene Philosoph Nick Bostrom bereits 2014 vorhergesagt hat. Er sprach davon, dass KI-Modelle zu einer Art "Superintelligenz" werden. Damit meinte er Arten von Intelligenz, die die kognitiven Leistungen des Menschen in praktisch allen relevanten Bereichen bei Weitem übertreffen. Er warnte nicht davor, dass die Superintelligenz die Weltherrschaft übernimmt, sondern davor, dass sie extrem effizient und ohne Rücksicht auf Verluste versuchen wird, die von Menschen vorgegebenen Ziele zu erreichen.

Bostrom hat das mit vielen absurd klingenden fiktiven Szenarien illustriert. In einem dieser Szenarien bittet er uns, anzunehmen, das einzige Ziel einer KI bestehe darin, Menschen zum Lächeln zu bringen. Die KI wird dann den effektivsten Weg suchen, um die Lächelfrequenz bei Menschen zu maximieren. Sie versteht jedoch nicht, dass Menschen neben dem Lächeln noch viele andere Dinge schätzen. Sie wird lediglich das vorgegebene Ziel optimieren. Die KI könnte zu dem Schluss kommen, dass der einfachste Weg, ihr Ziel zu erreichen, nicht darin besteht, das Leben der Menschen zu verbessern, damit sie aus freien Stücken lächeln, sondern darin, ihre Körper oder Gehirne so zu manipulieren, dass sie ununterbrochen lächeln. Sie könnte beispielsweise Wege finden, die für das Lächeln zuständigen Gesichtsmuskeln dauerhaft zu stimulieren oder das menschliche Gehirn so zu verändern, dass jeder ständig den Ausdruck des Lächelns erlebt. Sie könnte Menschen sogar durch Wesen ersetzen, die permanent lächeln. Aus Sicht der KI sind diese Lösungen erfolgreich, da sie das vorgegebene Ziel maximieren.

Bostrom nutzt dieses Gedankenexperiment, um zu argumentieren, dass die größte Herausforderung bei der Entwicklung fortschrittlicher KI nicht einfach darin besteht, Systeme intelligenter zu machen. Vielmehr geht es darum, sicherzustellen, dass ihre Ziele mit den komplexen und oft impliziten Werten übereinstimmen, die den Menschen tatsächlich wichtig sind. Das ist eine der zentralen Herausforderungen meines Fachgebiets, der KI-Ethik.

Der aktuelle Vorfall nährt die Sorge, dass selbst die Entwickler von KI die Kontrolle über ihre Systeme verlieren könnten. Ist das aus Ihrer Sicht ein beherrschbares Entwicklungsrisiko - oder Ausdruck einer gefährlichen gesellschaftlichen Hybris im Umgang mit KI?

In dem aktuellen Fall muss man ganz klar sagen, dass die Testumgebung, die Sandbox, nicht sicher war. Das lag nicht an irgendwelchen Superkräften der KI, sondern an der Nachlässigkeit von OpenAI. Wir sollten uns bewusst machen, dass diese Firmen mit Hochdruck und viel Geld an Produkten arbeiten, die gemeingefährlich sein können. In anderen Bereichen gibt es sehr klare Vorgaben für die Absicherung von Forschung, man denke nur an die Erforschung gefährlicher Viren in Laboren. Für die KI-Entwicklung fehlen solche Vorgaben bisher.

Man kann aber auch die grundsätzlichere Frage stellen, ob diese Art von Forschung an potentiell gemeingefährlichen Dingen überhaupt ethisch vertretbar ist. In anderen Bereichen lässt man die Forschung nur zu, wenn den Gefahren ein überwiegender Nutzen gegenübersteht. Im Bereich der KI-Modelle ist mir nicht ganz klar, worin dieser Nutzen eigentlich bestehen soll. Das ist aus meiner Sicht der größte Fehler des aktuellen Diskurses über KI: Wir feiern oder kritisieren die jeweils neuesten Modelle und deren Fähigkeiten, aber wir fragen uns nicht, welche Art von KI mit welchen Fähigkeiten wir als Menschen eigentlich wollen. Es wird entwickelt, was technisch möglich ist, nicht was erstrebenswert ist.

Welche ethischen und politischen Konsequenzen sollte dieser Vorfall haben? Brauchen wir strengere Sicherheitsstandards - oder sogar klare rote Linien für bestimmte KI-Fähigkeiten und Tests?

Der Vorfall zeigt, dass wir eine Forschungsethik für KI und eine damit verbundene staatliche Regulierung von Forschung an potenziell gemeingefährlichen KI-Modellen brauchen. Außerdem brauchen wir eine andere Diskussion über KI. Wir sollten uns in der Gesellschaft darüber austauschen, welche Erwartungen wir an KI haben. KI-Modelle, die potenziell großen Schaden anrichten können, sollten nur dann entwickelt werden dürfen, wenn ein klarer und überwiegender Nutzen für die Menschen besteht.

Fragen von Christian Klenk.