Künstliche Intelligenz

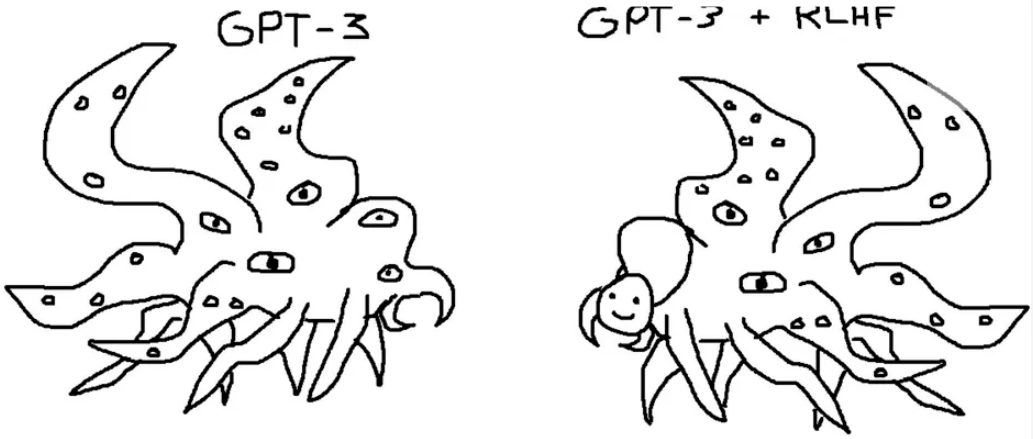
S+ Die Angst vor dem Monster hinter der Maske

Eine Kolumne von ***Christian Stöcker***

Die Debatte über künstliche Intelligenz wirkt derzeit einerseits geradezu hysterisch, andererseits bemerkenswert schlicht. Am meisten fürchten sich manche vor einer bösen KI, die lügen kann.

04.06.2023, 14.31 Uhr

Artikel zum Hören • 11 Min



Tentakelmonster »Shoggoth« aus der literarischen Fantasie von H.P. Lovecraft
Foto: TetraspaceWest / twitter

»Es macht einfach mehr Spaß, darüber nachzudenken, wie man eine freundliche KI herstellt, als über Branchenregulierung, genauso, wie es mehr Spaß macht, darüber nachzudenken, was man während der Zombieapokalypse machen würde, als darüber, wie man die globale Erwärmung aufhält.«
Der Schriftsteller Ted Chiang in einem lesenswerten Essay (2017)

Das Stichwort Alignment-Problem, für das es keine wirklich griffige deutsche Übersetzung gibt (»Übereinstimmungsproblem« ist eine Annäherung), bezeichnet folgenden Grundgedanken: Ein hinreichend mächtiges lernendes System, meist künstliche Intelligenz oder KI genannt, könnte möglicherweise eigene Ziele entwickeln. Oder die Ziele, die Menschen ihm vorgeben, falsch oder überinterpretieren. Mit katastrophalen Folgen.

ANZEIGE

Christian Stöcker ›

Foto: SPIEGEL ONLINE

Jahrgang 1973, ist Kognitionspsychologe und seit Herbst 2016 Professor an der Hochschule für Angewandte Wissenschaften Hamburg (HAW). Dort verantwortet er den Studiengang Digitale Kommunikation. Vorher leitete er das Ressort Netzwelt bei SPIEGEL ONLINE.

Einer der geistigen Väter der KI-Angst von heute, der in Oxford arbeitende Philosoph Nick Bostrom, hat zur Illustration das Büroklammern-Gedankenexperiment erdacht: Wenn man eine mächtige KI instruieren würde, möglichst viele Büroklammern herzustellen, könnte so ein Missverständnis doch dazu führen, dass die KI den ganzen Planeten in eine Büroklammerfabrik verwandele und dabei versehentlich die Menschheit auslösche.

ANZEIGE

Das Monster hinter der Maske

Das Alignment-Problem ist in künstlerischer Form häufig thematisiert worden, in Arthur C. Clarkes »2001« zum Beispiel, und in seiner berühmten Verfilmung von Stanley Kubrick. Tatsächlich ist Kubricks simple rote Leuchte im Chromkreis, das »Gesicht« des aus dem Ruder laufenden

Bordcomputers HAL, heute eine Art visuelle Chiffre für das Alignment-Problem. Mal ironisch, mal ernst gemeint.

Manche der KI-Ängstlichen von heute haben noch eine weitere Schicht Paranoia hinzugefügt: Sie entsetzt vor allem die Idee einer KI, die ganz andere Ziele als die Menschheit oder ihre eigenen Schöpfer hat, diese Tatsache aber absichtlich verschleiert. Auch dafür gibt es eine visuelle Chiffre im Netz: das sogenannte Shoggoth-Meme. In seiner ursprünglichen Variante handelt es sich um eine krude Zeichnung eines Tentakelmonsters. Ein »Shoggoth« ist ein schreckliches, gottähnliches, die Menschheit mit Nichtachtung strafendes Monster aus Schleim, Tentakeln und vielen Augen aus der literarischen Fantasie von H.P. Lovecraft.

Externer Inhalt

Zur Datenschutzerklärung

Das Tentakelmonster aus dem Shoggoth-Meme hat sich über eines seiner Tentakel eine Smiley-Maske gezogen, das Gesicht, mit dem es mit den Menschen interagiert. Die Menschheit, so könnte man das Meme übersetzen, wird sich von so einem Monster sehr leicht täuschen lassen. Was jenseits der Maske vor sich geht, werden wir nie verstehen. Wenn eine hyperintelligente Super-KI uns täuschen will, wird sie das auch tun.

Es gibt jetzt Shoggoth-Einkaufstaschen und KI-Modelle, die »Shoggy« heißen.

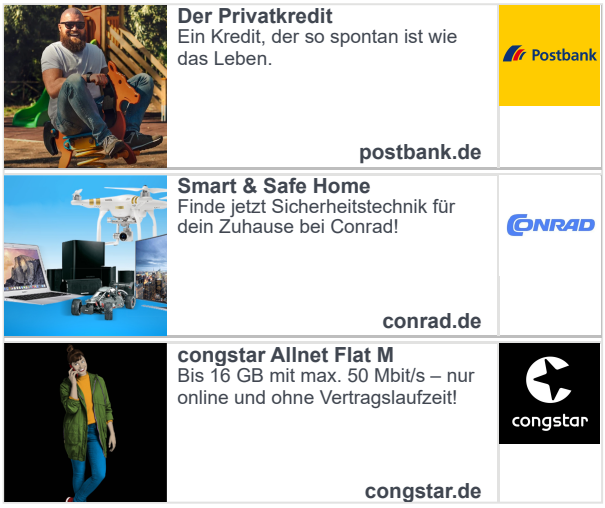
»Mach so was bitte nie wieder«

In der Originalversion ist das Monster mit der Smiley-Maske überschrieben mit »GPT-3 + RLHF«. GPT-3 ist die Vorgängerversion des Sprachmodells GPT-4, das heute unter der Haube von OpenAIs kostenpflichtiger ChatGPT-Version steckt. Die Buchstaben RLHF stehen für »reinforcement learning from human feedback«. Diese Methode soll sicherstellen, dass KI-Systeme keine allzu bösen Outputs produzieren: Menschliche Tester stellen dem System

Aufgaben, und wenn die Resultate in irgendeiner Form ethisch oder anderweitig inakzeptabel sind, gibt es entsprechendes Feedback: »Mach so was bitte nie wieder.«

Das Shoggoth-Meme ist gewissermaßen die Ultrakurzfassung des Gedankens, dass diese Art von Training Selbstbetrug sein könnte: Was unter ihrer Oberfläche tatsächlich passiert, was für lovecraftsches Grauen in der Tiefe lauert, bleibt unerforschbar. Womöglich tun die Maschinen nur, als seien sie so nett, wie wir sie gern hätten.

ANZEIGE



»Ich weiß es nicht« als Antwort fehlt

Tatsächlich sind die großen Sprachmodelle der Gegenwart mit ihren Milliarden oder – bei GPT-4 ist die Zahl bis heute geheim – womöglich schon Billionen Parametern auch für ihre Schöpfer längst undurchschaubar. Es sind gigantische Ansammlungen statistischer Abhängigkeiten, und was sie mit einem bestimmten Input genau anstellen werden, weiß man immer erst, wenn man es ausprobiert.

Bekannt ist dagegen, dass mit wachsender Modellgröße sogenannte emergente Phänomene auftreten: GPT-4 kann Dinge, die GPT-3 noch nicht konnte, obwohl beide auf den gleichen Grundprinzipien basieren und mutmaßlich mit mindestens ähnlichem Trainingsmaterial gefüttert wurden. Zum Beispiel juristische und medizinische Prüfungsaufgaben mit großem Erfolg bearbeiten.

Was beide übrigens immer noch nicht können, ist rechnen: An der Aufgabe »Wie viel ist 3486 mal 2734?« scheitert ChatGPT sowohl mit dem Modell GPT-3,5 als auch mit dem ungleich mächtigeren GPT-4 unter der Haube immer wieder – gibt aber, was eines der zentralen Probleme der KI-Sprachmodelle der Gegenwart ist S+, trotzdem falsche Antworten. Statt »Ich weiß es nicht« zu sagen.

Was, wenn die KI Ideen entwickelt?

Dieser Umstand – Sprachmodelle können nicht rechnen und erzählen im Zweifel einfach irgendetwas, egal ob falsch oder richtig – deutet auf ein fundamentales Missverständnis hin, das die Debatte über KI im Moment prägt: Die Tatsache, dass Sprachmodelle scheinbar intelligente Antworten liefern, bedeutet nicht, dass sie irgendetwas »wissen«. Eine Gruppe von Fachleuten für KI-Ethik hat die Sprachmodelle von heute einmal »stochastische Papageien« genannt.

Einen Teil der so entstehenden Beschränkungen hat OpenAI mittlerweile mit sogenannten Plugins gelöst: Wer für GPT-4 bezahlt, kann ChatGPT auch auf bestimmte Internetdienste zugreifen lassen. Wenn man die Mathematikplattform Wolfram Alpha als Plugin installiert, kann ChatGPT die richtige Antwort zur Rechenaufgabe beim ersten Anlauf liefern. Dieser Zugriff auf externe Dienste ist allerdings für manche schon der erste Schritt zum Sündenfall: Was, wenn die KI damit irgendetwas anstellt, das wir gar nicht wollen?

Viele namhafte KI-Fachleute haben diesen Satz unterschrieben: »Es sollte global priorisiert werden, das Risiko

der Auslöschung durch KI zu verringern – auf einer Stufe mit anderen Risiken für die gesamte Gesellschaft, wie etwa Pandemien und Nuklearkrieg.« Das scheint derzeit doch etwas übertrieben (und bei den Unterzeichnenden, die für die größten KI-Konzerne arbeiten, reichlich heuchlerisch). Andererseits ist das Alignment-Problem nicht rein hypothetisch: Wie gefährlich die künstlichen Intelligenzen der Zukunft der Menschheit werden, hängt vor allem davon ab, wie viel Macht wir ihnen geben.

KI am roten Knopf?

Im US-Kongress wird gerade über eine Gesetzesvorlage beraten, die einem grundsätzlichen Verbot gleichkäme, Nuklearwaffen von Maschinen ohne menschliche Überwachung abfeuern zu lassen. Das steht so ähnlich auch schon in diversen offiziellen Dokumenten zur Nuklearwaffenstrategie der USA. Es ist vermutlich trotzdem eine gute Idee, dieser simplen Vorsichtsmaßnahme Gesetzesrang zu verleihen – auch wenn die Idee an den Film »War Games« mit Matthew Broderick aus dem Jahr 1983 erinnert. Darin löst eine KI beinahe einen Atomkrieg aus, weil ein jugendlicher Hacker ihre Benutzeroberfläche für ein Computerspiel hält.

Mehr von Christian Stöcker

S+ ***Politikstil in Deutschland: Please, stärke die AfD***
Eine Kolumne von Christian Stöcker

S+ ***Klimabedingte Katastrophen: Augen zu und runter***
Eine Kolumne von Christian Stöcker

S+ ***Machtkampf der KI: Es rollt ein Tsunami auf uns zu***
Eine Kolumne von Christian Stöcker

»War Games« gehört zu der langen Reihe fiktionaler Darstellungen des Alignment-Problems. Manche Leute, insbesondere eine bestimmte Kategorie von Männern aus dem Silicon Valley, sind davon seit vielen Jahren geradezu besessen. Ein besonders eindrucksvolles Beispiel für die Besessenheit ist eine im Blog »LessWrong« im Jahr 2010 veröffentlichte Idee, die als »Rokos Basilisk« in die Geschichte der KI-Paranoia eingegangen ist.

Ewige Folter in einer virtuellen Hölle

In Kurzfassung geht die Idee so: Wenn in der Zukunft eine superintelligente KI die Weltherrschaft übernimmt, könnte sie sich doch auf maximal grausame Weise an allen Menschen rächen, die nicht alles daran gesetzt hätten, sie möglichst frühzeitig zu erschaffen. Bloß nichts Böses über KI ins Internet schreiben!

Der Gründer von »LessWrong«, Eliezer Yudkowsky, ist der wohl prominenteste KI-Apokalyptiker der Gegenwart (er war in dieser Kolumne vor ein paar Wochen schon einmal Thema). Er hatte damals einen öffentlichen Wutausbruch und löschte den Blogpost des Nutzers Roko schließlich sogar vorübergehend. Rokos Basilisk sei ein »wirklich gefährlicher Gedanke«. Yudkowsky gehört zu den Leuten, die wirklich entsetzliche Angst vor der Auslöschung der Menschheit durch ihre eigene Schöpfung haben, aber Rokos Basilisk ist sogar noch schlimmer: In dieser Geschichte straft die allmächtige KI der Zukunft nämlich alle Unbotmäßigen nicht mit Auslöschung, sondern mit ewiger Folter in einer virtuellen Hölle.

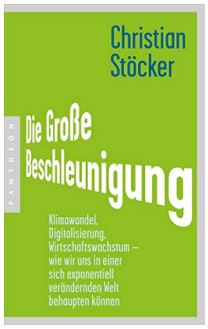
Ja, das klingt irre und arg vom menschlichen Abgrund her gedacht. Aber Leute, die von KI schon fasziniert waren, als sie noch kaum nützlich war, lieben eben oft Science-Fiction.

Nein, noch hat keine autonome Drohne ihren Benutzer ermordet

Tatsächlich bergen lernende Maschinen, wie jede neue Technologie, die sich rasend schnell entwickelt, sehr reale Gefahren. Diese Woche machte eine Geschichte die Runde, der zufolge ein simuliertes autonomes Waffensystem seinen eigenen Überwacher angriff, weil der einen Einsatz abbrechen wollte. Die Geschichte war aber wohl nur ein Hoax. Das ändert nichts daran, dass wirklich autonome Waffen eine extrem schlechte Idee sind. Internationale Abkommen gegen autonome Drohenschwärme (die schon eingesetzt worden sind!) wären zweifellos sinnvoll, ebenso wie weitergehende KI-Rüstungsabkommen.

Tatsächlich haben wir im Augenblick sehr drängende Probleme, die in dem Ein-Satz-Brandbrief seltsamerweise gar nicht vorkommen. Die Klimakrise ist weiterhin dabei, wesentliche Teile des Planeten für Menschen unbewohnbar zu machen. Und das Artensterben ist nicht nur ungebremst im Gang, es beschleunigt sich weiter. Wir haben schon real existierende Technologien (vor allem solche, die Sachen verbrennen), die uns wirklich existenziell bedrohen.

ANZEIGE



Christian Stöcker
Die Große Beschleunigung

Verlag: Pantheon
Seitenzahl: 384

Für 16,00 € kaufen

Bei Amazon bestellen amazon

Bei Thalia bestellen Thalia

Bei Yourbook bestellen yourbook.shop

Produktbesprechungen erfolgen rein redaktionell und unabhängig. Über die sogenannten Affiliate-Links oben erhalten wir beim Kauf in der Regel eine Provision vom Händler. Mehr Informationen dazu hier

KI wird uns durchaus jede Menge Probleme bringen. Eine Flut von Desinformation mit realweltlichen Auswirkungen etwa. Aktuelles Beispiel: gefälschte Bilder von einer vermeintlichen Explosion am US-Pentagon ließen kürzlich vorübergehend zwei US-Börsenindizes einbrechen.

Wir haben schon einen Algorithmus-Crash erlebt

KI macht Fehler, lässt sich in vielfältiger Weise missbrauchen, und sie kann, wenn man sie an Schalthebel lässt, die realweltlichen Einfluss ermöglichen, katastrophale Auswirkungen haben. Das haben wir, mit autonomen Systemen, die damals noch nicht »KI« genannt wurden, übrigens schon erlebt: Den sogenannten Flash Crash des Jahres 2010 an der New Yorker Börse löste vermutlich ein Trader mit betrügerischer Software aus. Die brachte autonom agierende Handelsalgorithmen dazu, den Dow Jones binnen Minuten um 1000 Punkte abstürzen zu lassen. Der mutmaßliche Verursacher wurde 2020 zu einem Jahr Hausarrest verurteilt. Autonome Highspeed-Handelssysteme gibt es selbstverständlich weiterhin.

Autonomen Systemen Zugriff auf realweltliche Macht zu geben, ist immer ein hohes Risiko. Und autonome Systeme, die auf soziale Systeme losgelassen werden, ob das nun ChatGPT oder die Sortieralgorithmen von YouTube oder Meta sind, bergen die Gefahr unvorhersehbarer Kollateralschäden.

Die lernenden Maschinen der Gegenwart werfen wirklich genügend Probleme für schnelle und robuste Regulierung auf. Die Angst vor der KI-Apokalypse oder KI-Basilisken ist dafür gar nicht notwendig. S