INTERVIEW

mit/with Karl von Wendt

von/by Reinhard Heil

Künstliche Intelligenz außer Kontrolle?

Artificial intelligence out of control?

In diesem Interview spricht Reinhard Heil mit dem Science-Fiction-Autor Karl von Wendt (Pseudonym Karl Olsberg) über die Möglichkeit, dass künstliche Intelligenz außer Kontrolle geraten und die Existenz der gesamten Menschheit gefährden oder zumindest ihr Entwicklungspotenzial nachhaltig und gravierend einschränken könnte. Unkontrollierbare künstliche Intelligenz, lange Zeit nur ein Thema der Science-Fiction, wird, vor allem seit dem großen Erfolg der auf sogenannten großen Sprachmodellen basierenden Chatbots, wie ChatGPT und Bard, in Politik und Medien zunehmend und sehr kontrovers diskutiert.

Reinhard Heil: Karl, Du bist Science-Fiction-Schriftsteller und – ich glaube das darf man durchaus so sagen – Aktivist. Vor Kurzem hast Du beide Eigenschaften zusammengeführt und „Virtua" veröffentlicht, einen Roman, dessen Untertitel „KI – Kontrolle ist Illusion" Deine Befürchtungen bezüglich der Risiken von Künstlicher Intelligenz (KI) auf den Punkt bringt.

Karl von Wendt: Ich würde mich nicht unbedingt als ‚Aktivist' bezeichnen. Ich sehe mich eher als Aufklärer. Dazu engagiere ich mich in einer internationalen Community von Menschen, die sich mit den Risiken der KI beschäftigen. Während das Thema in den USA und Großbritannien schon weit oben auf der politischen Agenda steht, wird es hier in Deutschland immer noch kaum ernst genommen.

Mein neuester Roman thematisiert zwar das Problem, das mir momentan die größten Sorgen macht – eine KI, die in mancher Hinsicht intelligenter ist als wir, uns manipuliert und so außer Kontrolle gerät. Aber „Virtua" ist nach wie vor ein Roman, der in erster Linie unterhalten soll und nicht die Zukunft vorhersagt. Ich habe ein umfangreiches Nachwort hinzugefügt, das meine Sicht auf das reale Problem grob schildert, aber trotzdem würde ich das Buch nicht als Aufklärungsmaterial sehen. Es soll eher nachdenklich machen und vielleicht den einen oder die andere dazu bringen, sich genauer über die Risiken hochentwickelter KI zu informieren. Dafür betreibe ich unter anderem mein Blog KI-Risiken.de.

Dass KI-Risiken in Deutschland nicht ernstgenommen würden, möchte ich so nicht stehenlassen. Es gibt eine große Anzahl von Stellungnahmen und Studien zu KI-Risiken von unterschiedlichen Institutionen. Unter anderem die Technikfolgenabschätzung (TA), der Deutsche Ethikrat und NGOs haben sich intensiv mit möglichen Risiken beschäftigt. Weniger Aufmerksamkeit erhalten allerdings, da stimme ich Dir zu, Risiken der unkontrollierbaren KI, zu denen es ja sehr unterschiedliche Positionen gibt: Beispielsweise spricht Elon Musk von einer existenziellen Bedrohung, Yann LeCun hingegen sieht darin vor allem wilde Spekulationen auf Basis methodisch zweifelhafter Studien und Sascha Lobo warnt, dass es ganz im Interesse der KI-Konzerne sei, auf spekulative zukünftige Risiken zu fokussieren, um von den gegenwärtigen abzulenken. Warum siehst Du die aktuell doch noch sehr spekulative Möglichkeit eines Kontrollverlustes als relevantes, drängendes Problem?

In der Tat meinte ich die gravierenden, sogar existenziellen Risiken einer unkontrollierbaren KI, die in Deutschland bisher kaum ernst genommen werden. Ich kann es gut verstehen, wenn man den Aussagen von Elon Musk oder Sam Altman misstraut, die auf solche existenziellen Risiken hinweisen. Bei ihnen liegt der Verdacht nahe, dass sie diese Aussagen aus geschäftlichem Kalkül oder Wichtigtuerei tätigen, zumal sie ja selbst alles tun, um solche Risiken überhaupt erst heraufzubeschwören. Allerdings sind sie bei Weitem nicht die Einzigen, die vor existenziellen Risiken warnen. Ein entsprechender Aufruf, solche Risiken ebenso ernst zu nehmen wie die Gefahren eines globalen Atomkriegs oder einer Pandemie, wurde im Mai 2023 von über 200 namhaften Wissenschaftler:innen unterschrieben, darunter absolute KI-Koryphäen wie die Turing-Preisträger Yoshua Bengio und Geoffrey Hinton, der oft als einer der Begründer moderner KI bezeichnet wird, oder Stuart Russell, der ein Standardwerk über neuronale Netze geschrieben hat und schon seit Jahren vor den Gefahren unkontrollierbarer KI warnt (Center for AI Safety 2023). Ein ähnlicher Aufruf kurz vor dem AI Safety Summit im November 2023 in London wurde zum Beispiel auch von Nobelpreisträger Daniel Kahnemann unterzeichnet, der zwar kein KI-Experte ist, sich dafür aber sehr gut mit den Schwachstellen des menschlichen Geistes auskennt, die es uns schwer machen, neuartige Risiken von großer Tragweite richtig einzuschätzen (Bengio et al. 2023). Wenn jemand

Keywords • *artificial intelligence, existential risk, governance, digital transformation*

https://doi.org/10.14512/tatup.33.1.64
Published online: 15. 03. 2024 (editorial peer review)

wie Sascha Lobo solche Leute als ‚nützliche Idioten' bezeichnet, ist das schon ziemlich arrogant und vermessen, finde ich. Mir wäre es am liebsten, wir würden im öffentlichen Diskurs nicht über die möglichen Motive dieser Leute spekulieren, sondern uns mit den konkreten Argumenten und Fakten auseinandersetzen, die für oder gegen ein existenzielles Risiko durch unkontrollierbare KI sprechen.

Was verstehst Du genau unter einer unkontrollierbaren KI?

Ich verstehe darunter eine KI, die in der Lage ist, automatisch Entscheidungen zu treffen, und alle Maßnahmen, die wir ergreifen, um diese Entscheidungen zu korrigieren oder zu unterbinden, konterkarieren oder umgehen kann. Eine solche KI wäre dann nicht mehr aufhaltbar oder korrigierbar, entweder, weil sie intelligenter ist als wir oder weil sie sich zum Beispiel ähnlich wie ein Virus in unserer technischen Infrastruktur ausgebreitet hat und wir sie nicht mehr daraus entfernen können. Wenn eine solche KI das falsche Ziel verfolgt, würde dabei höchstwahrscheinlich unsere Zukunft zerstört. Leider wissen wir nicht, wie wir ein Ziel so formulieren können, dass es wirklich gut für uns ist und keine Schlupflöcher oder Raum für Fehlinterpretationen bietet – ein Problem, das schon die alten Griechen kannten und das Goethe mit seinem Zauberlehrling sehr schön auf den Punkt gebracht hat. Es wird auch das Alignment-Problem, holprig übersetzt das Zielangleichungs-Problem, genannt.

Wenn man genauer hinschaut, wird sehr schnell offensichtlich, dass dieses Problem ungelöst und eine Lösung auch nicht in Sicht ist. Die Frage lautet daher eher, wie lange es noch dauert, bis wir in der Lage sind, eine potenziell unkontrollierbare KI zu entwickeln, und ob wir dann dumm genug sind, das auch zu tun. Letzteres beantwortet sich angesichts des jüngsten Dramas bei dem führenden KI-Konzern OpenAI wohl von selbst: Auch wenn es dabei anscheinend nicht primär um Sicherheit ging, haben die auf Sicherheit fokussierten Kräfte im Aufsichtsrat den Machtkampf gegen Sam Altman verloren. In einem globalen Wettrennen um Forschungserfolge und Marktführerschaft bleibt eben die Sicherheit gerne mal auf der Strecke, und allzu oft in der Geschichte hat blinde Gier die Vernunft besiegt.

Deiner letzten Aussage stimme ich hundertprozentig zu. ‚Gier frisst Hirn' ist wahrscheinlich die einzige echte anthropologische Konstante. In der TA vermeiden wir zwar konkrete Aussagen über die fernere Zukunft, ich frage Dich aber jetzt trotzdem nach dem Zeithorizont.

Die Frage, wie viel Zeit uns noch bleibt, kann niemand sicher beantworten, aber dass die Zeitlinie sich in den letzten Monaten drastisch verkürzt hat, dürfte offensichtlich sein, selbst wenn an den Spekulationen über einen Durchbruch beim mysteriösen Q*-Projekt von OpenAI nichts dran sein sollte. Yoshua Bengio schreibt in seinem Blog, dass er die Entwicklung einer Artificial General Intelligence, also einer allgemeinen KI auf menschlichem Niveau, im Jahr 2022 noch in 20–50 Jahren erwartete. Ein Jahr später hat sich dieser Erwartungshorizont auf fünf bis 20 Jahre verkürzt (Bengio 2023). Viele in der AI Safety Community haben noch kürzere Erwartungshorizonte. Uns bleiben also mit Glück noch höchstens zwei Jahrzehnte, wenn wir Pech haben nur noch wenige Jahre. Meine persönliche Erwartung ist da angesichts der sich überschlagenden Ereignisse eher auf der kürzeren Seite, ich befürchte, dass die kritische Phase potenziell unkontrollierbarer KI noch innerhalb dieses Jahrzehnts eintreten könnte. Aber selbst, wenn ich damit falsch liegen sollte, müssen wir uns auf diese Möglichkeit vorbereiten, damit wir nicht Gefahr laufen, wie bei Covid-19 und dem Klimawandel, schon wieder eine böse Überraschung zu erleben.

Karl von Wendt

hat 1988 über Künstliche Intelligenz (KI) promoviert, mehrere KI-relevante Start-ups gegründet und schreibt Blogbeiträge sowie unter dem Pseudonym „Karl Olsberg" Romane über die Risiken der KI.

Du hast zuvor von unkontrollierbarer KI als existenziellem Risiko gesprochen. Unter diesem Begriff werden Ereignisse gefasst, deren Eintreten die Existenz der gesamten Menschheit gefährden oder dauerhaft und schwerwiegend ihr Entwicklungspotenzial beschränken würde. Beispielsweise globale Naturkatastrophen aber auch von Menschen (mit-) verursachte Katastrophen wie extremer Klimawandel, Bioterrorismus oder ein atomarer Weltkrieg. Du hast gesagt, die Gefahr bestünde darin, dass eine KI nicht das ihr vorgegebene Ziel verfolgt oder zwar das richtige Ziel, aber mit den falschen Mitteln. Hierzu wird oft Nick Bostroms (2014) Behauptung angeführt, selbst eine mit der Herstellung von Büroklammern beauftragte allgemeine KI könnte sich der Kontrolle entziehen, damit niemand sie am Erreichen ihres Ziels hindern kann. Sind solche Gedankenspiele mit dem gegenwärtigen Stand der KI zu rechtfertigen, oder doch eher der Science-Fiction zuzurechnen?

Die Begründung dafür, dass unkontrollierbare KI ein reales Problem ist, liegt nicht in der Technik, sondern ergibt sich aus theoretischen Überlegungen, die auch in der Ökonomie in der so genannten Prinzipal-Agent-Theorie und in der Entscheidungstheorie zum Tragen kommen. Es ist prinzipiell äußerst schwierig, einem Agenten, also zum Beispiel einem Dienstleister, einen Auftrag so zu stel-

https://doi.org/10.14512/tatup.33.1.64

len, dass sichergestellt ist, dass dieser genau das tut, was man möchte. Jeder, der schon mal ein Haus gebaut hat, weiß das. Einerseits ist das ein Kommunikationsproblem – wie sage ich dem anderen, was ich genau will? Andererseits ein Problem des Interessenkonflikts – der Dienstleister will vielleicht möglichst wenig Leistung für das zugesagte Geld erbringen, um seinen eigenen Profit zu maximieren. Dann gibt es noch das Messproblem – wie stelle ich eigentlich fest, ob das, was der Dienstleister gemacht hat, das ist, was ich wollte? Baumängel treten zum Beispiel oft erst nach Jahren auf. Alle drei Probleme kommen bei KI noch deutlich stärker zum Tragen als bei menschlichen Dienstleistern, denn wir verstehen KI viel weniger als Menschen und man kann von einer KI auch keinen Schadensersatz fordern oder sie vor Gericht stellen. Wenn die KI dann auch noch intelligenter ist als wir und vielleicht besonders gut darin, zu lügen und Menschen zu manipulieren, wird es noch schwieriger.

Der ‚optimale Weltzustand', den die KI auf Basis ihres Ziels anstrebt, ist höchstwahrscheinlich nicht mit unseren Wünschen und oft auch nicht mit unserem Überleben vereinbar.

Voraussetzung unkontrollierbarer KI wäre also demnach, dass KI eigene Ziele entwickelt und zudem erkennt, dass es zum Erreichen dieser Ziele beiträgt, Kontrolle auszuüben?

Sobald die KI irgendein Ziel verfolgt und in der Lage ist, auf Basis eines komplexen Weltmodells einen Plan für die Zielerreichung zu erstellen, der auch sie selbst beinhaltet, kommt es nahezu zwingend zu einem Interessenkonflikt. Denn aus einem beliebigen Ziel ergeben sich so genannte instrumentelle Ziele, die für fast alle übergeordneten Ziele dieselben sind. Wenn die KI zum Beispiel abgeschaltet wird, kann sie ihr Ziel nicht mehr erreichen. Also ist es ein instrumentelles Ziel, nicht abgeschaltet zu werden. Wenn wir das Ziel der KI ändern, kann sie ihr ursprüngliches Ziel ebenfalls nicht mehr erreichen, also ist es ein instrumentelles Teilziel dieses Ziels, nicht geändert zu werden. Auch Macht und Einfluss zu erhöhen sind instrumentelle Ziele. Und die stehen unserem Wunsch entgegen, die KI kontrollieren, ihr Ziel ändern und sie notfalls abschalten zu können. Heutige KIs wie GPT-4 oder Gemini sind noch keine Agenten im hier gemeinten Sinn, sie erstellen keine langfristigen Pläne, die sie dann systematisch verfolgen. Aber wenn wir eine bestimmte Schwelle überschreiten, könnte es sein, dass eine solche agentische KI entsteht und sich unserer Kontrolle entzieht. Und dann ist es per se äußerst unwahrscheinlich, dass sie genau das tut, was wir wollen. Denn der ‚optimale Weltzustand', den die KI auf Basis ihres Ziels anstrebt, ist höchstwahrscheinlich nicht mit unseren Wünschen und oft auch nicht mit unserem Überleben vereinbar.

Das Büroklammer-Beispiel von Bostrom ist bewusst überzogen, aber ich glaube, ich könnte Dir zu nahezu jedem Ziel, das Du mir nennst, eine extreme Lösung beschreiben, bei der die Zukunft der Menschheit zerstört wird. Und ich bin keine superintelligente KI. Stuart Russell hat es mal so ausgedrückt: ‚Wenn wir in der Beschreibung des Ziels für die KI eine Variable vergessen, die uns wichtig ist, dann wird die KI diese Variable womöglich auf einen Extremwert setzen.' Ein naheliegendes Beispiel wäre der Klimawandel: Für fast alle Ziele wäre es aus Sicht der KI nützlich, möglichst viel Rechenleistung zu haben, was im Extremfall riesige weltweite Serverfarmen und eine rapide globale Erwärmung zur Folge haben könnte. Das könnte man natürlich bei der Zielformulierung spezifisch ausschließen, aber der CO_2-Anteil in der Atmosphäre ist selbst ein Beispiel für eine kritische Variable, die wir noch gar nicht kannten, als wir mit der Industrialisierung begannen.

Vielleicht führt der Begriff allgemeine KI hier auch in die Irre. Ist allgemeine KI tatsächlich eine Voraussetzung für den Kontrollverlust oder könnten nicht auf wenige Aufgaben spezialisierte Systeme, wie wir sie bereits besitzen, zu einem Kontrollverlust führen?

In der Tat ist es ein Problem, dass wir den Menschen als Maßstab nehmen, um die Leistungsfähigkeit einer KI und damit auch ihre potenzielle Gefährlichkeit zu beurteilen. KI funktioniert völlig anders als unsere eigene Intelligenz. So, wie ein Flugzeug nicht mit den Flügeln schlägt und keine Federn hat, ‚denkt' KI nicht wie wir.

Schon jetzt sind KIs Menschen in vieler Hinsicht überlegen, nicht nur in Spielen wie Schach oder Go, sondern auch zum Beispiel in der schieren Menge des Wissens, das sie verarbeiten können und in der Geschwindigkeit. Wer spricht schon wie GPT-4 dutzende Sprachen fließend und kann ein Essay in 30 Sekunden schreiben? Die agentische Planungsfähigkeit, die ich beschrieben habe, verbunden mit einer übermenschlichen Fähigkeit zur Manipulation von Menschen oder technischen Systemen könnten Eigenschaften einer potenziell unkontrollierbaren KI sein. Die KI könnte dann vielleicht immer noch vieles nicht, was ein Mensch kann, zum Beispiel ein Spiegelei braten. Aber wenn die KI geschickt genug wäre, könnte sie uns Menschen dazu bringen, zu tun, was ihrem Ziel nützt, so wie ein menschlicher Diktator die Massen manipulieren und ihnen seinen Willen aufzwingen kann. Es sind auch Szenarien denkbar, bei denen die KI sich selbst im Internet verbreitet und technische Systeme manipuliert, oder ein militärisches System könnte außer Kontrolle geraten und einen Weltkrieg verursachen (Future of Life Institute 2023). Das größte Problem dabei ist, dass wir bereits heutige

 https://doi.org/10.14512/tatup.33.1.64

KI kaum verstehen und deswegen extrem schwer vorhersagen können, wann und wie zukünftige, noch leistungsfähigere Systeme außer Kontrolle geraten könnten.

Nimmt man deine Befürchtungen ernst, so scheint die einzige Möglichkeit, den Untergang der Menschheit zu verhindern, darin zu bestehen, auf die Weiterentwicklung von KI gänzlich zu verzichten oder sie doch zumindest extrem streng zu regulieren. Im März 2023 veröffentlichte das Future of Life Institute einen offenen Brief, in dem gefordert wurde, die Entwicklung von KI Systemen, die mächtiger sind als GPT-4, für mindestens sechs Monate zu unterbrechen und diesen Zeitraum zu nutzen, um Sicherheitsprotokolle zu entwickeln, die eine sichere Weiterentwicklung gewährleisten. Das Moratorium kam zwar nie zustande, aber im Dezember 2023 wurde zumindest im Rahmen des AI-Acts der EU beschlossen, dass Hersteller sogenannter ‚general artificial intelligence systems', mit deren Einsatz große systemische Risiken verbunden sind, Modellevaluierungen durchführen, systemische Risiken bewerten und abschwächen, Schwachstellenanalysen durchführen und der Kommission über schwerwiegende Vorfälle berichten müssen. Zudem müsse die Cybersicherheit gewährleistet werden. Erscheinen Dir diese Maßnahmen hinreichend, um einen Kontrollverlust zu vermeiden?

Ich würde nicht sagen, dass die Weiterentwicklung von KI insgesamt gestoppt werden sollte. Im Gegenteil glaube ich, dass hochentwickelte KI in vielen Anwendungsgebieten extrem nützlich ist. Mein Lieblingsbeispiel ist AlphaFold – eine KI, die das so genannte Protein-Faltungsproblem gelöst und so die pharmazeutische und medizinische Forschung enorm vorangebracht hat. AlphaFold kann weit besser als der Mensch die räumliche Struktur von Proteinen auf der Basis der chemischen Formel vorhersagen, was sehr wichtig ist, um die Wirkung auf den Organismus einzuschätzen. Theoretisch kann das natürlich auch missbraucht werden, um zum Beispiel biologische Waffen zu entwickeln, aber es besteht keine Gefahr, dass AlphaFold oder ein ähnliches spezialisiertes System jemals instrumentelle Ziele entwickelt und unkontrollierbar wird. Ich glaube, dass fast alle wissenschaftlichen und gesellschaftlichen Probleme von solchen spezialisierten KIs gelöst werden könnten, ohne dass wir dabei die Zukunft der Menschheit riskieren. Eine Gefahr geht vor allem von allgemeinen KIs aus, die nahezu beliebige Probleme lösen und langfristige Pläne schmieden können. Solche ‚Universalgenies' brauchen wir eigentlich gar nicht. Natürlich sind sie in gewisser Hinsicht praktischer als spezialisierte KIs, weil man eben fast alles mit nur einer KI machen kann, und wirtschaftlich erscheinen sie sehr attraktiv. Aber sie sind eben auch viel gefährlicher.

Agentische Planungsfähigkeit verbunden mit einer übermenschlichen Fähigkeit zur Manipulation von Menschen oder technischen Systemen könnten Eigenschaften einer potenziell unkontrollierbaren KI sein.

Was den AI Act angeht, freue ich mich, dass die deutsche Forderung, ausgerechnet die sogenannten ‚foundation models', also zum Beispiel große Sprachmodelle wie GPT-4, von der Regulierung auszuklammern, sich offenbar nicht durchgesetzt hat. Ob allerdings die in der Beschlussvorlage genannten Regulierungen am Ende wirklich etwas bringen, muss sich noch zeigen. Ich bin da generell skeptisch, da wir ja nicht einmal im Ansatz verstehen, wie komplexe Systeme wie GPT-4 funktionieren, warum sie zum Beispiel bestimmte Entscheidungen so und nicht anders treffen. Ich glaube, ab einem bestimmten Punkt der Leistungsfähigkeit ist es nicht mehr möglich, bei einer allgemeinen KI vorauszusagen, ob sie unkontrollierbar werden könnte. Wo genau dieser Punkt liegt, weiß allerdings niemand. Daher wäre es aus meiner Sicht das Klügste, tatsächlich auf die Weiterentwicklung solcher allgemeinen KIs zu verzichten und die Energie und Investitionen stattdessen in spezialisierte KIs wie AlphaFold zu stecken, zumindest so lange, bis wir allgemeine KI besser verstehen und wissen, wie wir sie sicher kontrollieren können. Deshalb unterstütze ich Forderungen nach einem Moratorium – alles, was das aktuelle ‚Wettrennen übers Minenfeld' zwischen Google/Deepmind, Microsoft/OpenAI, Meta und noch ein paar anderen verlangsamt, ist hilfreich. Darüber hinaus wünsche ich mir mehr Forschung zum besseren Verständnis der notwendigen und hinreichenden Bedingungen für die Unkontrollierbarkeit von KI, so dass wir besser als heute ‚rote Linien' definieren können, die niemand bei klarem Verstand überschreiten wollen würde.

Literatur

Bengio, Joshua (2023): FAQ on catastrophic AI risks. In yoshuabengio.org, 24. 06. 2023. Available online at https://yoshuabengio.org/2023/06/24/faq-on-catastrophic-ai-risks, last accessed on 09. 01. 2024.

Bengio, Yoshua, et al. (2023): Managing ai risks in an era of rapid progress. In: arxiv.org, 12. 11. 2023. https://doi.org/10.48550/arXiv.2310.17688

Bostrom, Nick (2014); Superintelligence. Paths, dangers, strategies. Oxford: Oxford University Press.

Center for AI Safety (2023): Statement on AI risk. Available online at https://www.safe.ai/statement-on-ai-risk, last accessed on 09. 01. 2024.

Future of Life Institute (2023): Artificial escalation. (Video) Available online at https://www.youtube.com/watch?v=w9npWiTOHX0, last accessed on 09. 01. 2024.