01.12.2023

Transkript

„Ist KI auf dem Weg zur Superintelligenz? Fortschritte bei der Forschung zu AGI und Alignment“

Expertin und Experten auf dem Podium

- **PD Dr. Jessica Heesen**
  Leiterin des Forschungsschwerpunkts Medienethik, Technikphilosophie und KI, Eberhard Karls Universität Tübingen
- **Prof. Dr. Holger Hoos**
  Alexander von Humboldt Professor für Methodik der Künstlichen Intelligenz, Rheinisch-Westfälische Technische Hochschule Aachen (RWTH)
- **Prof. Dr. Kristian Kersting**
  Leiter des Fachgebiets Maschinelles Lernen, Technische Universität Darmstadt
- **Prof. Dr. Marcus Liwicki**
  Vizerektor für künstliche Intelligenz und Leiter der Forschungsgruppe maschinelles Lernen, Luleå University of Technology, Schweden
- **Bastian Zimmermann**
  Redakteur für Digitales und Technologie, Science Media Center Germany,
  und Moderator dieser Veranstaltung

Mitschnitt

- Einen Videomitschnitt finden Sie unter: https://www.sciencemediacenter.de/alle-angebote/press-briefing/details/news/ist-ki-auf-dem-weg-zur-superintelligenz-fortschritte-bei-der-forschung-zu-agi-und-alignment/
- Falls Sie eine Audiodatei oder eine Sprecheransicht des Videomittschnitts benötigen, können Sie sich an redaktion@sciencemediacenter.de wenden.

Transkript

Moderator [00:00:00]

Hallo, liebe Journalistinnen und Journalisten, herzlich willkommen hier zu unserem Press Briefing zur Debatte über Superintelligenz, AGI und Alignment. Mein Name ist Bastian Zimmermann. Ich bin Redakteur hier beim Science Media Center, und mit mir habe ich heute noch eine Expertin und drei Experten. Herzlich willkommen erst mal an Sie. Schön, dass Sie da sind. Ich stelle Sie gleich noch im Detail vor.

Ja, die letzten Wochen, das Drama um OpenAI, der Rauswurf und die Wiedereinstellung von Sam Altman, Spekulationen um Q*, das alles hat ja zuletzt die Debatte um Generelle Künstliche Intelligenz, deren sinnvolle Gestaltung, aber eben auch um die Befürchtungen zu Themen wie AGI und Superintelligenz weiter angeregt. Das Thema kam ja Anfang und Mitte des Jahres schon einmal im Kontext der offenen Briefe auf, die vor den Gefahren von KI gewarnt haben. Damals gab es dann oft das Argument, dass diese Debatte über "diffuse" Zukunftsthemen die realen aktuellen Probleme und Risiken im Bereich von KI verschleiert, wie zum Beispiel Biases, prekäre Bedingungen für die Personen, die die Daten annotieren, und die große Marktmacht weniger Firmen. Die Probleme sind ja noch nicht wirklich gelöst. Gerade die Marktmacht und das Problem der kaum vorhandenen Rechenschaftspflicht der Unternehmen haben sich ja zum Teil noch verstärkt. Wir haben jetzt gesehen, dass Sam Altman ja offenbar nicht gefeuert werden kann. Trotzdem ist die gesellschaftliche Debatte unserer Meinung nach mittlerweile so weit, dass man über AGI und Superintelligenz reden kann, ohne sich vorwerfen zu müssen, dass man andere wichtige Aspekte der Debatte verschleiert. Deswegen haben wir eben heute, einen Tag nach dem ersten Geburtstag von ChatGPT, hier unser Briefing, in dem wir über AGI, Superintelligenz und Alignment reden wollen.

Bevor ich gleich zum Vorstellen komme, noch der Hinweis an die Journalistinnen und Journalisten: Die Fragen stellen Sie bitte einfach über die Fragefunktion von Zoom. Das ist dann einfacher, dann können Ihre Kolleginnen und Kollegen die Fragen auch immer sehen, dann gibt es weniger Doppelung. Also da bitte das F&A-Modul benutzen und nicht den Chat.

Dann zu unserer Expertin und den Experten. Danke noch mal, dass Sie alle hier sind. Ich stelle Sie jetzt kurz vor. Wir haben hier einmal Jessica Heesen, die Leiterin des Forschungsschwerpunkts Medienethik, Technikphilosophie und KI an der Universität Tübingen. Sie wird uns heute aus ethischer Perspektive ihre Einschätzungen liefern. Dann haben wir drei Experten aus der KI-Forschung direkt. Einmal Holger Hoos, Alexander von Humboldt Professor für Methodik der Künstlichen Intelligenz an der RWTH Aachen. Dann Kristian Kersting, den Leiter des Fachgebietes Maschinelles Lernen an der TU Darmstadt. Und Marcus Liwicki, Vizerektor für Künstliche Intelligenz und Leiter der Forschungsgruppe Maschinelles Lernen an der Lulea University of Technology in Schweden.

Damit wir eine Basis für das Thema haben, über die gleichen Sachen reden und den, sage ich mal, Grundpfeiler der Debatte schon mal legen, haben wir ein paar kurze Eingangsfragen pro Person. Wir fangen an mit Kristian Kersting und der Frage: Was ist denn eigentlich aktuell der Stand der Forschung zu AGI und Superintelligenz? Wie weit ist man da gerade aus deiner Sicht?

Kristian Kersting [00:03:08]

Es wurde ja gerade gesagt, dass ChatGPT gestern sein Einjähriges hatte. Ich glaube, es kann immer noch nicht eine Geburtstagskerze auspusten. Denn ich glaube einfach, dass viele, viele menschliche Fähigkeiten in dem System noch nicht abgedeckt sind. Vielleicht geht es aber manchmal auch gar nicht so sehr um die Frage, ob AGI nun besser ist als der Mensch. Sondern es geht darum, dass die Systeme Sprünge in ihren Fähigkeiten hinbekommen haben, die wir alle vielleicht vorher gar nicht so direkt gesehen haben. Oder vielleicht war es eine Utopie für uns. AGI haben wir sicherlich

nicht. Q* Ist für mich einfach der nächste schöne Verkaufsschlager. Ich glaube nicht, dass wir da sind. Ich glaube, das wird noch ziemlich lange brauchen.

Das muss dann aber nicht heißen, dass wir nicht jetzt schon auch daran arbeiten: "Was wäre, wenn?", und uns auch darüber ein bisschen Gedanken machen. Ich finde es sehr vernünftig, jetzt zum Beispiel über Regulierung zu diskutieren und die auch durchzuführen. Das finde ich völlig normal. Aber wir haben noch keine AGI. Und ich persönlich glaube auch nicht daran, dass wir in absehbarer Zeit AGI in dem Sinne, wie es normalerweise diskutiert wird, erreichen werden, weil die Aufgaben viel zu schwierig sind. Wir werden aber viele, viele Fortschritte sehen.

Moderator [00:04:26]

Gut, dabei belassen wir es erst mal, vielen Dank. An Marcus Liwicki die Frage: Was ist denn eigentlich dieses Alignment, von dem immer geredet wird und das OpenAI offenbar für eine wichtige Methode hält, um KI ethisch zu gestalten?

Marcus Liwicki [00:04:44]

OpenAI arbeitet ja immer weiter an neuen Technologien, um bessere KI, bessere Chatbots herzustellen. Und man muss dann immer überlegen, was ist eigentlich mit dem menschlichen Faktor? Wie können wir die KI – man spricht manchmal von diesem KI-Monster – wie kann man das bändigen, sodass es menschliche Werte widerspiegelt? Und da gibt es mehrere Möglichkeiten.

Das eine, was derzeit fast überall angewendet wird, auch in ChatGPT, ist, man trainiert die KI mit menschlichem Feedback. Wir trainieren ein Sprachmodell, und wenn wir irgendein Sprachmodell – das werden wir nachher auch noch hören von unserer Kollegin Jessica – wenn man irgendein Sprachmodell trainiert, dann wird das wahrscheinlich Sachen sagen, die ethisch nicht ganz korrekt sind. Und dann heißt es: Hier solltest du nicht antworten oder da mal lieber ein bisschen vorsichtig sein. Das nennt man menschliches Feedback. Da sind also ein paar Leute bei OpenAI, die schauen sich an, wie ChatGPT funktioniert und was es für Antworten gibt. Und dann sagt man: Hier lieber das oder hier lieber so antworten oder generell nicht versuchen, diesen Bias Mann, Frau [zu machen] oder irgendwelche Kulturen hervorzuheben. Frühere Versionen von ChatGPT haben zum Beispiel Amerikaner gegenüber einigen Afrikanern bevorzugt, und da gab es ganz viele Probleme. Also dieses menschliche Feedback reinzubringen, das ist der erste Schritt.

Der zweite Schritt ist dann eben, dass man die KI trainiert, um menschliche Entscheidungen zu verbessern. Dass man also die Entscheidungen, die Menschen treffen und die eventuell einen Bias haben könnten, dass man da den Menschen hilft, bessere Entscheidungen zu treffen.

Und der dritte Schritt, den OpenAI sich auch als Ziel setzt, ist, selbst zu forschen, um besser in den ersten beiden Schritten zu werden, also dass man das gesamte Ding automatisiert. Und ich denke, die ersten beiden Schritte werden in sehr, sehr vielen Anwendungen erforscht und praktisch auch gelebt. Mit dem dritten ist es relativ schwierig. Denn wenn man dann irgendwann den Menschen rausnimmt aus dem Feedback Loop, kann es sein, dass die KI sich in eine Richtung entwickelt, die dann nicht mehr diese menschlichen Werte hat. Aber das ist eher noch ein bisschen Utopie als tatsächlich realisierbar.

Moderator [00:07:25]

Ja, vielen Dank erst mal. Jessica Heesen, wie sieht es denn mit dem Alignment-Ansatz aus? Wie würden Sie das aus ethischer Perspektive beurteilen?

Jessica Heesen [00:07:37]

Ich kann vielleicht erst noch mal kurz reagieren auf das, was Marcus gerade gemeint hat, weil da wieder die menschlichen Werte genannt wurden. Ich muss sagen, ich als Ethikerin habe ganz große Probleme mit den menschlichen Werten, weil die müssen erst ausgehandelt werden, die sind kulturell unterschiedlich. Und nicht alles, was Menschen tun, hat was mit allgemeinen Werten zu tun. Es ist zum Beispiel auch ein Wert, viel Geld zu verdienen. Es gibt ganz unterschiedliche Motivlagen, und auch Rechtspopulisten haben menschliche Werte. Das heißt also, es ist immer schon mit Vorsicht zu genießen, wenn man so viel mit solchen Begriffen wie "menschliche KI" um sich wirft oder "human zentrierte KI" oder eben "Alignment". Da denkt man, wenn man irgendwie was ausrichtet nach Menschen, dann ist es schon an sich gut. Und das ist eben einfach überhaupt nicht der Fall.

Aber trotzdem, was ich beobachte bei der Debatte, ist generell was ganz Interessantes, was fast so eine Art Marketingprinzip ist, nämlich, dass dieser Begriff Alignment einfach auch genutzt wird, um die Diskurshoheit zu haben in einem bestimmten Gebiet. Denn was ich beobachte, wenn man sich genauer anschaut, was OpenAI meint mit Alignment und was generell in der ganzen Diskussion passiert, dann sieht man, das sind Diskussionen, die gibt es schon lange. Die Prinzipien der AI-Ethik, der KI-Ethik sind schon sehr lange bekannt, werden eben einfach nicht besonders stark rezipiert, sagen wir mal im Silicon Valley. Weil natürlich sowas wie Ethik durch Design ist eine ganz anerkannte Regel. Die zieht sich überall auch durch die Regulierung, auch bei uns in Deutschland und in der EU, also dass man sagt, man muss in der Entwicklung von Anfang an schauen, dass es eben nicht zu unerwarteten Nebeneffekten kommt. Denn darum geht es ja vor allen Dingen.

Und was mich am meisten wundert, ist, es gibt ja auch in der Informatik selber schon einen etablierten Forschungszweig, den nennt man FAccT, das steht dann für Fairness, Accountability and Transparency. Und ich habe nicht den Eindruck, dass es eine Verbindung gibt zwischen diesen unterschiedlichen Forschungslinien. Sondern OpenAI sagt einfach, wir machen jetzt unser Alignment-Projekt, und definiert ganz spezifisch für sich, was die Problemlagen sind. Und versteht das Ganze dann ganz stark aus dem maschinellen Lernen heraus. Und dazu gehört dann eben auch, das beobachtet man ja, dass OpenAI und auch viele andere, die in der KI-Forschung stark sind, die imaginieren dann, was hier gerade schon gesagt wurde, so eine Art Monster oder so eine bösartige Superintelligenz. Wo ich mich immer frage, warum soll die eigentlich bösartig sein? Aber gut, die wird eben imaginiert. Und praktischerweise liefert dann OpenAI auch gleich die Instrumente mit, um dem zu begegnen. Also das ist eigentlich ein super Marketinggag, könnte man sagen. Also das ist der eine Punkt.

Und vielleicht auch doch noch mal – du hattest es ja gerade schon gesagt, Bastian – zu der Frage: Was wird alles vergessen, wenn man so damit umgeht, wie OpenAI das tut, oder wie die Diskussion gerade verläuft? Dazu würde ich noch mal sagen, dass es sich hierbei um eine sehr technokratische Sichtweise handelt. Man sagt hier, wir lösen das Problem von KI und Ethik und Alignment, indem wir das maschinelle Lernen verbessern. Das ist natürlich ein anerkennenswertes und wichtiges Vorhaben. Aber es gibt eben doch diese vielen anderen Themen in Bezug auf KI und KI-Sprachmodelle oder Foundation Models, die angewendet werden in ganz unterschiedlichen Zusammenhängen. Und da muss man ganz einfach, wie das auch in der Ethik üblich ist, fragen: Wem nützen diese Anwendungen eigentlich und wem schaden sie? Wenn man nur ganz allgemein da so rumfischt und sagt, es muss ein Alignment geben, dann werden genau diese politischen Betrachtungsweisen oder die Betrachtungsweisen, die in die Richtung gehen, dass KI auch immer ein soziotechnisches System ist, die werden einfach vernachlässigt. Und das ist eben auch etwas, was einfach zu dünne ist, sagen wir mal, wenn man aus der Perspektive der Ethik darauf schaut.

Moderator [00:11:06]

Ja, vielen Dank. Holger, an dich dann noch die Frage: Inwiefern hältst du die Forschung zu AGI und Superintelligenz eigentlich für potenziell gefährlich? Müsste da aus deiner Sicht mehr reguliert werden?

Holger Hoos [00:11:20]

Ja, das ist natürlich immer eine schwierige Frage. Aber ich lehne mich jetzt einfach mal ziemlich weit aus dem Fenster und erkläre dann auch ein bisschen, warum ich das tue. Also, ich finde Forschung, die sich zum Ziel setzt, heutzutage AGI zu erzeugen, gehört verboten. Punkt. Und der Grund dafür ist der folgende: dass, egal wie nah wir dran sind, wenn wir AGI erreichen, wir als Gesellschaft nicht bereit sind, damit auch nur annäherungsweise verantwortungsvoll umzugehen. Und ich finde, die Debatte in der Öffentlichkeit, aber auch in wissenschaftlichen Kreisen zeigt das sehr, sehr deutlich. Deswegen kann man sich eigentlich nur wünschen, dass entweder Kristian recht hat, und ich stimme ihm dazu, dass wir noch lange nicht da sind. Ich sage jetzt nicht, dass wir nicht da sind, das ist ganz offensichtlich. Aber dass wir noch lange nicht da sind. Oder dass, wenn irgendwann wirklich eine sprunghafte Entwicklung stattfindet – was, glaube ich, nach den Überraschungen im letzten November niemand mehr so ganz grundsätzlich ausschließen möchte oder nur sehr wenige Leute – dass wir dann vorher zu Sinnen kommen und sagen: Es gibt Grenzen, über die sollte man derzeit einfach nicht hinausgehen, weil die gesellschaftlichen Risiken zu groß sind gegenüber den Vorteilen, die man möglicherweise daraus ziehen könnte.

Das ist sicherlich was, das kann Jessica professioneller beurteilen. Aber alles, was sie gerade zu dem Thema gesagt hat, zum Beispiel Value Alignment, also Alignment mit Werten – die richtige Frage, wie Jessica genau gesagt hat, ist: wessen Werte? Um wessen Werte geht es hier eigentlich? Und ich persönlich sehe keinerlei Anzeichen dafür, dass man sich bei OpenAI zum Beispiel Mühe gibt, eine breite und diverse Palette von Werten zugrunde zu legen. Ich glaube auch, das kann man von einer Technologiefirma wie OpenAI jetzt nicht wirklich verlangen. Ich glaube, das braucht eine breitere gesellschaftliche Debatte.

Ich glaube weiterhin, dass die Gefahr durch AGI nicht erst dann entsteht, wenn man wirklich sozusagen die Schwelle erreicht oder überschritten hat. Ich glaube, die Gefahr besteht schon sehr viel früher. Ein Teil der Gefahr besteht schon jetzt. Die kommt dann, wenn Leute glauben, dass man schon Fähigkeiten realisiert hat in diesem System, die es noch gar nicht hat. Und diese Irrmeinung wird ganz stark dadurch befördert, wie wir Sprache verwenden. Wir reden, selbst technisch versierte Menschen reden häufig von "der KI", wenn wir von einem KI-System oder KI-Tool reden. Das vermenschlicht. Das lädt uns dazu ein und verführt uns dazu, diesem System Fähigkeiten zuzusprechen und zuzuschreiben und uns dann auch entsprechend zu verhalten im Umgang mit dem System, aber auch in der Anwendung der Systeme, Fähigkeiten, die diese Systeme objektiv noch gar nicht haben.

Das letzte Problem, was ich ansprechen möchte, ist: Warum ist das so, dass wir jetzt von ChatGPT zum Beispiel so beeindruckt sind, wenn es doch jede Menge Anzeichen dafür gibt, dass da in Wirklichkeit sehr große Schwächen vorhanden sind? Kristian hat schon auf einige hingewiesen. ChatGPT kann im Wesentlichen nur sehr, sehr eingeschränkt logisch denken oder Schlüsse ziehen. Die ganze Kahneman-System-2-Problematik ist bei ChatGPT nur extrem schwach vertreten. Und das System-2 [beinhaltet] ja genau die Sachen, die menschliche Intelligenz von anderen Formen der Intelligenz, im Tierreich zum Beispiel, unterscheidet. Also mal ganz ganz grob gesagt: Lernen kann schon jede Ameise. Aber bis diese ganze KI-Debatte vor 20 Jahren so richtig losging, als das maschinelle Lernen ganz starke Fortschritte und auch sehr beeindruckende Fortschritte gemacht hat, die auch sehr, sehr nützlich sind – da möchte ich nicht falsch verstanden werden – aber bis zu dem Zeitpunkt war es ja so, dass man menschliche Intelligenz nicht in erster Linie durch Lernfähigkeit definiert hat, sondern durch logische und sprachliche Fähigkeiten. Und bei den sprachlichen

Fähigkeiten sind wir genau bei ChatGPT. Ich glaube, wir sind als Menschen alle darauf angelegt, die Leistungsfähigkeit und den Bildungsstand und auch auf eine gewisse Art und Weise, wie ernst man jemanden nehmen muss, aufgrund der sprachlichen Ebene, auf der man kommuniziert, irgendwie heuristisch abzuschätzen. Und da punktet ChatGPT natürlich ganz enorm. Denn ChatGPT kann sich schon in allen möglichen Sprachen sehr gewandt ausdrücken. Und ich denke, dadurch schreiben wir ChatGPT dann auch intellektuelle Eigenschaften und Fähigkeiten zu, die bei Menschen mit so einem sprachlichen Ausdrucksniveau in der Regel da wären, aber bei ChatGPT und ähnlichen Systemen eben nicht.

Moderator [00:15:53]

Wir haben auch schon einige konkrete Nachfragen. Vielleicht an dich die erste, Kristian. Wie kommt man denn jetzt zur AGI? Welche Schritte müssten auf dem Weg dahin zurückgelegt werden? Welche Probleme müssten gelöst werden? Und in der Debatte um Q* ging es ja auch darum, dass das angeblich besser darin sein soll, gewisse mathematische Gleichungen zu lösen. Welchen Stellenwert hat denn sowas für die Fähigkeit zu generalisieren?

Kristian Kersting [00:16:17]

Na ja, ich glaube, es geht darum, genau wie beim Schachspielen. Solange wir Schachspielen nicht gelöst hatten mit der Maschine, war das eine große Frage, ob wir damit zeigen, dass sie intelligent ist. Jetzt ist es die Mathematik oder das Lösen von gewissen Gleichungen. Und natürlich wäre es eine starke Unterstützung in der Forschung beispielsweise oder auch im Ingenieursbereich. Die Frage ist – und ich glaube, da wäre ich mit Jessica sehr d'accord – ob wir das überhaupt wollen? Und ich glaube, wir brauchen eben auch diesen Diskurs darüber, was wir wollen.

Ich würde aber, bevor ich gleich noch mal sage, was fehlt für AGI, darauf hinweisen: Meiner Meinung nach – und da würde mich auch interessieren, wie Juristen das sehen – dürfen wir gar nicht an AGI als Ziel arbeiten, zumindest [dürfen wir es nicht] erreichen. Denn das Klonen ist in Deutschland verboten. Und ich glaube, es macht keinen Unterschied, ob ich biologisch klone oder mit irgendwelchen Silicon-based Schaltkreisen. Also das ist meine Prämisse. Ich finde es als utopisches Ziel spannend. Das motiviert zu fragen: Was macht Intelligenz aus? Denn daran möchte ich noch mal erinnern: Wir müssen verstehen, dass die Frage AGI nicht das Spannende ist, sondern für mich ist das Spannende: Was ist Intelligenz? Das kann mir auch in der Psychiatrie helfen, in der Psychologie, in den Kognitionswissenschaften. Es kann mir in der Ökonomie helfen. Und was gerade passiert, weil wir über Regulierung gesprochen haben, wir reden ganz viel darüber, dass uns US-amerikanische Kollegen Vorschläge machen, die aber aus einem anderen Kulturkreis kommen. Oder die dann plötzlich ähnliche Arbeiten aber wieder ganz spannend finden, nur weil ich sie Verhaltensökonomie nenne. Also da müssen wir einfach aufpassen, dass wir nicht einfach sagen, das eine wollen wir nicht, aber wir nennen es anders und dann machen wir es. Generell sehe ich keinen Sinn darin, ein AGI-System zu bauen. Ich glaube, wir dürfen es nicht. Aber als Zielsetzung, "Was ist Intelligenz?", da gibt es viele spannende Fragen, die uns alle helfen.

Was fehlt? Ganz viel. Ich glaube, es fehlt das Gefühl, es fehlt auch das Fingerspitzengefühl. Und damit meine ich, dass wir auch greifen können, auch viel einfacher interagieren können. Und das darf man nicht unterschätzen. Wenn Sie mit Leuten aus der Robotik reden, die haben da sehr viel Respekt davor, wie man wirklich greift und wie man damit dann auch noch Dinge modelliert. Oder denken Sie ans Cello spielen, Violine spielen, und Sie wollen das ja alles mit einem System, was dann jeweils vielleicht noch mal was trainieren muss. Aber es soll ein System sein, weil das alles kann. Ich glaube, da fehlt [viel].

Mathe – ja, ich glaube, Mathe ist toll, ist spannend, aber ist für mich nicht das Wesentliche. Ich finde soziale Normen viel spannender. Wir haben eben viel über Alignment gesprochen – genauso [ist es

mit den] sozialen Normen. Darin [steckt] das eigentlich Schwierige. Weil wir nun mal entscheiden müssen: Wer bestimmt, was die perfekte Welt ist? Darauf wollte ich eben hinweisen. Ich glaube nicht, dass alles nur Lernen ist. Ich glaube, wir können ähnlich wie beim Menschen unterschiedliche Zentren haben. Und ein System kann das andere regulieren. Und dann muss es vielleicht nicht abdriften. Aber da sind ganz große Fragestellungen drin. Ich glaube, es fehlt mehr als das, was wir haben.

Und wir dürfen uns nicht so sehr auf diesen Verkaufs-Hit einlassen. Und Q* – also ganz ehrlich, das meine ich auch nicht böse, ich möchte mich nicht über Algorithmen unterhalten, wo gar keiner weiß, welcher Algorithmus das überhaupt ist. Weil dann möchte ich gerne über Z* reden, mal gucken, was der ist. Wir sollten mal anfangen, über uns und darüber, was wir hier machen, zu reden und darüber, was wir in Europa haben wollen. Dann gerne auch mit den Amerikanern. Da gibt es so viele tolle Forschungskooperationen. Aber ich rede nicht über einen Algorithmus, wo keiner weiß, was der überhaupt ist. Das sind zwei Symbole.

Moderator [00:19:58]

Das ist natürlich ein Konflikt. Es ist natürlich sehr interessant, was das jetzt ist. Das ist gerade für den Journalismus jetzt auch relevant und …

Kristian Kersting [00:20:06]

Ja, aber wie soll ich ihnen denn sagen, was das ist, wenn das Einzige, was veröffentlicht wurde, Q* ist und alle anfangen [zu spekulieren]: "Ist das Q-Learning mit ein bisschen Suche?" Ja, das haben wir auch schon. Das nannte man AlphaZero.

Moderator [00:20:19]

Das stimmt. Aber was wir hier machen, ist ja der Versuch, herunterzubrechen, was kann man sagen, was kann man noch nicht sagen?

Kristian Kersting [00:20:24]

Ja, deswegen sage ich ja, ich rede gerne darüber, was ist AGI, was ist nicht AGI. Aber ich finde, über Q* können wir nicht viel reden, weil es ist nicht mehr als die beiden Symbole. Mehr kann ich einfach nicht sagen. Und was ich eben nur meine, ist, wir können regulieren, wir sollen regulieren, wir müssen nur aufpassen. Es war für mich verwunderlich, dass Elon Musk sechs Monate pausieren wollte, und nach sechs Monaten hat er plötzlich sein eigenes System. Das muss man auch zur Kenntnis nehmen. Neben der Frage, was wir vielleicht sehr viel stärker regulieren sollten.

Moderator [00:20:52]

Holger, du hattest noch eine Anmerkung. Und vielleicht an dich auch noch mal die Frage nach dem Stellenwert von dem mathematischen Können. Das ist ja schon nicht irrelevant.

Holger Hoos [00:21:01]

Genau darauf wollte ich nämlich auch eingehen. Also es ist ja nicht so, dass wir keine KI-Systeme hätten, die mathematische Gleichungen sehr gut lösen könnten oder sehr gut in der Logik sind. Ich meine, wir dürfen nie vergessen: Alles, was wir zurzeit an Informatik Hardware haben, diese ganzen

Chips, auf denen jetzt gerade Zoom läuft, die würden gar nicht existieren, wenn es nicht KI-Methoden gäbe, die im logischen Schließen – was in dem Fall für die Korrektheit dieser Hardware verantwortlich ist – die eben so stark sind, wie sie nun mal sind. Also es gibt KI-Systeme, die in diesem Bereich extrem stark sind, und diese KI-Systeme sind auch in den letzten 20 Jahren gewaltig viel besser geworden. Das ist nur eine Technologie, die bislang keiner irgendwie vernünftig mit großen Sprachmodellen zusammengebracht hat. Es sind einfach andere Systeme.

Und Kristian hat das ganz richtig gesagt: Es ist völlig unklar, warum es ein großer Vorteil sein sollte, diese ganzen Fähigkeiten in ein System zu packen. Es kann manchmal viel besser sein, wenn man spezialisiertere Werkzeuge für bestimmte Zwecke hat. Und es gibt sehr gute Mathe- und Logikwerkzeuge, die für die Gesellschaft auch schon wahnsinnig wertvoll sind. Über die wird nur aus irgendwelchen Gründen nicht so viel geredet. Und manchmal frage ich mich, ob die Gründe nicht was damit zu tun haben, wer diese Forschung macht. Denn in dieser Art von Forschung sind die großen Unternehmen, die ja auch Dinge verkaufen möchten, eben nicht führend. Sondern da sind eher akademische Forschungsgruppen in der Führung, nach wie vor. Das als eine Anmerkung.

Die zweite Anmerkung zum Thema Q*. Ich möchte mich da ganz stark Kristian anschließen. Ich sehe auch im Q&A die Frage danach. Ich würde jetzt mal einfach sagen – es sei denn, Marcus oder Jessica überraschen uns hier ganz doll – niemand, den ich kenne, und ich würde sogar noch weiter gehen, niemand, der nicht in OpenAI genau an diesen Sachen arbeitet, weiß, was dahintersteckt. Und nach allem, was ich sehe, würde ich sagen, ist es erst mal nur Marketing Hot Air und nicht mehr. Um mich davon zu überzeugen, dass da auch nur ein bisschen mehr dahintersteckt, würde ich gerne sehen, was dieser Algorithmus kann und auch, wie er funktionieren sollte. Und da haben wir natürlich ein großes Problem damit, dass diese Art von Spitzenforschung verstärkt nur noch hinter verschlossenen Türen von ein paar großen globalen Firmen stattfindet. Weil nämlich dann genau diese legitimen Fragen, ja auch für die Gesellschaft legitimen Fragen, nicht mehr beantwortet werden können. Und zwar vielleicht auch langfristig nicht mehr beantwortet werden können. Und das ist einer der Gründe, weshalb die derzeitige Entwicklung, gar nicht nur in Bezug auf AGI, extrem gefährlich ist. Weil nämlich immer mehr von der Technologie, die auf die Gesellschaft so oder so einen großen Einfluss hat, hinter verschlossenen und zwar hermetisch verschlossenen Labortüren entwickelt wird. Und da hilft es auch nichts, wenn die LLMs jetzt plötzlich Open Source sind. Denn damit weiß man immer noch nicht genau, wie sie funktionieren, auf welchen Daten sie trainiert wurden und wie sie sich im Ernstfall verhalten. Und genau deswegen brauchen wir nicht nur Open-Source-LLMs und generative KI-Modelle, sondern wir brauchen einen komplett transparenten Prozess, der zu diesen Modellen führt.

Moderator [00:23:58]

Das führt schon fast zur nächsten Frage. Aber noch kurz eine Anmerkung von Marcus.

Marcus Liwicki [00:24:05]

Ich möchte den Bogen ein bisschen schließen. Also erst mal, Q* ist wirklich mehr so ein Marketinggag, weil AlphaStar kam ja schon 2019 von Google DeepMind raus. Ich wollte noch ein bisschen mehr Fundament geben. Holger hat es schon angemerkt, auch ich kenne einige Leute, die in DeepMind und auch in OpenAI arbeiten, und selbst die wissen nicht, wie dieser Algorithmus funktioniert. Ich war ja selbst Coautor damals von den DeepMind-Gründern. Also wir sind da so ein bisschen drin in dieser Community. Und das mit der Mathematik und Logik: Die Sprachmodelle haben schon Entry-Level-Examen in Universitäten bestanden. Das ist nichts grundsätzlich Neues. Wie Holger schon gesagt hat, man kann mathematische Probleme lösen. Wenn es jetzt wirklich kryptografische Probleme sind, dann wär es wirklich noch etwas anderes. Aber wir wollen mal nicht vermuten, was da passiert. Das ist wirklich eine Richtung, wohin wir nicht gehen sollten. Nun noch zurück zur Spitzenforschung. Ich denke, was da passiert, ist keine Spitzenforschung, sondern vor allen Dingen

Spitzenengineering. Denn die Forschung, die Modelle, mit denen sie arbeiten, die werden ja nicht dort neu entwickelt. Das ist mehr ein Zusammenpuzzeln von existierenden Technologien, aber mit wahnsinnig großen Datenmengen. Und dadurch werden die Algorithmen erst richtig stark. Wenn man es vereinfachen will: Was da passiert, sind statistische Modelle, die einfach nur reproduzieren, was sie schon gelesen haben, was sie schon gesehen haben in den Eingangsdaten. Das ist also nichts Neues. Wenn [...] in diesen Daten viele Mathematikexamen drin waren, dann kann das Modell irgendwann auch Mathematik. Aber das ist noch nicht Intelligenz.

Moderator [00:26:00]

Wir haben jetzt noch die generelle Frage, gerade wenn OpenAI eine Blackbox ist, wie oder woran würde man eigentlich erkennen, dass AGI erreicht ist. An dich, Jessica, die Frage: Es ist ja immer der Versuch, die Unternehmen transparenter zu machen, die Ethik da hineinzukriegen, das irgendwie nachzuvollziehen, was die machen. Siehst du da eine Chance oder kriegen wir erst mit, dass, wenn es irgendwann AGI gibt, [...] dann einfach irgendwann Sam Altman in der "New York Times" auftaucht und das erzählt?

Jessica Heesen [00:26:33]

[...] Er kann das einfach erzählen, weil es überhaupt nicht definiert ist, was überhaupt AGI ist. [...] Er hat es selber ja auch gesagt, er wüsste nicht genau, was es ist, und er hat seinen eigenen Begriff davon. Aber ich möchte das gerne einmal ein bisschen herunterbrechen, auch aus meiner [...] philosophisch-ethischen Perspektive. Ich stecke ja nicht so drin in der mathematischen Diskussion oder was überhaupt alles im Detail darin liegt. Aber was ich sehe und was auch immer wieder von OpenAI und anderen Firmen und Entwicklern und Entwicklerinnen gesagt wird, ist, dass es die Befürchtung gibt, dass künstliche Intelligenz diesen Sprung macht in alle möglichen Anwendungsbereiche. Das ist ja der Sinn der Sache. Es geht ja nicht nur um eine generelle Intelligenz, sondern auch um ein Tool und ein Instrument, das für alle möglichen Bereiche anwendbar ist. Und sagen wir mal, das Negativszenario, mit dem viele leben, ist ja, dass KI uns missversteht. Wir müssen leider immer diese menschlichen Begriffe benutzen, [dass es] unbeabsichtigte Nebenfolgen gibt der Programmierung der KI, dass sie sich verselbstständigt. Und ein Beispiel in dem Zusammenhang ist, man sagt einer KI, sie soll bitte ein Mittel erfinden gegen Lungenkrebs. Und dann ist die Konsequenz: Die KI fängt an alle Menschen zu töten, weil wenn es keine Menschen mehr gibt, gibt es auch keinen Lungenkrebs. Solche Szenarien schweben dann ja herum, und da ist es mir absolut schleierhaft, wie das funktionieren soll. Das ist einfach vollkommen unrealistisch und das hat nur etwas mit Science Fiction zu tun. Das möchte ich gerne an der Stelle noch mal sagen und das muss man sich bewusst machen.

Und die eigentliche Gefahr liegt genau darin, in dieser Anthropomorphisierung, dieser Vermenschlichung von KI und das Ganze in dieser monsterartigen Weise darzustellen. Und das hat dann auch gar nicht mehr viel mit Transparenz zu tun. Ich bin sowieso ein bisschen kritisch, was diese ganzen Transparenzwünsche [betrifft]. Wir verstehen die ganze Technologie um uns herum eigentlich überhaupt nicht. Ich kann Ihnen noch nicht einmal so genau erklären, wie die Schreibtischlampe hier auf meinem Tisch funktioniert. Und trotzdem funktioniert sie und ich kriege einen elektrischen Schlag und so weiter. Was wir brauchen, sind sichere Technologien, Technologien, die generell nachprüfbar sind durch Expertinnen und Experten. Wir brauchen standardisierte Verfahren, bevor sie auf den Markt kommen. Aber es muss nicht immer alles für alle transparent sein, sondern das ist ein ganz genau abgestuftes Verfahren von Transparenz. Und das ist natürlich dann schon wichtig.

Moderator [00:28:47]

Und noch mal in den Raum, an die anderen: Seht ihr da noch irgendwelche Möglichkeiten, das früher mitzukriegen? Wie macht ihr das eigentlich, um da jetzt neue Sachen mitzukriegen? Was könnt ihr den Journalistinnen und Journalisten dafür empfehlen? Müssen die jetzt den ganzen Tag auf Arxiv herumsitzen und gucken, was für neue Studien erscheinen? Oder auf Twitter? Oder gibt es da irgendwelche Tipps, die man geben kann, um so etwas früh mitzukriegen? Vielleicht kurz Holger und dann Kristian, und dann haben wir auch schon einige weitere Fragen, wie ich sehe.

Holger Hoos [00:29:15]

Auf gar keinen Fall auf Arxiv, weil da kann ja jeder alles hinstellen und da wird auch eine ganze Menge Unsinn hochgeladen. Über X und Twitter möchte ich mich jetzt überhaupt gar nicht äußern, aber ich denke, wir wissen alle, was man davon zu halten hat. Ganz im Ernst, ich bin ein bisschen old school. Ich glaube, es gibt im Wesentlichen nur eine Methode, das zu erkennen, weil es eigentlich nach wie vor nur eine wirklich operationalisierte Definition von Intelligenz gibt, die aber auch viele, viele Probleme hat und wo man sich auch lange darüber streiten kann. Das ist der Turing-Test. Da gibt es eine lange Literatur zu, die auch sagt, warum das gut wäre oder nicht gut. Aber es ist relativ klar, dass man in dem Augenblick etwas hat, was überzeugend nicht nur nach irgendwelchen oberflächlichen Definitionen bestehen kann. Dann hat man etwas, [...] worüber man sich diese Arten von Gedanken machen sollte. Und nun gibt es ja Leute, die sagen, den Turing-Test bestehen Chatbots schon lange. Da gibt es ausnahmsweise mal gerade ein Papier auf Arxiv, was vor einer Woche ungefähr rausgekommen ist, wo Leute sich das einmal ernsthaft angeguckt haben. Leute, die auch wirklich [kein] Interesse daran haben, die Technologie zu verkaufen, sondern ganz im Gegenteil mit KI-Wissen darangehen, aber das kritisch hinterleuchten, und die kommen zu dem Schluss, dass man nicht nah daran ist, den Turing-Test zu bestehen. Was immer man davon hält, das ist zumindest mal ein Indiz.

Das Zweite ist für mich immer die Sache mit dem autonomen Fahren. Kristian hat das vorhin schon gesagt. Was AGI nach jeder vernünftigen Definition leisten könnte, müsste in der Lage sein, mit der Welt zu interagieren, wie intelligente oder normal intelligente Menschen. Das heißt, insbesondere müsste so etwas in der Lage sein, autonom zu fahren, logischerweise, das kann ja jeder 17-Jährige im Prinzip. Und beim autonomen Fahren sagen mittlerweile alle wirklichen Experten auf dem Gebiet: Wir sind nicht nah dran. In dem Augenblick, wo man vollautonom fahren kann, wo diese ganzen Elon-Musk-Versprechungen wirklich einmal erfüllt sind, da wäre man ein ganzes Stückchen näher dran, weil nämlich autonomes Fahren und die Interaktion, die man dort mit der Welt hat, eine sehr, sehr hohe Latte ist. Das sieht man normalerweise gar nicht. Aber es gibt einen guten Grund, weshalb wir nicht Sechsjährigen erlauben, Auto zu fahren, sondern erst 17-Jährigen oder 16-Jährigen. Und warum auch verstärkt gesagt wird, Leute in den hohen 80ern, die vielleicht nicht nur perzeptuell ein bisschen gefordert sind, sondern wo es auch intellektuell manchmal nicht mehr so ganz [auf der Höhe] ist, die sollten vielleicht auch nicht mehr fahren. Das ist in der Tat gar keine so schlechte Latte. Und da [...] sind wir ganz weit von weg, tatsächlich autonom auf irgendeine befriedigende Art und Weise zu fahren. Das sieht man ja auch daran, dass diese Experimente in San Francisco, die schon eine ganze Weile lang laufen, wo Leute immer sagen, wie super, so langsam nicht mehr so gut aussehen, und [es gibt ja] auch schon die ersten Firmen, die dort jetzt gerade pleitegehen.

Moderator [00:32:04]

Kristian dazu noch.

Kristian Kersting [00:32:06]

Neben dem, was Holger gerade gesagt hat und alle anderen eigentlich auch, glaube ich – wie kriegt man das wirklich hin? Man muss es wirklich akzeptieren. Informatik, KI, mit all den Überschneidungen zu anderen Themengebieten ist eine etablierte Wissenschaft, über die man sich leider auch mit ein bisschen Schweiß informieren muss. In der Physik würden sie diese Frage nicht stellen, sondern würden akzeptieren, dass sie was lesen müssen oder mit vielen Leuten reden müssen. Das müssen wir einfach akzeptieren. Und dann muss man solche Dinge, die wir heute machen, also Science Media Center, einfach weiter vorantreiben. Und deswegen haben wir auch zusammen mit Holger den KI Klub gegründet. Das ist sicherlich eine kleine Sache. Es gibt CLAIRE, es gibt ELLIS. Ich glaube, man muss sich dieses Netzwerk aufbauen, wo man einmal anrufen kann und in privater Atmosphäre auch einmal fragen darf: Sag mal, macht das Sinn oder nicht? Und das kenne ich aus anderen Wissenschaftsbereichen, und das müssen wir hier einfach auch aufbauen.

Und dann noch einmal darauf hingewiesen, weil das auch gerade gesprochen worden ist: Ich weiß auch nicht, ob man eine AGI erkennen würde. Ich würde es glauben, weil irgendjemand würde so bollestolz sein und das mitteilen müssen, wahrscheinlich. Aber ich glaube auch, dass wir aufpassen müssen, auch nicht AGI nur mit dem Menschen zu definieren, sondern es könnte ja auch Systeme geben, die Dinge anders machen als der Mensch. Aber das war klar. Also mit dem Fingerspitzengefühl, was ich meinte, war, dass man mit der Welt interagieren können muss und sie auch manipulieren kann. Manipulieren im Sinne von einfach nur verändern. Und das fehlt den meisten Systemen ganz aktuell. Und dann eben dieses mit kurz antrainiert Auto fahren finde ich auch ein schönes Beispiel. Können wir aber auch mit dem Kochen machen, mit so vielen Dingen, die wir alle können, und das können die nicht. Und wenn ich immer sage, das System kann keine Pizza backen, dann lachen immer alle. Aber ruhig mal länger drüber nachdenken: Es ist gar nicht so einfach, eine Pizza zu backen und das einer Maschine beizubringen.

Moderator [00:33:57]

Gut, Marcus, noch kurz.

Marcus Liwicki [00:34:00]

Ich wollte noch sagen, neben dem Turing-Test, wo ich auch der Meinung bin, der ist noch nicht bestanden. Obwohl ich teile jetzt im Chat zwei Links. Der eine ist ein Nature Paper, was zeigt, dass der Turing-Test bestanden wurde, wobei ein Experte da sogar sagt: Wenn du mich als Experte fragst, würde ich immer erkennen, dass das ein Sprachmodell und keine natürliche Intelligenz ist. Also das Papier ist wie viele andere Nature-Papiere am Ende doch "Rubbish". Und dann gab es das andere Papier von Hutter, das ich noch erwähnen will, das zeigt, da gibt es noch mehr Tests als den Turing-Test. Es gibt einige Tests, die man verwenden kann, um festzustellen, wie und in welcher Perspektive künstliche Intelligenz ist. Wir hatten das ja im Vorhinein mit ein paar E-Mails auch besprochen. Es ist relativ schwierig, und wir haben jetzt noch keine tolle Einstufung, wie weit wir schon sind bei der AGI, und ich bin auch einer von denen, die lieber Applied AI machen und das in speziellen Anwendungsgebieten erforschen, als generell die AGI zu entwickeln. Das ist ein Ziel, was ich nicht habe.

Moderator [00:35:09]

An Jessica die Nachfrage: Die Frage im Chat lautete, wie du die Risiken von AGI für den Medizinbereich einschätzt, ob mögliche Gefahren unterschätzt werden? Und ich würde jetzt noch einmal hinzufügen, wenn du die KI-Leute siehst, die sagen, erst einmal halblang machen, AGI ist jetzt noch nicht um die Ecke, würdest du sagen, das ist vielleicht etwas zu optimistisch? Oder müsste man da

doch etwas vorsichtiger herangehen? Und müsste man, auch wenn die Gefahr gering ist, noch mehr Schranken einbauen und gucken. Oder wie schätzt du generell die Gefahr ein?

Jessica Heesen [00:35:42]

Gerade für den medizinischen Bereich glaube ich, dass AGI da überhaupt keine Rolle spielt, weil es gibt schon wirklich fantastische Möglichkeiten für den Medizinbereich, und die brauchen auf so etwas überhaupt nicht zu setzen. Ich sehe da auch überhaupt keinen Vorteil. Wir haben natürlich in der Diagnostik sehr gute Sachen in Bezug auf KI oder auch in Bezug [auf die] Erleichterung des ärztlichen Alltags und so weiter. Was man allerdings in der Medizin und in der Therapie sieht, das ist im Moment so eine kleine Modeerscheinung, dass man versucht menschenähnliche Avatare zu erstellen, die dann auch wie Ärzte und Ärztinnen aussehen und Auskünfte geben über Therapieformen oder die damit mit Patientinnen und Patienten sprechen und solche Geschichten. Das hat natürlich überhaupt nichts mit AGI zu tun, aber es geht hier um eine andere Form von Menschenähnlichkeit, wo man sich auch noch einmal darüber unterhalten müsste, was daran ethisch problematisch sein könnte. Was war die andere Frage zu AGI?

Moderator [00:36:34]

Ob du generell sehen würdest, da müsste man vorsichtiger sein, auch wenn die Wahrscheinlichkeit gering ist, ob man da noch mehr Sicherheitssachen einführen müsste. Ist natürlich schwierig, aber…

Jessica Heesen [00:36:44]

Man muss sich vor Augen halten, dass es bei dieser ganzen Alignment-Problematik ja um AI Safety geht. Das ist ja eine Unterabteilung von KI-Sicherheitsfragen. Das ist ein altes Thema auch der Ingenieur-Ethik und auch der […] gesetzlichen Haftungsverpflichtung, die man hat gegenüber Produkten, die man erschafft. Und da sehe ich durchaus Risiken. Und deswegen haben wir ja auch in Bezug auf die EU-Regulierung jetzt diesen risikobasierten Ansatz, wo man sagt, man muss noch einmal ganz genau dahin schauen, welche möglichen Schäden so ein KI-System anrichten könnte. Und jetzt nicht unbedingt, weil es menschenähnlich ist, sondern weil es schon eine gewisse Autonomie gerade gibt und es auch außer Kontrolle geraten kann, es Feedbackschleifen geben kann, die man einfach nicht unter Kontrolle hat, wo man Grenzen einziehen muss und so weiter. Man muss schon sehr genau hinschauen, wie diese Systeme arbeiten und dann ist es eine Frage der Produktsicherheit und die ist natürlich enorm wichtig, weil KI auch im Medizinbereich zum Beispiel großen Schaden anrichten kann. Man denke zum Beispiel an Operationsroboter. Aber die sind sehr, sehr stark kontrolliert. Da soll man nicht denken, dass man da jetzt irgendwelche KI-Geschichten wild auf die Menschheit loslässt, sondern im Bereich Mobilität oder eben auch Medizin unterliegen diese Anwendungen sehr starken Regulierungen.

Moderator [00:38:00]

Ich sehe hier auch noch einige Fragen im Chat zum AI Act, noch nicht ganz ein Spoiler, aber höchstwahrscheinlich werden wir am 7. Dezember dazu ein anderes Press Briefing haben. Und da wir jetzt noch einige andere Fragen haben und dann auch Juristinnen und Juristen dabeihaben würden, würde ich jetzt erst einmal die Fragen für den AI Act hintenanstellen. Wenn wir am Ende noch Zeit haben, können wir gerne dazu kommen. Aber ich habe die Befürchtung, das wird nicht mehr so sein. […] Holger, bei dir gab es jetzt zweimal die Nachfrage zum AGI-Verbot. Wie das funktionieren soll und welche Risiken du da siehst, auch wenn es jetzt unwahrscheinlich ist, das zusammenzubringen: Es ist einerseits sehr unwahrscheinlich und wir sind noch nicht da, aber Forschung daran müsste eigentlich reguliert oder verboten werden.

Holger Hoos [00:38:49]

Ja, das ist ein bisschen wie beim menschlichen Klonen. Kristian hat ja vorhin schon darauf hingewiesen. Das ist ja auch verboten, geächtet – [das] würde man in Deutschland niemals gefördert bekommen. Wenn man sich damit beschäftigt, würde man in Deutschland an keiner öffentlichen Einrichtung einen Job bekommen mit menschlichen Klonen und man würde vermutlich auch in keiner Biotechnologiefirma in Deutschland arbeiten können, zumindest nicht, wenn man dann tatsächlich seine Arbeit darauf konzentriert. Und das könnte man natürlich im Prinzip mit AGI-gerichteter Forschung ganz genauso machen. Es ist natürlich eine drastische Sache, ist mir auch völlig klar, nach so etwas zu fragen.

Aber wir müssen uns über Folgendes im Klaren sein: Ich finde es übrigens höchst gefährlich, wenn Leute wie Sam Altman sagen, na ja, AGI, wir wissen alle gar nicht so genau, was das wirklich ist. Ich bin mir sicher, dass es zumindest eine Definition von AGI gibt, früher ja auch gerne als starke künstliche Intelligenz bezeichnet, auf die man sich schon verständigen kann. Und das ist KI, die den Menschen auf der ganzen intellektuellen Bandbreite erreicht oder übertrifft. Und wenn man sich einfach vorstellt, wie das wäre, wenn man solche Systeme hätte, dann wird einem sehr klar, was das für gesellschaftliche Auswirkungen hätte.

Eine Auswirkung wäre, dass die Ungleichheit in der Gesellschaft, die ja schon in vielen Orten in der Welt ganz extrem ist, noch viel extremer würde. Was würde denn Elon Musk machen, wenn er den Roboter, von dem er letztes Jahr gesprochen hat, realisieren könnte? Der Roboter, der in seinen Fabriken jede menschliche Tätigkeit ausführen kann. Wie viele Menschen würden in diesen Fabriken noch arbeiten? Und natürlich gibt es jetzt Leute, die sagen: Wie wunderbar, das löst alle unsere Probleme, weil nicht nur, dass wir keine Arbeit mehr machen müssen und alle nur noch am Strand liegen können, sondern dann wird auch noch die Wissenschaft besser, das Klimaproblem ist gelöst, neue Materialien lösen alle Probleme in der Welt. Selbst unter so einer utopischen Vision, von der ich nicht weiß, warum man daran glauben soll und nicht an wesentlich dunklere Versionen der Zukunft, weil immerhin menschliche Intelligenz hat uns ja auch in eine Situation gebracht, die nicht nur und unbeschränkt positiv ist. Warum sollte Maschinenintelligenz uns automatisch in eine bessere Situation bringen? Aber selbst wenn das nicht der Fall wäre, selbst wenn wir nur den ganzen Tag am Strand liegen könnten … Es ist ja mittlerweile sehr gut bekannt, dass Menschen ihre Erfüllung, ihren Zweck zu einem großen Teil daraus beziehen, wie sie sich nützlich machen im Sinne von Arbeit. Und natürlich gibt es Jobs, die man gerne den Maschinen überlässt, überhaupt gar keine Frage. Aber für die meisten von diesen Jobs braucht man keine AGI. Und wenn man plötzlich alle Jobs von Maschinen erledigen lassen kann, dann haben wir ein gewaltiges Problem als Gesellschaft.

Das nächste Problem ist: Es gibt aus meiner Sicht – und das meine ich jetzt gar nicht spekulativ, sondern rein technologisch gesehen –, überhaupt gar keinen Grund daran zu glauben, dass in dem Augenblick, wo man tatsächlich AGI, also dem Menschen gleichwertig erreicht hat, warum es dann nicht sehr schnell auch weitergehen sollte. Und in dem Fall […] würde ich mir schon gewaltige Sorgen über Kontrollverlust machen. In dem Augenblick, wo man gegen Kontrollverlust schützt, indem man in die Systeme Sicherungen einbaut, muss man sich darüber im Klaren sein. Und da gibt es ja von Hollywood auch mal ausnahmsweise eine relativ gute Darstellung, wie so was schiefgehen kann, dass man mit Intelligenz sehr viele Sicherungen aushebeln kann.

Ich will das jetzt nicht weiter vertiefen, aber noch einmal: Ich denke nicht, dass das Problem ist, dass wir kurz davor stehen, AGI zu erreichen, überhaupt gar nicht. Aber genau wie [man bei bestimmten existenziellen] Risiken, so wie Asteroideneinschlag, nicht anfangen sollte, sie dann ernst zu nehmen, wenn man den Asteroiden, der in drei Jahren einschlägt, entdeckt hat und sich einigermaßen sicher ist. Genau deshalb denke ich, sollte man auch hier vorsichtig sein. Und das ist genau der Grund, weshalb wir auch beim menschlichen Klonen vorsichtig sind, zumindest einer der Gründe.

Moderator [00:42:42]

Wir haben ja noch einige sehr interessante Fragen, deswegen muss ich die Zusatzanmerkungen einmal kurz abbinden. Wir haben noch zweimal die Nachfrage, Kristian an dich, zur generellen Definition von AGI. Wie kann man AGI überhaupt definieren? Meine Hinzufügung wäre jetzt noch, wie würde man sie von Superintelligenz unterscheiden? Und hier gibt es jetzt auch die Frage, ob alle menschlichen Fähigkeiten wie Fingerspitzengefühl et cetera in diesem Fall und bei dieser Definition wichtig sind oder ob nicht auch eine KI denkbar ist, die ganz andere, nicht menschliche Fähigkeiten, eben kein Fingerspitzengefühl hat, aber uns trotzdem in vielen geistigen Bereichen überlegen wäre?

Kristian Kersting [00:43:22]

Man kann sich sicherlich verschiedene Abstufungen überlegen. Wir haben das ja auch schon in der Diskussion mitbekommen, dass keine Einstimmigkeit darüber besteht, was AGI ist. Aber die, die ich kenne und die ich gelesen habe, ist: Man denke sich einer Aufgabe und die Aufgabe kann durch das System gelöst werden. Jetzt kann man sich die nächste Aufgabe überlegen und die kann auch gelöst werden. Man hat diese Utopie, dass es ein System gibt, das egal was, die Aufgabe löst. Für mich folgt daraus, dass es AGI nicht geben kann, auch um den Gegenpol zu Holger zu setzen, weil wir sehen alle, was passiert, wenn wir Intelligenz aufeinander loslassen und es gibt bei vielen Dingen eben nicht die eine richtige Antwort, sogar beweisbar. Es wird aus der Logik sehr wohl klar, dass vielleicht gewisse Dinge wahr oder falsch sind, wir aber den Beweis nicht finden werden. In dem Sinne ist diese Utopie von AGI, wenn damit einhergeht, man kann damit die Wahrheit finden, zum Scheitern verurteilt. Deswegen glaube ich, gibt es eigentlich die Definition, dass AGI bedeutet: Jede Aufgabe, die ein Mensch lösen kann und vielleicht sogar der Intelligenteste, der Beste, was auch immer, da können wir lange drüber reden, was das wieder bedeutet, dass es eigentlich darum geht und da finde ich eben zumindest bei uns schließt das Verbot des Klonens das aus für mich. Aber dazu muss ich dann auch sagen: Klar Holger, wir müssen darüber nachdenken, was passiert, wenn ein Asteroid auf uns stürzt. Aber das Interessante ist, wir diskutieren Asteroideneinschlag nicht gleichwertig mit Klimawandel. Ich finde, was passiert ist durch die Diskussion und die ganzen Briefe, dass wir auf einmal über mögliche AGI genauso wichtig diskutieren wie über Klimawandel und wir können noch andere Themen anführen. Dass wir das mehr und mehr machen, um abzulenken oder Nebelkerzen zu werfen, um ein bisschen den Gegenpol jetzt auch zu spielen, davor habe ich ein bisschen Angst. Ich möchte lieber die aktuellen Probleme auf FAccT, auf Aies, auf AAAI – das sind all diese komischen Konferenzen – angehen und versuchen Beiträge zu bieten und nicht darüber zu diskutieren, was passiert, wenn ein AGI von einem Asteroid auf der Welt landet. Das wollte ich jetzt sagen als Gegenpol, auch um die Diskussion weiterzuführen.

Moderator [00:45:47]

Na ich glaube wir haben kein Problem damit, dass es zu wenig zu diskutieren gibt. Marcus, an dich die Frage aber sonst auch gerne an alle. Es wurde am Rande gesagt, dass vor allem Unternehmen an AGI forschen. Gibt es da keine unabhängigen wissenschaftlichen Projekte zu?

Marcus Liwicki [00:46:04]

Die gibt es auch und es gibt sogar Bücher, die publiziert wurden. Ein sehr bekanntes Buch ist das von Ray Kurzweil von Anfang der 2000er Jahre ["The Singularity Is Near"]. Er hat quasi einen Beweis aufgeführt, dass man 2045 die Artificial Singularity [erreicht] , also diese Superhuman General AI. Dieser Beweis beinhaltet verschiedene Annahmen zum Beispiel, dass ein Exabyte of Data nur 1.000 Euro kostet oder Dollar. Und heutzutage ist man noch nicht einmal im Peta-Bereich bei 500.000 Dollar oder so. Wir sind also sehr, sehr weit weg von den Annahmen, die da gemacht

wurden. Das heißt, diese KI bräuchte eine ganze Menge Energie und ich glaube, da sind auch andere Forscher ....Genevieve Bell, sie ist [unter anderem] bei Intel, sie war erst in Stanford und ist dann aber als Forscherin für die menschlichen Werte zu einer Firma gewechselt und sie hat angemerkt: Wir können uns vielleicht vorstellen, dass es eine Zukunft gibt, wo KIs und Menschen sehr nett miteinander umgehen, vielleicht nicht die Super-KI, aber die Super-KI selbst ist noch so weit weg, weil die Energy Needs so groß sind, dass wir das noch nicht realisieren können. Allein jetzt, wenn ich eine Chatbot, ChatGPT-Anfrage mache, brauche ich zehnmal mehr Energie als wenn ich eine Google-Anfrage mache. Und das wird noch schlimmer, je größer die Modelle in der Zukunft werden.

Moderator [00:47:42]

Und haben die dann einfach nicht so eine gute Publicity wie Open AI oder sind die einfach noch nicht so weit, oder haben die noch nicht genug Compute, um da mithalten zu können?

Marcus Liwicki [00:47:52]

Ich denke, die Forschung wird vielleicht nicht wirklich diese AGI erreichen, weil viele Forscher wissen, wie Holger schon gesagt hat, wir kommen da nicht wirklich hin. Jürgen Schmidhuber, er ist ein sehr berühmter Forscher, der sagt sogar in 2040 haben wir die KI, die Super-KI. Er hat jetzt aber nicht die Forschung selbst gemacht, wie man diese erreicht, sondern einfach nur beweisen wollen, dass man 2040 da sein kann. Aber alle diese Annahmen, die in diesen Papieren gemacht werden, die sind bereits widerlegt. Also wir kommen da nicht so schnell hin.

Kristian Kersting [00:48:27]

Ich wollte nur darauf hinweisen: Kurzweil ist sehr prominent bei Google, also zu sagen, Kurzweil hätte nicht den Zugriff auf Infrastruktur, hätte ich ein bisschen witzig gefunden. Und dann müssen wir dazu sagen, er ist einer der prominentesten Vertreter des technologischen Posthumanismus. Wir sind doch genau da, wo man auch mal drüber nachdenken muss, woher AGI kommt, nämlich von dieser Utopie, dass wir den Menschen in eine neue Bewusstseinsform oder Daseinsform oder was auch immer [bringen]. Da sind ja noch so viel mehr Implikationen drin, insbesondere dann auch irgendwann die Frage: Wer von uns sollte denn überhaupt in dieses neue Paradies eingehen und wer nicht? Da kommen wir jetzt wirklich ganz schnell in diese ganzen komischen Dinge, die ich auch bei AGI-Diskussionen nicht verstehe: Diese Hoffnung, wir bauen uns mal den besten Planeten, aber nur du und du darst mitmachen und ihr da nicht, da wirds dann echt kritisch. Also nur darauf hingewiesen: Wenn wir das jetzt als den Ursprung nehmen und als ein Beispiel, dann müssen wir das auch mit diskutieren.

Moderator [00:49:52]

Holger noch und dann kommen wir zu den Abschlussfragen.

Holger Hoos [00:49:55]

Ich würde mich aus dieser Diskussion lieber gerne rausziehen, weil ich finde, da wird es dann wirklich ein bisschen abstrus. Aber zu dem, Kristian, was du davor gesagt, möchte ich kurz Stellung nehmen. Ich stimme hundertprozentig mit dir überein, dass selbst wenn man sich über Asteroideneinschläge und analoge Dinge Gedanken macht, die eine relativ geringe Wahrscheinlichkeit haben – im Bereich Asteroideneinschlag ist die Wahrscheinlichkeit gering, aber niemand glaubt, dass sie

Null ist – ist das bei AGI ein bisschen anders. Da könnte man vielleicht argumentieren, dass sie philosophisch gesehen tatsächlich sogar Null ist oder auch nicht. Aber ich möchte dir ganz klar zustimmen, dass niemand, der sich über Asteroideneinschläge Gedanken macht, dadurch rechtfertigen sollte, sich nicht um den Klimawandel zu kümmern. Und genau so ist es auch bei AGI, es gibt viel dringlichere Probleme.

Und Bastian, du hast das ja auch am Anfang unserer Diskussion gesagt: Es gibt viel, viel dringlichere Probleme, zum Beispiel all den Schaden, den man anrichten kann mit der Art von KI-Technologie, die wir schon heute haben. Ich selber sehe mit großer Besorgnis den Wahlen, die an vielen Stellen der Welt nächstes Jahr stattfinden [entgegen]. Ich glaube, irgendjemand hat mal ausgerechnet, zwei Milliarden Menschen werden nächstes Jahr wichtige Regierungen wählen und das wird das erste Mal sein, dass generative AI im Wesentlichen in großem Maßstab eingesetzt werden kann – zumindest theoretisch zur Manipulation des politischen Prozesses. Das macht mir wirkliche Sorgen und dazu braucht es keine AGI, dazu braucht es nur das, was wir jetzt gerade schon haben, auf eine bestimmte Art und Weise eingesetzt. Also da kann ich nur zustimmen. Trotz allem bin ich persönlich der Meinung, dass es eine sehr gute Sache ist, dass es eine Task Force gibt, dass es Leute bei der NASA gibt, dass es ernsthafte Astrophysikerinnen und Astrophysiker gibt, die sich über Asteroideneinschlag Gedanken machen. Und genauso finde ich es auch absolut legitim und auch nützlich und weise, wenn es eine gewisse Anzahl von KI-Forscherinnen und -Forscher gibt, die sich über AGI-Risiken Gedanken machen. Das darf aber nicht die öffentliche Debatte dominieren, da stimme ich völlig zu.

Moderator [00:51:54]

Gut, danke. Ich würde sagen, wir kommen jetzt mal schnell zu den Abschlussstatements, und zwar fangen wir mit dir an, Jessica. Die Frage: Was haltet ihr eigentlich in der aktuellen Debatte für den einen wichtigsten Aspekt, den ihr den Journalistinnen und Journalisten jetzt hier noch mit auf den Weg geben wollen würdet?

Jessica Heesen [00:52:12]

Was ich eine bemerkenswerte Beobachtung finde, ist, dass alle denken, wenn etwas intelligent ist, ist es so wie der Mensch. Wir sehen ja, dass wir eigentlich gar nicht so intelligent handeln, sondern es gibt ja auch viele andere intelligente Wesen außer dem Menschen. Es ist noch viel mehr, was zum Menschen dazugehört. Was uns an AGI so Angst macht, ist auch das Bewusstsein, dass wir eben gerade häufig nicht rational arbeiten und uns so eine kalte Rationalität einfach verängstigt, die viel schlauer ist als wir und das konsequent durchzieht. Diese Gleichsetzung von Intelligenz mit dem Menschsein, das ist eine ganz einfache Botschaft. Das geht völlig nach hinten los und ist vollkommen unberechtigt.

Moderator [00:52:50]

Danke. Dann die gleiche Frage an dich, Marcus.

Marcus Liwicki [00:52:55]

Wie wir jetzt schon mehrfach heute festgestellt haben: Wir müssen uns auch um den Klimawandel und Energie kümmern und die KI oder Technik allgemein nimmt schon jetzt circa zehn Prozent des Energiebedarfs der Welt ein. Und das wird viel mehr, vor allen Dingen auch durch KI. Und als kurze Zweitbemerkung: Wir können nicht davon ausgehen, dass irgendeines dieser maschinellen Modelle 100 Prozent perfekt ist. Wir müssen immer damit rechnen, dass da irgendetwas falsch läuft,

dass da irgendetwas von den Ausgaben nicht wirklich verlässlich ist. Also bleibt kritisch, wenn ihr mit der KI redet oder auch wenn ihr Google oder Open AI Artikel lest.

Moderator [00:53:42]

Danke. Genau dann Holger, deine Meinung dazu noch.

Holger Hoos [00:53:47]

Da würde ich mich als allererstes völlig Marcus anschließen und auch Jessica anschließen wollen. Ich glaube, wir wollen gar keine KI, die uns in all unseren Schwächen und Stärken reproduzieren kann und gleich ist. Wir wollen KI, die uns hilft, unsere Schwächen zu erkennen und auszugleichen und die uns die Bereiche, in denen wir stark sind, im Wesentlichen weiterhin überlässt. Das ist meine Definition von menschzentrierter KI und die ist kompatibel mit der der Europäischen Union. Das ist eine gute Sache. Deswegen sehe ich die Weiterentwicklung von ChatGPT, was ja deutlich darauf hin zielt, menschenähnliches Verhalten und Fähigkeiten zu erzeugen, im Grunde genommen als eine Art Sackgasse. Ich finde, wir sollten uns viel stärker darauf konzentrieren, Systeme zu erzeugen, die uns bei der Bewältigung der Probleme unserer Zeit helfen.

Eines dieser Probleme ist natürlich der große Energieverbrauch nicht nur von KI-Systemen, sondern von allen möglichen informationsverarbeitenden Systemen. Da ist die Richtung, die zurzeit verfolgt wird, immer größere Modelle immer aufwendiger zu trainieren, hochgradig problematisch. Ich finde, wir sollten gerade in Europa, aber auch weltweit einen großen Fokus darauf setzen, kleinere Systeme zu bauen, die vielleicht aus kleineren Mengen hochwertigerer Daten bessere Schlüsse ziehen können, aufgrund dieser Daten besser operieren können und uns dabei helfen können, uns den Herausforderungen unserer Zeit zu stellen. Ich würde mir von den Journalistinnen und Journalisten wünschen, dass sie nicht nur die sensationalistischen Meldungen, die aus diesen Großunternehmen kommen, weitertragen und feiern, sondern insbesondere auch Bemühungen, die in diese von mir gerade skizzierte Richtung gehen. Denn das ist es, was unsere Gesellschaft und was die Menschheit viel dringender braucht.

Moderator [00:55:35]

Danke. Und Kristian, dein Abschlusswort noch.

Kristian Kersting [00:55:38]

Ich glaube, nach der ganz großen Definition werde ich zu Lebenszeiten AGI nicht mehr sehen. Ich glaube, das wird keiner so schnell sehen. Erster Punkt. Zweiter Punkt: Trotzdem glaube ich daran, dass es ganz wichtige Entwicklungen sind, die uns helfen werden. Ich glaube daran auch als Investor von Aleph Alpha, das wollte ich noch klarstellen, damit das im Nachgang nicht falsch verstanden wird. Diese Entwicklungen lösen die Probleme nicht alleine, sie werden uns helfen, viele der Probleme angehen zu können, weil ich glaube, beim Klimawandel zum Beispiel gibt es keinen, der den Überblick über all die Disziplinen hat, den wir aber vielleicht brauchen, um hoffentlich doch noch irgendeine Lösung zu finden. Ein Beispiel, es gibt viele andere.

Und dann würde ich zustimmen, wir brauchen nicht die großen Modelle, möchte aber ein Plädoyer dafür halten, dass wir vielleicht an irgendeiner Stelle in Europa einen Ort schaffen, wo offene Modelle, die in der öffentlichen Hand liegen, gebaut werden können. Auch vielleicht größere. Das muss dann die Forschung zeigen, wie wir das hinkriegen. Holger, da stimme ich dir zu, hoffentlich mit weniger Energie. Aber trotzdem [sollten] wir diese Modelle haben. Und dann können wir aus den offenen Modellen entweder durch Testen verstehen, wie es funktioniert, aber auch besser

geschlossene Modelle ableiten. Denn ich möchte darauf hinweisen, wenn alles offen ist, dann wird die Monetarisierung schwierig. Und ich glaube, wir brauchen Monetarisierung in dem aktuellen Wirtschaftssystem. Dafür würde ich gerne plädieren und daher auch den Blick nicht immer nach Amerika richten, sondern auch nach Deutschland, denn wir haben mit CLAIRE und auch mit ELLIS tolle Initiativen in Deutschland, in Europa, die was hinkriegen und gerne auch bei uns bei hessian.AI einmal gucken. Das muss ich jetzt einmal sagen, weil wir solche Modelle aufbauen, auch im Nachtrainieren, also weiter trainieren von existierenden Modellen, sodass man deutsche Sprachmodelle zum Beispiel hat. Und ich glaube, das ist wichtig, damit wir mitreden können.

Und wir müssen nicht alles mitmachen. Wir sind in der Forschung top. Redet also bitte alle mehr auch mit uns allen hier und nicht nur mit den vier, die hier sitzen, sondern es gibt noch viele andere tolle Kolleginnen und Kollegen in Deutschland. Deswegen auch noch mal Danke an das Science Media Center, weil ihr glaube ich echt uns helft an diese Plattform zu bekommen, um zu zeigen, wie gut wir in Deutschland in Europa sind.

Moderator [00:57:49]

Ja gut. Erst mal vielen Dank. Ja, dann ist die Zeit hier jetzt auch schon wieder vorbei. Vielen Dank nochmal allen Journalistinnen und Journalisten für die vielen Fragen. Wir haben jetzt leider nicht alle geschafft, aber es gibt noch andere Briefings, die wir dazu machen werden. Und wie ja gerade auch schon gesagt wurde, gerne einige Leute auch persönlich fragen. Ich glaube, die haben auch immer Interesse daran. Vielen Dank auch an Sie Vier erst mal für Ihre Zeit, an meine Kolleginnen und Kollegen für die Unterstützung. Heute werden wir so schnell wie möglich die Aufzeichnung des Briefings auf unsere Homepage stellen. Das Transkript werden wir voraussichtlich Montag haben. Falls Sie schon schneller eine Videoaufzeichnung einer Audioaufzeichnung oder das maschinell erstellte Transkript haben wollen, finden Sie in der Reminder-Mail von heute Morgen den Link, wo Sie das dann darüber abrufen können. Dann erst mal vielen Dank für Ihre Zeit. Ich wünsche Ihnen später ein schönes Wochenende und auf Wiedersehen.

Ansprechpartner in der Redaktion

Bastian Zimmermann
Redakteur für Digitales und Technologie

Telefon +49 221 8888 25-0
E-Mail redaktion@sciencemediacenter.de

Impressum

Die Science Media Center Germany gGmbH (SMC) liefert Journalisten schnellen Zugang zu Stellungnahmen und Bewertungen von Experten aus der Wissenschaft – vor allem dann, wenn neuartige, ambivalente oder umstrittene Erkenntnisse aus der Wissenschaft Schlagzeilen machen oder wissenschaftliches Wissen helfen kann, aktuelle Ereignisse einzuordnen. Die Gründung geht auf eine Initiative der Wissenschafts-Pressekonferenz e.V. zurück und wurde möglich durch eine Förderzusage der Klaus Tschira Stiftung gGmbH.

Nähere Informationen: **www.sciencemediacenter.de**

Diensteanbieter im Sinne MStV/TMG
Science Media Center Germany gGmbH
Schloss-Wolfsbrunnenweg 33
69118 Heidelberg
Amtsgericht Mannheim
HRB 335493

Redaktionssitz
Science Media Center Germany gGmbH
Rosenstr. 42–44
50678 Köln

Vertretungsberechtigter Geschäftsführer
Volker Stollorz

Verantwortlich für das redaktionelle Angebot (Webmaster) im Sinne des §18 Abs.2 MStV
Volker Stollorz

science
media center
germany