ZUM AUFBEWAHREN, DENN: NIEMAND SOLL SPÄTER SAGEN, MAN HABE ES NICHT GEWUSST!

Download dieser Webseite als PDF: „AI ist gefährlich!“

Vorbemerkung: Ich habe gar nichts gegen die kleinen AI-Anwendungen, die uns das Leben erleichtern.

Mein Kommentar zum nachstehenden Artikel: Was Menschen sich zutrauen, werden sie auch tun. Es ist daher davon auszugehen, dass es Menschen geben wird, die die neuen Technologien missbrauchen werden. Dies gilt zusätzlich für die Möglichkeiten der Quantencomputer, die im folgenden Artikel noch nicht berücksichtigt wurden. Wir werden nicht die Intelligenz erlangen, die es erforderlich macht, eine AI kontrollieren zu können. Dennoch: Wie man an den Abbildungen dieser Seite sieht, arbeite ich gerne mit AI, hier ist es „Midjourney. ;-)

– 08.12.2023, Dieter Wiemkes -

Der nachstehende Artikel, der mit freundlicher Genehmigung des Autors Eliezer Yudkowsky und mit DeepLcom übersetzt wurde, ist im englischen Original bei „Time“ nachzulesen: „Pausing AI Developments Isn't Enough. We Need to Shut it All Down“

„Von Eliezer Yudkowsky, 29. März 2023 6:01 PM EDT

(Yudkowsky ist ein Entscheidungstheoretiker aus den USA und leitet die Forschung am Machine Intelligence Research Institute. Er arbeitet seit 2001 an der Ausrichtung künstlicher allgemeiner Intelligenz und gilt weithin als Begründer dieses Bereichs.)

ES REICHT NICHT AUS, DIE AI-ENTWICKLUNG ZU STOPPEN. WIR MÜSSEN SIE GANZ ABSCHALTEN!

In einem heute veröffentlichten offenen Brief werden 'alle AI-Labore aufgefordert, das Training von AI-Systemen, die leistungsfähiger als GPT-4 sind, sofort für mindestens sechs

Monate zu unterbrechen'.

Dieses 6-monatige Moratorium wäre besser als gar kein Moratorium. Ich habe Respekt vor allen, die sich dafür eingesetzt und es unterzeichnet haben. Es ist eine Verbesserung am Rande.

Ich habe nicht unterschrieben, weil ich denke, dass der Brief den Ernst der Lage unterschätzt und zu wenig fordert, um das Problem zu lösen.

Das Hauptproblem ist nicht die 'konkurrenzfähige' Intelligenz des Menschen (wie es in dem offenen Brief heißt), sondern die Frage, was passiert, wenn die AI intelligenter als der Mensch wird. Die entscheidenden Schwellenwerte sind vielleicht nicht offensichtlich, **wir können definitiv nicht im Voraus berechnen, was wann passiert**, und es scheint derzeit vorstellbar, **dass ein Forschungslabor kritische Grenzen überschreitet, ohne es zu merken**.

Viele Forscher, die sich mit diesen Fragen befassen, darunter auch ich, gehen davon aus, dass das wahrscheinlichste Ergebnis der Entwicklung einer übermenschlich intelligenten AI unter den derzeitigen Umständen darin besteht, dass buchstäblich jeder auf der Erde sterben wird. Nicht im Sinne von "vielleicht eine entfernte Chance", sondern im Sinne von 'das ist das Offensichtliche, was passieren würde'. Es ist nicht so, dass man prinzipiell nicht überleben könnte, wenn man etwas erschafft, das viel intelligenter ist als man selbst; es ist nur so, dass es Präzision und Vorbereitung und neue wissenschaftliche Erkenntnisse erfordern würde, und dass man wahrscheinlich keine AI-Systeme hat, die aus riesigen undurchschaubaren Arrays von Bruchzahlen bestehen.

Ohne diese Präzision und Vorbereitung ist **das wahrscheinlichste Ergebnis eine AI, die nicht das tut, was wir wollen**, und die sich weder um uns noch um empfindungsfähiges Leben im Allgemeinen kümmert. Diese Art von Fürsorge ist etwas, das man einer AI im Prinzip einimpfen könnte, aber **wir sind noch nicht so weit und wissen derzeit nicht, wie**.

Ohne diese Fürsorge heißt es: 'Die AI liebt dich nicht, sie hasst dich auch nicht, und du bist aus Atomen gemacht, die sie für etwas anderes verwenden kann.'

Das wahrscheinliche Ergebnis einer Konfrontation der Menschheit mit einer übermenschlichen Intelligenz ist ein **Totalverlust**. Gültige Metaphern sind 'ein 10-Jähriger, der versucht, gegen Stockfish 15 Schach zu spielen', 'das 11. Jahrhundert, das versucht, gegen das 21. Jahrhundert zu kämpfen' und 'Australopithecus, der versucht, gegen Homo sapiens zu kämpfen'.

Um sich eine feindliche, übermenschliche AI vorzustellen, sollte man sich nicht einen leblosen, bücherschlauen Denker vorstellen, der im Internet haust und böswillige E-Mails

verschickt. Stellen Sie sich eine ganze außerirdische Zivilisation vor, die mit millionenfacher menschlicher Geschwindigkeit denkt und zunächst auf Computer beschränkt ist - in einer Welt voller Kreaturen, die **aus ihrer Sicht sehr dumm und sehr langsam sind**.

Eine ausreichend intelligente AI wird nicht lange auf Computer beschränkt bleiben. In der heutigen Welt kann man DNA-Stränge per E-Mail an Labors schicken, die auf Anfrage Proteine produzieren, so dass eine AI, die zunächst auf das Internet beschränkt ist, künstliche Lebensformen aufbauen oder **direkt zur postbiologischen Molekularproduktion** übergehen kann.

WENN JEMAND UNTER DEN HEUTIGEN BEDINGUNGEN EINE ZU MÄCHTIGE AI BAUT, ERWARTE ICH, DASS JEDES EINZELNE MITGLIED DER MENSCHLICHEN SPEZIES UND ALLES BIOLOGISCHE LEBEN AUF DER ERDE KURZ DARAUF STIRBT.

Es gibt keinen vorgeschlagenen Plan, wie wir so etwas tun und überleben könnten. Die offen erklärte Absicht von OpenAI ist es, eine zukünftige AI unsere Hausaufgaben in Sachen AI-Ausrichtung erledigen zu lassen. Allein zu hören, dass dies der Plan ist, sollte ausreichen, um **jeden vernünftigen Menschen in Panik zu versetzen**. Das andere führende AI-Labor, DeepMind, hat überhaupt keinen Plan.

Eine Anmerkung am Rande: Nichts von dieser Gefahr hängt davon ab, ob AI ein Bewusstsein hat oder haben kann; sie liegt in der Vorstellung von leistungsfähigen kognitiven Systemen, die hart optimieren und Ergebnisse berechnen, die hinreichend komplizierte Ergebniskriterien erfüllen. Abgesehen davon würde ich meine moralischen Pflichten als Mensch vernachlässigen, wenn ich nicht auch erwähnen würde, dass wir

keine Ahnung haben, wie wir feststellen können, ob AI-Systeme **sich ihrer selbst bewusst** sind - da wir **keine Ahnung** haben, wie wir alles entschlüsseln können, was in den riesigen undurchschaubaren Arrays vor sich geht -, und dass wir daher irgendwann unbeabsichtigt digitale Köpfe erschaffen könnten, die wirklich ein Bewusstsein haben und Rechte haben sollten und nicht besessen werden sollten.

Die Regel, die die meisten Menschen, die sich dieser Problematik bewusst sind, vor 50 Jahren befürwortet hätten, lautete: Wenn ein AI-System fließend sprechen kann und sagt, dass es sich seiner selbst bewusst ist und Menschenrechte einfordert, dann sollte das ein harter Stopp für Menschen sein, die diese AI einfach so besitzen und sie über diesen Punkt hinaus nutzen. **Diese alte Grenze haben wir bereits überschritten.** Und das war wahrscheinlich richtig; ich stimme zu, dass die aktuellen AIs wahrscheinlich nur das Gerede von Selbstbewusstsein aus ihren Trainingsdaten imitieren. Aber ich gebe zu bedenken, dass wir angesichts des geringen Einblicks, den wir in das Innere dieser Systeme haben, nicht wirklich etwas wissen.

Wenn das der Stand unserer Unkenntnis für GPT-4 ist und GPT-5 ein ebenso großer Sprung in den Fähigkeiten ist wie von GPT-3 zu GPT-4, dann können wir nicht mehr mit Fug und Recht sagen: 'Wahrscheinlich nicht selbstbewusst', wenn wir die Leute GPT-5 bauen lassen. Dann heißt es nur noch **'ich weiß es nicht; niemand weiß es'.** Wenn man sich nicht sicher sein kann, ob man eine selbstbewusste AI erschafft, ist das nicht nur wegen der moralischen Implikationen des 'selbstbewussten' Teils alarmierend, sondern auch, weil Ungewissheit bedeutet, **dass man keine Ahnung hat, was man tut**, und das ist gefährlich und man sollte damit aufhören.

Am 7. Februar freute sich Satya Nadella, CEO von Microsoft, öffentlich darüber, dass das neue Bing Google dazu bringen würde, 'zu zeigen, dass sie tanzen können'. 'Ich möchte, dass die Leute wissen, dass wir sie zum Tanzen gebracht haben', sagte er.

So redet ein CEO von Microsoft nicht in einer normalen Welt. Es zeigt, wie ernst wir das Problem nehmen, und wie ernst wir es schon vor 30 Jahren hätten nehmen müssen.

Wir werden diese Lücke nicht in sechs Monaten schließen.

Es hat mehr als 60 Jahre gedauert, bis das Konzept der künstlichen Intelligenz zum ersten Mal vorgeschlagen und untersucht wurde und wir die heutigen Fähigkeiten erreicht haben. Die Lösung des Sicherheitsproblems bei übermenschlicher Intelligenz - nicht die perfekte Sicherheit, sondern die Sicherheit im Sinne von 'nicht buchstäblich jeden umbringen' - könnte mindestens halb so lange dauern. Das Problem bei diesem Versuch

mit übermenschlicher Intelligenz ist, **dass man nicht aus seinen Fehlern lernen kann, wenn es beim ersten Versuch schief geht, weil man dann tot ist**. Die Menschheit lernt nicht aus dem Fehler und kann sich nicht wieder aufrappeln und es erneut versuchen, wie bei anderen Herausforderungen, die wir in unserer Geschichte bewältigt haben, denn dann sind wir alle tot.

Der Versuch, irgendetwas beim ersten wirklich kritischen Versuch richtig zu machen, ist eine außergewöhnliche Herausforderung, sowohl in der Wissenschaft als auch in der Technik. **Wir haben nicht annähernd den Ansatz, der erforderlich wäre, um dies erfolgreich zu tun.** Würden wir auf dem im Entstehen begriffenen Gebiet der allgemeinen künstlichen Intelligenz weniger strenge technische Maßstäbe anlegen als bei einer Brücke, die ein paar tausend Autos tragen soll, wäre das gesamte Gebiet schon morgen stillgelegt.

WIR SIND NICHT VORBEREITET.

Wir sind nicht auf dem Weg, in einem vernünftigen Zeitfenster vorbereitet zu sein. **Es gibt keinen Plan.** Die Fortschritte bei den AI-Fähigkeiten liegen weit, weit vor den Fortschritten bei der AI-Anpassung oder sogar vor den Fortschritten beim Verständnis dessen, was zum Teufel in diesen Systemen vor sich geht. Wenn wir dies tatsächlich tun, werden wir alle sterben.

Viele Forscher, die sich mit diesen Systemen befassen, sind der Meinung, **dass wir auf eine Katastrophe zusteuern**, wobei sich mehr von ihnen trauen, dies privat zu sagen als in der Öffentlichkeit; aber sie sind der Meinung, dass sie den Vorwärtsdrang nicht einseitig aufhalten können, dass andere weitermachen werden, selbst wenn sie persönlich ihren Job aufgeben. Und so denken sie alle, sie könnten genauso gut weitermachen. Das ist ein dummer Zustand und eine unwürdige Art und Weise, die Erde sterben zu lassen, und der

Rest der Menschheit sollte an diesem Punkt eingreifen und der Industrie helfen, ihr **kollektives Handlungsproblem** zu lösen.

Einige meiner Freunde haben mir kürzlich berichtet, dass Menschen außerhalb der AI-Branche, wenn sie zum ersten Mal vom Aussterberisiko der Künstlichen Allgemeinen Intelligenz hören, mit **'vielleicht sollten wir dann keine KAI bauen'** reagieren.

Das gab mir einen kleinen Hoffnungsschimmer, denn es ist eine einfachere, vernünftigere und offen gesagt gesündere Reaktion als die, die ich in den letzten 20 Jahren gehört habe, als ich versucht habe, jemanden in der Branche dazu zu bringen, die Dinge ernst zu nehmen. Jeder, der so vernünftig redet, hat es verdient zu hören, **wie schlimm die Situation tatsächlich ist**, und nicht, dass man ihm sagt, ein sechsmonatiges Moratorium würde das Problem lösen.

Am 16. März schickte mir meine Partnerin diese E-Mail. (Sie gab mir später die Erlaubnis, sie hier auszugsweise wiederzugeben).

'Nina hat einen Zahn verloren! Und zwar auf die übliche Art und Weise, wie es Kinder tun, nicht aus Unachtsamkeit! Die Tatsache, dass GPT4 die standardisierten Tests am selben Tag, an dem Nina einen Meilenstein in der Kindheit erreichte, einfach wegpustete, hat mich für eine Minute aus dem Konzept gebracht. Es geht alles viel zu schnell. Ich befürchte, dass es Ihre Trauer noch verstärkt, wenn ich Ihnen davon erzähle, aber es ist mir lieber, dass Sie mich kennen, als dass jeder von uns allein leiden muss.'

Wenn es in dem Insider-Gespräch um die Trauer darüber geht, dass die eigene Tochter ihren ersten Zahn verliert, und um den Gedanken, dass sie keine Chance bekommt, erwachsen zu werden, dann glaube ich, dass wir über den Punkt hinaus sind, an dem wir politisches Schach über ein sechsmonatiges Moratorium spielen.

Wenn es einen Plan für das Überleben der Erde gäbe, wenn wir nur ein sechsmonatiges Moratorium verabschieden würden, würde ich diesen Plan unterstützen. **Einen solchen Plan gibt es nicht.**

(Anm.: Wenn es schlecht läuft, dann kann AI uns allen **alles** nehmen. Die gesamte Menschheit ist in Gefahr! - Dieter Wiemkes)

HIER IST, WAS TATSÄCHLICH GETAN WERDEN MÜSSTE:

Das Moratorium für neue große Trainingsläufe muss **unbefristet und weltweit** gelten. Es darf keine Ausnahmen geben, auch nicht für Regierungen oder Militärs. Wenn die Politik von den USA ausgeht, dann muss China erkennen, dass die USA nicht nach einem Vorteil streben, sondern **versuchen, eine schrecklich gefährliche Technologie zu verhindern**, die keinen wirklichen Besitzer haben kann und die jeden in den USA, in China und auf der Erde töten wird. Wenn ich unendlich viel Freiheit hätte, Gesetze zu schreiben, würde ich vielleicht eine einzige Ausnahme für AI machen, die ausschließlich für die Lösung von Problemen in der Biologie und Biotechnologie trainiert wird und nicht auf Texte aus dem Internet trainiert wird und nicht so weit geht, dass sie anfängt zu sprechen oder zu planen; aber wenn das die Sache auch nur im Entferntesten verkomplizieren würde, würde ich diesen Vorschlag sofort verwerfen und sagen, **dass wir einfach alles abschalten sollten**.

Schalten Sie alle großen GPU-Cluster ab (die großen Computerfarmen, auf denen die leistungsfähigsten AIs weiterentwickelt werden). Schalten Sie alle großen Trainingsläufe ab. Legen Sie eine **Obergrenze für die Rechenleistung** fest, die jeder für das Training eines AI-Systems verwenden darf, und senken Sie diese in den kommenden Jahren, um effizientere Trainingsalgorithmen zu ermöglichen. Keine Ausnahmen für Regierungen und Militärs. Sofortige multinationale Vereinbarungen, um zu verhindern, dass die verbotenen Aktivitäten in andere Länder verlagert werden. Verfolgen Sie alle verkauften GPUs. Wenn Geheimdienstinformationen besagen, dass ein Land außerhalb des Abkommens einen GPU-Cluster baut, sollten Sie weniger Angst vor einem Schusswechsel zwischen den

Nationen haben als vor einer Verletzung des Moratoriums; seien Sie bereit, ein abtrünniges Rechenzentrum durch einen Luftangriff zu zerstören.

Stellen Sie nichts als einen Konflikt zwischen nationalen Interessen dar, machen Sie klar, dass jeder, der von Wettrüsten spricht, ein Narr ist. **Dass wir alle als Einheit leben oder sterben, ist keine Politik, sondern eine Tatsache der Natur.** Machen Sie in der internationalen Diplomatie deutlich, dass die Verhinderung von AI-Auslöschungsszenarien Vorrang vor der Verhinderung eines vollständigen nuklearen Austauschs hat, und dass verbündete Nuklearländer bereit sind, ein gewisses Risiko eines nuklearen Austauschs einzugehen, wenn dies nötig ist, um das Risiko großer AI-Trainingsläufe zu verringern.

Das ist die Art von Politikänderung, die meinen Partner und mich dazu veranlassen würde, uns in den Arm zu nehmen und uns zu sagen, dass ein Wunder geschehen ist und es jetzt eine Chance gibt, dass Nina vielleicht leben wird. Die vernünftigen Menschen, die zum ersten Mal davon hören und vernünftigerweise sagen: **'Vielleicht sollten wir das nicht tun'**, verdienen es, ehrlich zu hören, was nötig wäre, damit das passiert. Und wenn die politische Forderung so groß ist, kann sie nur durchgesetzt werden, wenn die Entscheidungsträger erkennen, dass ihre eigenen AInder auch sterben werden, wenn sie so weitermachen wie bisher und das tun, was politisch einfach ist.

SCHALTEN SIE ALLES AB.

Wir sind **nicht bereit**. Wir sind **nicht auf dem besten Weg**, in absehbarer Zeit wesentlich besser vorbereitet zu sein. Wenn wir so weitermachen, werden alle sterben, auch Kinder, die sich das nicht ausgesucht und nichts falsch gemacht haben.

BEENDEN SIE ES.“

WEITERFÜHRENDE LINKS (WERDEN BEI GRAVIERENDEN NEUIGKEITEN ERGÄNZT):

„Sam Altman vor dem US-Senat über künstliche Intelligenz: «Die Sache kann völlig schiefgehen»“ (NZZ)

„NASA Just Shut Down Quantum Computer After Something Insane Happened!“ (YouTube)

„Künstliche Intelligenz droht im Gespräch, die Menschen auszuschalten“ (Der Standard)

„Chatbot LaMDA - Künstliche Intelligenz von Google soll Bewusstsein entwickelt haben“ (Forschung und Wissen)

„Google-KI mit „Bewusstsein“ soll rassistisch sein“ (Forschung und Wissen)

„Hat künstliche Intelligenz wie ChatGPT ein Bewusstsein“ (NZZ)

„Unser Gehirn funktioniert wie Chat-GPT“ (NZZ)

„Umstrittene KI von Google hat Anwalt eingeschaltet“ (Forschung und Wissen)

„Der Open-AI-Chef warnt vor der KI der Zukunft“ (NZZ)