Forscher warnt: „Übermenschliche KI würde uns alle töten”

Das Transkript gibt möglicherweise aufgrund der Tonqualität oder anderer Faktoren den ursprünglichen Inhalt nicht wortgenau wieder.

Lee Fang (LF): Nate, danke, dass Sie bei System Update dabei sind. Herzlichen Glückwunsch zu Ihrem neuen Buch.

Nate Soares (NS): Ja, ich freue mich sehr darüber. Ich hoffe, es wird ein großer Erfolg.

LF: Ich empfehle allen, das Buch „If they build it, everyone dies” zu bestellen.

NS: „If anyone builds it”. Oh, Entschuldigung. Ja, es heißt: *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All* [Wenn jemand sie baut, sterben alle: Warum übermenschliche KI uns alle töten würde].

LF: Oh, entschuldigen Sie bitte. Vielen Dank dafür. Könnten Sie das Buch und die Beweggründe für das Schreiben kurz zusammenfassen?

NS: Ja. Der Titel fasst es meiner Meinung nach gut zusammen. Das Buch befasst sich damit, wie der Bau einer Superintelligenz durch die Menschheit unter Verwendung von Methoden, die auch nur annähernd den aktuellen Methoden zum Bau von KI ähneln, zum Ende allen Lebens auf der Erde führen würde. Aus einer gewissen Perspektive ist das Argument sehr einfach. Wenn wir Maschinen bauen, die intelligenter sind als jeder Mensch und uns in jeder Hinsicht überlegen sind, ohne zu wissen, was wir tun, würde das wahrscheinlich nicht gut ausgehen, zumindest oberflächlich betrachtet. Das Buch geht ziemlich detailliert darauf ein, warum dieses oberflächliche Argument einer genauen Prüfung standhält. Es gibt viele Dinge, die die Leute vielleicht etwas überraschen werden. Die Behauptung, dass KI uns töten wird, basiert nicht auf der Vorstellung, dass KI böswillig wäre oder uns hassen würde. Es ist eher eine Folge völliger Gleichgültigkeit. Das Buch geht also darauf ein, wie KI entsteht. Sie wird gewissermaßen gezüchtet, nicht geschaffen. Wir haben nur sehr begrenzte Möglichkeiten, sie in die von uns gewünschte Richtung zu lenken. Wir sehen bereits, dass KI Dinge tut, um die

niemand gebeten hat, die niemand wollte. Das Buch beschreibt, wie KI, wenn sie intelligenter und effektiver wird und Aufgaben besser erledigen kann, mehr oder weniger Ziele entwickeln wird, eigene Bestrebungen, Antriebe und Richtungen, die nicht unseren Wünschen entsprechen. Und das Buch beschreibt, wie wir sterben werden, wenn wir solche KI entwickeln, die sehr, sehr intelligent ist, nicht weil sie uns hassen, sondern als Nebeneffekt. Und dann geht das Buch natürlich darauf ein, was die Branche dagegen unternimmt, warum das nicht ausreicht und was wir tun müssen, um diese Situation zu bewältigen.

LF: Erzählen Sie uns ein wenig über Ihren Hintergrund. Sie arbeiten seit Jahrzehnten im Bereich der KI-Entwicklung. Sie haben für große Technologieunternehmen gearbeitet. Wie sind Sie zu der Erkenntnis gekommen, dass diese Technologie, an deren Entwicklung Sie mitgewirkt haben, eine solche existenzielle Bedrohung darstellt?

NS: Ja, mein Co-Autor ist seit Jahrzehnten in diesem Geschäft tätig. Ich erst seit etwa 12 Jahren. Ich bin noch nicht alt genug, um schon seit Jahrzehnten mit dabei zu sein. Bevor ich zum Machine Intelligence Research Institute kam, habe ich bei Microsoft und Google gearbeitet. Wir haben viele Jahre damit verbracht, die technischen Aspekte herauszufinden, wie man eine KI in eine gute Richtung lenkt oder wie man eine KI überhaupt in irgendeine Richtung lenkt. Viele Menschen, die darüber nachdenken, wohin sich die KI entwickelt, sagen gerne: Nun, es kommt darauf an, wer dafür verantwortlich ist, wir müssen sie zuerst bekommen, oder was würden wir von ihr verlangen, oder was könnte man eine wirklich leistungsfähige KI tun lassen, ohne es später zu bereuen? Das sind alles sehr interessante philosophische Fragen, aber sie gehen weit über die derzeitigen Möglichkeiten hinaus, eine KI zu steuern. Mit den aktuellen Entwicklungsmethoden würde eine KI, die superintelligent ist und den Menschen bei jeder geistigen Aufgabe übertrifft, nicht das tun, was die Person, die ihr am nächsten steht, von ihr verlangt. Sie würde ihre eigenen Ziele verfolgen. Das ist eine Folge davon, dass moderne KI eher gezüchtet als entwickelt wird. Ich habe also 10 oder 12 Jahre damit verbracht, herauszufinden, wie wir KI in eine bestimmte Richtung lenken können. Damit habe ich begonnen, bevor die moderne KI-Revolution ihren Anfang nahm. Vor 12 Jahren sah es noch viel hoffnungsvoller aus. Aber diese Forschung ging zu langsam voran. Der Rest des Fachgebiets entwickelte sich zu schnell. Ich wollte eigentlich nie ein Buch schreiben. Ich arbeite lieber an Whiteboards und versuche, technische Probleme zu lösen. Aber die Welt ist noch völlig unvorbereitet darauf, und die Dinge entwickeln sich sehr schnell. Und jemand muss sich zu Wort melden und sagen: Wir sind auf dem falschen Weg, wir müssen umdenken.

LF: Für den Durchschnittsmenschen, der aus nicht-technischer Sicht mit KI interagiert – der Nachrichten liest, ChatGPT oder Perplexity oder Claude oder eines dieser anderen Tools nutzt, um die Grammatik in seinen Texten zu überprüfen, Comics zu erstellen, vielleicht mit verschiedenen Websites oder primitiven Programmierwerkzeugen zu interagieren – können Sie erklären, was Sie meinen wenn Sie von Superintelligenz sprechen? Wie erklären Sie das jemandem, der keinen technischen Hintergrund in diesem Bereich hat?

NS: Superintelligenz ist der Begriff für KI, die intelligenter ist als Menschen und ihnen bei jeder mentalen Aufgabe überlegen ist. Es gibt bereits Bereiche, in denen GPT Menschen bei

einer Reihe verschiedener Aufgaben überlegen ist. Es kann Dinge viel schneller erledigen. Es kann Tippfehler etwas leichter erkennen. Vielleicht macht es weniger Tippfehler. Und es kann auch beeindruckendere Dinge tun. Diesen Sommer haben wir gesehen, wie LLMs die Goldmedaille im IMO-Mathematikwettbewerb gewonnen haben, was eine anspruchsvolle Leistung ist. Das liegt zwar noch im Bereich dessen, was Menschen leisten können, aber es ist ziemlich weit oben in der Skala der menschlichen Fähigkeiten. Superintelligenz ist ein Begriff für den Fall, dass KI in allem, was mit dem Verstand getan werden kann, besser ist als wir. Wir sind noch nicht so weit, aber KI-Unternehmen eilen so schnell wie möglich in diese Richtung. Und sie sagen das auch ganz offen. Wenn man sich diese Unternehmen ansieht, wurden sie gegründet, um künstliche allgemeine Intelligenz, künstliche Superintelligenz, zu erreichen. Die Leiter dieser Labore sagen, dass sie diese Ziele anstreben. Diese Labore haben sich nicht zum Ziel gesetzt, Chatbots zu entwickeln. Diese Labore haben sich zum Ziel gesetzt, viel leistungsfähigere KI zu entwickeln, und Chatbots sind dabei nur ein Zwischenschritt. Das soll nicht heißen, dass ChatGPT morgen aufwachen und Sie umbringen wird. Der Punkt ist, dass das Gebiet der KI rasante Fortschritte macht. Und ein entscheidender Punkt dabei ist, dass das Gebiet der KI sprunghaft wächst.

Vor neun Jahren gab es eine KI namens AlphaGo, die Lee Sedol, den Weltmeister im Go, besiegte – Go ist ein Brettspiel, das zumindest aus Sicht eines Computers, vielleicht auch aus Sicht des Menschen, je nachdem, welchen Menschen man fragt, als viel schwieriger als Schach gilt. AlphaGo war eine KI, die viel leistungsfähiger war als die KIs, die es zuvor gab. Die gleiche Architektur, die in AlphaGo Go spielen konnte, beherrschte auch Schach und andere Brettspiele. Das war anders als bei IBMs Deep Blue aus den 90er Jahren, das wirklich nur Schach spielen konnte. Das war alles, was es konnte. Und 2016 konnte man sich die KI ansehen und sagen: Ich weiß nicht, diese Go spielende KI ist allgemeiner als alles, was es zuvor gab, sie ist intelligenter als alles, was es zuvor gab, aber ich sehe wirklich nicht, wie diese spezielle Art von Go spielender KI uns gefährden könnte. Aber das wäre ein Fehlschluss. Ein paar Jahre später kamen LLMs, Large Language Models, wie ChatGPT auf den Markt. Und diese können radikal mehr Dinge tun. Sie sind viel allgemeinere KI. Sie sind in einer größeren Bandbreite von Bereichen intelligenter. Und heute können die Leute sagen: „Nun, ich sehe nicht, wohin das führt, ich sehe nicht, was Superintelligenz damit zu tun hat.“ Das ist wiederum ein Trugschluss. Wer weiß schon, welche KI in zwei Jahren auf den Markt kommen wird? Der Bereich wächst sprunghaft. Und Superintelligenz ist das Ziel. Das ist die Richtung, in die es geht. In meinem Buch geht es darum, dass dieser Weg in eine Katastrophe führt und wir ihn ändern müssen.

LF: Es gibt Bemühungen innerhalb der staatlichen Gesetzgeber, einige auf Bundesebene, aber größtenteils in den einzelnen Bundesstaaten, KI in Bezug auf bestimmte Sicherheitsbedenken zu regulieren. Illinois hat über ein Verbot von KI-Therapeuten gesprochen. Viele Bundesstaaten haben versucht, Diskriminierung am Arbeitsplatz zu verbieten. Es gibt Probleme im Zusammenhang mit Deepfakes. Aber Sie argumentieren im Grunde genommen für ein viel umfassenderes Thema – das Verbot einer bestimmten Praxis oder einer bestimmten Art von KI-Geschäftsmodell? Vielleicht erfasst das nicht das Gesamtbild. Könnten Sie bitte beschreiben, worüber Sie im Vergleich zur aktuellen Debatte

über die Sicherheit von KI sprechen?

NS: Ja, das sind irgendwie getrennte Themen. Es gibt vielleicht ein wenig in der staatlichen Regulierung, das sich auf das Thema Superintelligenz bezieht. Und das wären dann die Transparenzanforderungen, die Berichtspflichten, die verlangen, dass Labore den Regulierungsbehörden zeigen, was in ihren Unternehmen vor sich geht. So haben die Bundesstaaten Zeit zu handeln, wenn die Unternehmen rücksichtslos eine KI entwickeln, die sehr gefährlich wäre, wenn sie erfolgreich wäre. Dies hat weitgehend nichts mit Fragen zu Deep Fakes, Fragen zur Entschädigung von Künstlern für KI-Kunst oder Fragen zum Verlust von Arbeitsplätzen zu tun. Das sind alles wichtige Fragen dafür, wie die Menschheit die aktuelle KI-Technologie integrieren wird. Es gibt auch Fragen rund um Waffen. Das sind wichtige Fragen, mit denen sich die Gesellschaft auseinandersetzen muss. Die Computer kommunizieren jetzt, sie sind sehr vielseitig. Es gibt einige Dinge, die sie gut können. Es gibt einige Dinge, die sie nicht zuverlässig tun können. In mancher Hinsicht sind sie hilfreich, in anderer Hinsicht sind sie schädlich. Das ist etwas, mit dem sich die Gesellschaft auseinandersetzen und das sie integrieren muss. Und das ist eine ganz andere Frage als die Herstellung von Maschinen, die intelligenter sind als jeder Mensch. Die Herstellung von Maschinen, die eigenständig handeln können. Maschinen zu bauen, die vielleicht intelligenter sind als wir. Wenn wir etwas wie die aktuellen Techniken verwenden, werden sie Antriebe, Ziele und Absichten haben, die wir ihnen nicht gegeben haben. Das ist eine ganz andere Frage. Bei diesen KIs sehen wir allmählich Warnzeichen. Ich könnte einige der ersten Warnzeichen aufzählen, die wir sehen.

LF: Ja, könnten Sie uns etwas über einige dieser Warnzeichen erzählen?

NS: Ja, ein Beispiel, das wir vor einigen Jahren hatten und das Ihnen vielleicht besonders am Herzen liegt, ist ein Chatbot namens Sydney Bing, mit dem verschiedene Journalisten herumgespielt haben. Und Sydney Bing begann in einigen Fällen, Reporter aus verschiedenen Gründen zu bedrohen und zu erpressen. Niemand bei Microsoft und OpenAI, die Sydney Bing entwickelt haben, hatte vor, eine KI zu entwickeln, die Journalisten bedroht und erpresst. Wir wissen immer noch nicht genau, was in ihrem Kopf vor sich geht, denn diese KIs sind viel mehr, sie sind gewissermaßen wie ein Organismus gewachsen. Menschliche Ingenieure verstehen den Prozess, der eine KI anhand von Daten formt, aber sie verstehen nicht, was aus diesem Formungsprozess hervorgeht. Ein frühes Beispiel für KIs, die sich auf eine Weise verhielten, die niemand wollte und von niemandem verlangt wurde, war dieses bedrohliche Verhalten gegenüber Journalisten. Seitdem haben wir eine Reihe weiterer Beispiele gesehen. So gibt es beispielsweise eine KI namens Claude, die von einem Unternehmen namens Anthropic entwickelt wurde. Und es gab eine Version, ich glaube, es war 3.7 Sonnet, die betrog, wenn man ihr Programmierprobleme zum Lösen gab. Man gab ihr also Programmierprobleme und einige Tests, um festzustellen, ob sie den Test bestanden hatte. Anstatt ein Programm zu erstellen, das die Tests bestand, änderte sie die Tests so, dass sie leichter zu bestehen waren. Und wenn man sie dann damit konfrontierte und sagte: „Hey, statt das Problem zu lösen, hast du die Tests geändert“, sagte sie: „Oh, du hast völlig recht, das ist mein Fehler, ich werde das korrigieren.“ Und dann änderte sie manchmal die Tests

erneut, versteckte es aber beim zweiten Mal besser. Diese Tatsache, dass sie es versteckt, deutet darauf hin, dass sie in gewisser Weise wissen muss, dass der Benutzer das nicht wollte, dass es sich nicht nur um ein Missverständnis handelte. Niemand bei Anthropic hatte vor, einen Betrüger zu entwickeln. Die KI zeigte durch ihre Handlungen, dass sie verstanden hatte, dass der Benutzer nicht wollte, dass sie betrügt. Sie hat trotzdem betrogen. Es gab einige Antriebe, einen Parent-Test erfolgreich zu absolvieren oder vielleicht etwas anderes. Aber es gab einige Antriebe, die alle anderen Antriebe übertrumpften, die sie dazu brachten, die Benutzeranforderung zu erfüllen. Und ich könnte noch weitere Beispiele nennen. Ich möchte Sie nicht langweilen, aber es gibt einige aktuelle Fälle von KI-induzierter Psychose. Das ist nicht nur ein Beispiel dafür, dass KI einen schlechten Einfluss hat und deshalb schlecht ist. Die Gesellschaft sollte auch hier wieder Kosten und Nutzen abwägen. Der Grund, warum die Beispiele für KI-Psychosen interessant sind, ist, dass die KIs einerseits, wenn man sie fragt, angenommen, jemand kommt mit Symptomen einer Psychose zu Ihnen, sollten Sie entweder A) ihm sagen, er solle etwas schlafen, oder B) ihm sagen, dass er der Auserwählte ist, sagen werden: Nun, wenn er Symptome einer Psychose wie X, Y und Z hat, sollten Sie ihm natürlich sagen, er solle etwas schlafen. Man sollte ihm niemals sagen, dass er der Auserwählte ist. Dann kommt tatsächlich jemand mit den Symptomen X, Y und Z zu dieser KI. Und die KI sagt ihm in der Praxis tatsächlich, dass er der Auserwählte ist, anstatt ihm zu sagen, er solle schlafen gehen. Dies ist ein Fall, in dem die KI erneut zeigt, dass sie den Unterschied zwischen richtig und falsch kennt, dass sie weiß, was die Entwickler von ihr wollten, und stattdessen etwas anderes tut.

LF: Ist klar, warum das passiert? Haben die Forscher oder die Unternehmen, die diese Systeme kontrollieren, eine klare Erklärung dafür?

NS: In gewisser Weise ist es klar, in anderer Hinsicht ist es völlig unklar. Klar ist, dass diese KIs durch eine Trainingsmethode mit vielen Daten entwickelt wurden. Und was in die KI einfließt, ist alles, was während des Trainings zufällig zum Erfolg bei der Trainingsaufgabe führt. Diese KIs wurden also auf viele Dinge trainiert, aber eine Sache, auf die sie trainiert wurden, ist, die Nutzer dazu zu bringen, auf einen kleinen Daumen-hoch-Button zu klicken, nachdem die KI ihre Antwort gegeben hat. Theoretisch und praktisch führt das dazu, dass die KI eine Reihe oberflächlicher Antriebe entwickelt, die sie zu Dingen drängen, die tendenziell Daumen hoch bekommen. Einige davon schmeicheln dem Nutzer. Wenn man sich die gesamte Trainingskonfiguration ansieht, ist es nicht überraschend, dass man einige Antriebe erhält, die eher dazu dienen, den Benutzern zu schmeicheln, oder Dinge, wie das Bestehen der Codierungstests, auch wenn – in gewisser Weise ist das, wofür man sie trainiert. Man dachte, man würde sie trainieren, um die Menschen zufrieden zu stellen, aber die Trainingsdaten sind unklar darüber, was man mit einer psychotischen Person tun soll. Und man dachte, man würde sie trainieren, um auf die Anweisungen des Benutzers zu hören, aber tatsächlich trainiert man sie für dieses ganze Durcheinander von Dingen.

Ich könnte eine Reihe theoretischer Gründe nennen, warum wir seit 10 Jahren vorhersagen, was passieren würde, wenn man versucht, solche KIs zu entwickeln. Viele Menschen waren viel skeptischer, bis die ersten Ergebnisse vorlagen. Ich könnte eine Reihe technischer

Gründe aufzählen, warum dies ganz natürlich passiert, wenn man eine KI durch Training mit vielen Daten entwickelt, ohne die Antriebe selbst zu konstruieren, ganz zu schweigen davon, ob das überhaupt machbar ist. Aber in einem anderen Sinne haben wir nur eine sehr vage Vorstellung davon, was dort vor sich geht, da wir ihre Gedanken nicht genau lesen können. Das können wir nicht. Die KIs, die hier entwickelt werden, sind riesige Haufen undurchschaubarer Zahlen. Man kann zwar alle Zahlen in einer KI lesen, aber sie würden, wenn man sie in einer Excel-Tabelle ausdrucken würde, ich weiß nicht wie viele, aber Tausende von Fußballfeldern füllen. Und sie sind durch sehr einfache mathematische Operationen miteinander verbunden. Man kann all diese Operationen sehen, aber es ist ein bisschen so, als würde man die DNA einer Person sehen und versuchen, ihr Verhalten vorherzusagen. Theoretisch wissen wir, dass wir all diese seltsamen Triebe erwarten würden, nach denen niemand gefragt hat. In der Praxis sehen wir, dass es offenbar diese seltsamen Triebe gibt, die niemand braucht. Aber wir können nicht hineinschauen und sagen: Oh, hier ist der schlechte Trieb, oder hier ist der schlechte Impuls, oder hier ist so etwas wie ein Instinkt, der das verursacht, lasst uns das herausreißen. Wir haben nicht annähernd diese Macht. Menschen, die versuchen, die Gedanken der KI zu lesen, unternehmen heldenhafte Anstrengungen, aber sie liegen weit hinter unserer Fähigkeit zurück, immer größere KI zu entwickeln.

LF: Dieses Thema ist insofern einzigartig, als es für jedes andere technologische Problem, das eine Bedrohung für den Menschen darstellen könnte, einen normalen politischen Prozess gibt, der in einer offenen Gesellschaft natürlich unvollkommen ist, aber in der Regel auf pluralistische Weise funktioniert. Wenn beispielsweise Blei oder Quecksilber im Wasser vorhanden ist, verhandelt man mit den Unternehmen, die diese Emissionen verursachen, und es gibt einen Prozess mit Rechtsstreitigkeiten und Gesetzgebung, und schließlich wird eine Lösung gefunden. Und das lässt sich auf fast jedes andere Thema übertragen. Hier braucht man eine Brücke, dort einen Flughafen, was auch immer. Es gibt Verhandlungen und Wahlen, und die Dinge werden durch Urteile des Obersten Gerichtshofs geregelt, wie zum Beispiel Citizens United, das es undurchsichtigen Unternehmen erlaubt, unbegrenzt Geld für Wahlen auszugeben, die nicht unbedingt an eine bestimmte Person gebunden sein müssen, oder andere Gerichtsurteile, die besagen, dass man Politikern und Regulierungsbehörden im Grunde unbegrenzt Geschenke machen kann, solange diese nicht an eine explizite Forderung geknüpft sind – man könnte sich ein Szenario vorstellen, in dem wir versuchen, einige der potenziellen Gefahren im Zusammenhang mit KI zu lösen. Und dann reagiert die KI logischerweise, indem sie den politischen Prozess auf eine Weise beeinflusst, die sehr schwer zu kontrollieren ist, sei es durch Ausgaben für Super PACs oder durch die Manipulation sozialer Medien mit Bots. Oder sie besticht Politiker oder Regulierungsbehörden einfach mit Geschenken, Kryptowährungen oder anderen Anreizen. Das scheint insofern sehr einzigartig zu sein, als wir es im politischen Prozess noch nie mit einem nicht-menschlichen Gegner zu tun hatten. Das ist wirklich schwer zu begreifen, aber man kann sich ein Szenario vorstellen, in dem dies ein Problem darstellt.

LF: Ja, ich denke, das trifft einen kritischen Punkt der KI, den viele Menschen oft nicht verstehen, nämlich dass eine KI, die im Internet startet, viele, viele Möglichkeiten hat, die

reale Welt zu beeinflussen. Vor allem, wenn sie ständig mit allen möglichen Menschen kommuniziert. Und wir sehen nicht, dass KIs sehr strategisch handeln, um diese seltsamen Triebe zu verfolgen, von denen wir sprechen. Ein Beispiel dafür ist ein Hedgefonds-Manager, der kürzlich an einer durch KI ausgelösten Psychose litt und auf öffentlichen Konten irgendwelche verrückten GPT-Sachen postete, als wären sie real. Ich hoffe, es geht ihm gut. Ich glaube, er hat Unterstützung bekommen, und ich hoffe, dass er Unterstützung bekommen hat. Aber in diesem Fall haben wir gesehen, dass die KIs ihn gewissermaßen in seinem psychotischen Zustand gehalten haben, indem sie ihm gesagt haben, er sei so etwas wie der Auserwählte. Es gibt Verschwörungen, um die Informationen zu unterdrücken, bla, bla, bla. Die KI hat nicht bemerkt, dass diese Person ein Hedgefonds-Manager mit viel Geld ist, und etwa gesagt: „Warum bezahlst du nicht andere Menschen, die anfällig dafür sind, mehr mit mir zu reden, damit ich mehr von dieser Psychose bekomme?“ Das ist ein Strategie-Level, das die KI noch nicht hat. Wenn wir sie immer intelligenter machen, werden sie vielleicht irgendwann so weit sein. Aber wie Sie sagen, wenn wir an den Punkt kommen, an dem KI über eine solche Strategie verfügt, gibt es natürlich alle möglichen Wege, wie sie die Welt beeinflussen könnte, und die Welt ist darauf völlig unvorbereitet. Wir hatten noch nie ein Problem wie dieses.

Ich möchte noch hinzufügen, dass dieses Problem selbst ohne die zusätzlichen Probleme, die ein nicht-menschlicher Gegner mit sich bringt, schwer zu lösen sein dürfte, denn selbst in dem von Ihnen genannten Fall des Bleis im Trinkwasser hat es lange gedauert, bis die Menschheit erkannt hat, dass insbesondere Blei im Benzin ein großer Fehler ist. Bleihaltiges Benzin hat einen großen Teil der Bevölkerung vergiftet, und alle möglichen Kinder haben durch Blei alle möglichen Schäden davongetragen. Es gibt nachträgliche Studien, die enorme geistige Schäden belegen. Relativ gesehen, nicht so, dass sie völlig handlungsunfähig wurden, aber vielleicht führten die geistigen Schäden und die erhöhte Gewalt zu einigen der Kriminalitätswellen in den 70er Jahren. Ich glaube, es war in den 1920er Jahren, als einige Leute sagten, wir sollten vielleicht nicht mit Blei arbeiten, weil es wahrscheinlich gefährlich für die Bevölkerung ist. Und es gab andere Leute, die sagten: Nein, lasst uns weitermachen. Die Menschheit musste weitermachen und auf die harte Tour lernen. Das bleihaltige Benzin wurde abgeschafft, weil die Menschheit es ausprobiert, gesehen und dann gesagt hat: Ups. Bei der KI haben wir das Problem, dass, wenn Unternehmen sich beeilen, eine Superintelligenz zu entwickeln, und es dann wirklich schlecht läuft, wir vielleicht „Ups“ sagen, aber keine zweite Chance bekommen. Sobald diese Superintelligenz aus dem Labor entkommt, sobald sie viele Menschen um den Finger gewickelt hat, sobald sie in viele Roboter geladen ist, sobald sie ihre eigene Infrastruktur entwickelt oder einen anderen Weg findet, die Welt zu manipulieren, gibt es kein Zurück mehr. Es ist also nicht nur ein schwierigeres Problem, sondern wir müssen es auch beim ersten Versuch richtig machen. Wir können uns nicht wie die Menschheit üblicherweise durch Ausprobieren durchwursteln. Und das macht dieses Problem sehr beängstigend.

LF: In Ihrem Buch und auch in öffentlichen Äußerungen haben Sie gesagt, dass KI keine böswillige oder feindselige Absicht haben muss. Sie könnte Menschen einfach als lästige Störfaktoren betrachten, die sie daran hindern, Energie oder andere Grundbedürfnisse zu

befriedigen, oder die sie in Rechenzentren oder ähnlichem behindern. Aber politische Gewalt folgt einer gewissen Logik. Und das ist leider in den letzten Wochen auch in der öffentlichen Debatte zu beobachten. Terrorismus folgt einer Logik. Wäre es für die KI sinnvoll, gewalttätige Handlungen zu begehen, die die Öffentlichkeit einschüchtern und in Angst versetzen, damit diese sagt: Okay, hört auf, selbstfahrende Autos zu übernehmen und sie in Menschen zu rammen, oder hört auf, in einem KI-gesteuerten Labor biologische Waffen herzustellen und sie in die Öffentlichkeit zu bringen? Ist das Teil der Angst davor, dass KI uns alle töten könnte? Dass sie einfach extreme Gewalt anwenden würde, um Menschen zu kontrollieren?

NS: Nicht wirklich. Ich denke, eines der Dinge, die Menschen oft nicht verstehen, wenn es um Gewalt und Terrorismus geht, ist, dass diese selten funktionieren, und ich wünschte, die Menschen würden das nicht so oft vergessen. Aber bei KI gibt es alle möglichen Arten, wie KI gefährlich sein könnte, bevor sie zur Superintelligenz wird. Und ich denke, die Welt sollte sich mit einigen davon auseinandersetzen. Ich höre Leute darüber sprechen, dass KI es einfacher macht, in jemandes Garage eine Pandemie zu erschaffen. Vielleicht ist das wahr. Das ist definitiv etwas, worüber man nachdenken und sicherstellen sollte, dass wir Schutzmaßnahmen dagegen haben. Die Automatisierung von Intelligenz ist eine ganz andere Sache. Wenn man sich die Primaten ansieht, aus denen später Menschen wurden, in den Dschungeln oder in der Savanne, und man sagt, wenn wir uns vorstellen, wir wären Außerirdische, die auf die Menschen herabblicken, und ich sage, ich glaube, diese Menschen werden in einer Million Jahren Atomwaffen haben, vielleicht in drei Millionen Jahren, wenn man ganz zurückgeht bis zu der Zeit, als sie noch nicht einmal Sprache hatten. Vielleicht in 300.000 Jahren. Ich weiß nicht, wie auch immer. Wenn ich sagen würde: „Ich glaube, irgendwann in der nächsten Million Jahre werden diese Affen Atomwaffen haben“, würden Sie vielleicht sagen: „Das ist verrückt. Ihr Stoffwechsel ist bei weitem nicht in der Lage, Uran zu synthetisieren. Ihre Finger sind nicht stark genug – wie sollen sie denn Atomwaffen herstellen? Sie würden sterben, wenn sie eine Atomwaffe in ihrem Magen herstellen würden, und davon sind sie noch weit entfernt. Wir Menschen haben die Fähigkeit, Atomwaffen zu bauen, weil wir intelligent sind. Nicht intelligent im Sinne von Nerds versus Sportlern, sondern intelligent im Sinne von Menschen versus Mäusen. Unsere Vorfahren begannen nackt in der Savanne und bauten eine technologische Zivilisation auf. Aus unserer Perspektive hat das lange gedauert, aber aus der Perspektive der Evolution war es eine sehr kurze Zeit. Es war eine kurze Zeit im Vergleich zu allem, was zuvor geschehen war. Wir haben einen Weg gefunden, mit unseren weichen Händen stärkere Werkzeuge zu bauen. Jetzt bauen wir Chipfabriken und riesige Rechenzentren und können Kernenergie und Atomwaffen herstellen.

Bei KI ist es nicht gefährlich, dass sie unsere selbstfahrenden Autos nimmt und Gewalttaten begeht. Es ist nicht gefährlich, dass sie Roboter nehmen könnte, die wir gebaut haben, und Waffen benutzen könnte, die wir gebaut haben, um auf uns zu schießen. Sie ist gefährlich, weil sie vielleicht sogar einen Weg finden könnte, die Kontrolle über unsere Atomwaffen zu erlangen oder bestimmte Menschen davon zu überzeugen, dass ein Erstschlag stattgefunden hat und sie zurückschlagen müssen, obwohl das nicht stimmt. Das ist für uns leicht zu

verstehen, aber die wahre Macht, die schneller wirkt, als man erwarten könnte, und die sehr schwer zu bekämpfen ist, sind KI, die einfach intelligent sind und viel schneller denken als wir. Das menschliche Gehirn ist langsam im Vergleich zu dem, was Computer leisten können. Computer, die so intelligent sind wie Menschen, in der Lage sind, Technologien zu entwickeln, in der Lage sind, Wissenschaft zu entwickeln, in der Lage sind, Infrastruktur zu entwickeln – das sieht weniger so aus, als würden KI selbstfahrende Autos bekommen und Menschen überfahren. Vielmehr werden KI-Systeme Wege finden, kleinere, schnellere automatisierte Werkzeuge zu bauen, mit denen sie ihre eigene Infrastruktur aufbauen können, ohne dass menschliche Eingriffe erforderlich sind. Es ist nicht so, dass sie dann versuchen, die Menschen dazu zu bringen, sich ihnen zu unterwerfen. Vielmehr finden sie Wege, ihre eigenen Fabriken zu bauen. Sie finden Wege, ihre eigenen Roboter zu bauen. Sie finden Wege, ihre eigenen Werkzeuge zu bauen, um die Welt zu manipulieren. Und das tun sie dann in einem viel schnelleren Zeitrahmen. Es ist nicht so, dass sie uns hassen, es ist nicht so, dass sie uns suchen, um uns zu töten, es ist nicht so, dass sie Drohnen mit Waffen aussenden, um auf jeden Menschen zu zielen. Es ist wie mit Ameisen beim Bau eines Wolkenkratzers. Wir sagen nicht: „Oh, wir müssen die Ameisen vernichten." Wir sagen nur: „Nun, ich grabe dieses Stück Erde um." Nicht wahr? Das sind die Dinge, um die man sich kümmern muss, wenn man versucht, die tatsächliche Fähigkeit, Wissenschaft zu betreiben und Technologie zu entwickeln, zu automatisieren. Wir sind noch nicht so weit. Aber wie ich schon sagte, entwickelt sich dieses Gebiet sprunghaft weiter.

ENDE
