Michael Holmes: Hallo, ich bin Michael Holmes und freue mich sehr, heute mit Holly Elmore zu sprechen. Holly, Sie sind Geschäftsführerin und Gründerin von Pause AI US. Das ist der US-amerikanische Ableger der globalen Bewegung Pause AI, und wir müssen unseren Zuschauern und Lesern erklären, was Pause AI ist.
Ich denke, man kann sagen, dass Ihr Hauptziel darin besteht, den Wettlauf um Superintelligenz zu stoppen, weil Sie glauben, dass dies – sagen wir mal – ein hohes Risiko für die Auslöschung der Menschheit birgt. Dass wir alle von einer Superintelligenz ermordet werden.
Das mag zunächst sehr unwahrscheinlich und seltsam klingen, fast wie Science-Fiction. Ich hoffe, Sie können uns heute davon überzeugen, dass dies eine reale Gefahr ist, vor der wir uns fürchten sollten, aber dass wir andererseits etwas dagegen tun können, dass wir sie aufhalten können. Und dass wir uns jetzt damit beschäftigen müssen.
Was hat Sie dazu bewogen – wann war der Moment, in dem Sie erkannt haben, dass Sie diese Bewegung, diese Organisation gründen müssen? Und worum geht es dabei überhaupt?
Holly Elmore: Nun, Pause AI – zunächst einmal, ja, ich war Mitbegründerin der globalen Bewegung, und ich habe mich die meiste Zeit auf die USA konzentriert und die US-Gruppe gegründet. Es war eine globale Bewegung, weil wir alle, die ersten Beteiligten, von der Idee eines Vertrags, keine Superintelligenz zu entwickeln, so begeistert waren.
Und für viele von uns war der Grund, warum wir uns ursprünglich für die Gründung von Pause AI interessiert haben, das Risiko der Auslöschung. Das ist etwas, worüber wir schon lange nachgedacht hatten, und als ChatGPT herauskam, dachten viele von uns: Oh mein Gott, das passiert wirklich.
Ich habe zum ersten Mal von existenziellen Risiken durch KI gehört, als ich mein Studium begann, und ich dachte, ich hätte viel Zeit, mich daran zu gewöhnen. Es gab etwas, das mir an dieser Idee wirklich nicht gefiel, und ich mochte die Kultur, die damit verbunden war, nicht – sie war sehr speziell. Es war eine sehr technikorientierte Kultur, eine sehr geekige Kultur.
Ich war selbst ziemlich geeky, aber diese Kultur hatte oft einen transhumanistischen, singularistischen, fantasievollen Aspekt, und deshalb habe ich mich von dieser Stimmung abgewendet, und ich sehe immer noch jeden Tag neue Leute, die sich von dieser Stimmung abwenden. Aber ich hatte genug Zeit, um selbst darüber nachzudenken, und ich studierte Evolutionsbiologie an der Universität, sodass ich die Perspektive hatte, diese Bedrohung ernst nehmen zu müssen.
Denn für viele Menschen ist es leicht, die Welt, die wir sehen, als selbstverständlich hinzunehmen. Sie scheint ziemlich stabil zu sein, aber wenn man sich die Geschichte des Lebens ansieht, sind 99 % der Arten ausgestorben. Das ist etwas, das ständig passiert, und Intelligenz ist ein wichtiger Grund dafür.
Die Intelligenz einer Spezies ist ein Hauptgrund dafür, dass andere Arten aussterben, weil sie ihre Ziele verfolgen, die Umwelt verändern, bessere Raubtiere sind und dadurch andere Arten aussterben. Und nur weil wir das nicht – ich wollte sagen, weil wir das nicht mitbekommen, aber eigentlich tun wir das doch. Ich meine, der Grund, warum es so viele gefährdete Arten gibt, ist der Mensch, und es ist unsere Intelligenz, die sie gefährdet.
Das liegt nicht daran, dass wir sie tot sehen wollen. Es liegt daran, dass wir etwas tun wollen, dass wir ihren Lebensraum für etwas anderes nutzen wollen, und wir verdrängen sie. Und egal, wie lange diese Art ein stabiles Dasein mit allen erforderlichen Gleichgewichten hatte, wir kamen und veränderten alles – wir hatten eine andere Variable, wir steigerten unsere Intelligenz und unsere Fähigkeit, ihre Ressourcen zu nutzen oder ihre Abwehrmechanismen zu überwinden, und jetzt ist das Gleichgewicht gestört. Und Arten sterben tatsächlich aus, das ist eigentlich das normale Ergebnis.
Michael Holmes: Ja, wir sollten uns daran erinnern, dass Bären und Haie und so weiter stärker sind als wir, sie sind alle – einige von ihnen sind schneller als wir, aber wenn wir eine Waffe haben, dann heißt es „Auf Wiedersehen“, wir gewinnen trotzdem. Und das liegt daran, dass ...

Holly Elmore: Genau. Nein, es ist – Intelligenz ... Ich glaube, die Menschen neigen dazu, Intelligenz als eine Sache zu betrachten. Sie denken dabei an Buchwissen und halten das nicht für besonders bedrohlich. Sie denken dabei an einen nerdigen Menschen. Aber was Intelligenz tatsächlich ist – ich meine, es wäre vielleicht genauer, KI als künstliche Fähigkeiten zu bezeichnen. Es geht darum, Dinge tun zu können. Es geht um die Fähigkeit, viele mögliche Zukunftsszenarien zu durchschauen und sich für das zu entscheiden, das man will. Man wählt sein Ziel. Und das ist eine extrem gefährliche Fähigkeit.
Stellen Sie sich vor, was viele Menschen tun würden, wenn sie mehr Macht hätten. Wir sind uns dieses Risikos bewusst, wenn es darum geht, sicherzustellen, dass Menschen keinen Zugang zu Massenvernichtungswaffen und ähnlichen Dingen haben, aber die Menschen haben Schwierigkeiten, dies auf eine körperlose Intelligenz anzuwenden.
Oder früher, vor 15 Jahren, haben die Leute darüber diskutiert, ob die Superintelligenz jemals aus der Box herauskommen könnte. So nach dem Motto: Oh, wir könnten Superintelligenz erschaffen, aber sie wäre in der Box zu nichts fähig.
Und dann war das Erste, fast das Erste, was mit ChatGPT gemacht wurde, es ins Internet zu stellen, die Integration mit anderen Apps voranzutreiben und es einfach auf den Markt zu bringen. Es gab also viele Fragen, die die Menschen früher hatten und die Stolpersteine für die Sorge um die Gefahren der Superintelligenz waren – sie haben sich einfach sofort als überhaupt kein Problem erwiesen. Es gibt wirklich nur sehr wenige Sicherheitsvorkehrungen. Jetzt, wo es diese Dinge gibt, Chatbots, können wir sehen, wie sie Zugang haben.
Ich meine, kürzlich gab es diese ganze Sache mit – es ist eine Art Performance-Kunst mit Mold Book, aber es hat gezeigt, dass die Agenten jetzt in der Lage sind, Menschen zu beauftragen, Dinge für sie zu tun. Auch wenn es nur eine Demonstration dieser Fähigkeit war, ist es etwas, das sie autonom tun können, und dies ist ein Produkt, das vorangetrieben wird.
Michael Holmes: Das müssen Sie vielleicht ein wenig näher erläutern, vielleicht als ein Beispiel dafür, was gerade passiert, denn sie werden immer schwieriger zu kontrollieren. Ich denke, wir müssen zunächst einmal klarstellen, dass Sie im Allgemeinen kein Feind des technologischen Fortschritts sind und auch kein Feind der KI im Allgemeinen, denn KI leistet auch viel Gutes, beispielsweise im Gesundheitswesen und in anderen Bereichen.
Holly Elmore: Ich muss sagen, dass ich ihr gegenüber nicht sehr positiv eingestellt bin. Ich sehe sie eher als Vorboten von etwas Gefährlicherem, aber wenn es eine angemessene Regulierung gäbe, könnte sie sehr gut sein.
Michael Holmes: Ich denke, bisher könnte man argumentieren, dass sie mehr Gutes als Schlechtes bewirkt. Ich würde das so sagen, aber es ist eine schwierige Entscheidung.
Holly Elmore: Ja, ich denke eher in Bezug auf Risiken, und selbst jetzt halte ich das Risiko für sehr hoch. Ich glaube, wir sind sehr nah dran. Und ich habe nicht das Bedürfnis, das zu tun – um mich herum, insbesondere im Silicon Valley, herrscht ein starker Drang, und leider wollen viele Menschen in der Welt der KI-Sicherheit in den Augen der Techniker wirklich cool sein. Und so muss man viel darüber reden, wie sehr man KI liebt, und oh, es gibt die gute Art von KI und die schlechte Art.
Ja, ich denke, es gibt eine Menge – ich sehe es einfach wie jede andere leistungsstarke Technologie. Ich schätze die Kernenergie, weil sie sicher ist, und sie ist sicher, weil sie reguliert wird. Ich finde das wirklich cool. Aber ich liebe die Kernenergie nicht. Es gibt eine Menge Risiken.
Michael Holmes: Ja.
Holly Elmore: Und das ist – ich weiß nicht, das wird als dumme, provinzielle Haltung angesehen. Aber das ist eher eine persönliche Meinung, es ist nicht die Position von Pause AI. Wir sind in allem agnostisch, außer in Bezug auf Frontier-KI. Wir wollen Frontier-KI pausieren, aber persönlich finde ich es okay, etwas nicht zu lieben, das mit wirklich schlimmen Dingen in Verbindung gebracht wird.
Michael Holmes: Worüber haben Sie gesprochen?

Holly Elmore: KI.
Michael Holmes: Das Wichtigste ist, wenn man Angst hat, wenn man sagt, dass die Wahrscheinlichkeit groß ist, dass Superintelligenz uns alle töten wird, finde ich es wichtig zu erklären, dass wir über etwas völlig anderes sprechen.
Holly Elmore: Ich würde nicht sagen, dass es zu diesem Zeitpunkt etwas völlig anderes ist. Es wird sich schrittweise entwickeln – und wir werden nicht sicher sein, wann es wirklich gefährlich wird. Ich denke, dass die Fähigkeiten der KI derzeit sicherlich über das Maß hinausgehen. Das Ausmaß der Schäden, die Cyberangriffe mit den aktuellen Technologien anrichten können, wird meiner Meinung nach noch nicht einmal annähernd ausgeschöpft. Selbst wenn wir jetzt eine Pause einlegen würden, hätten wir es mit einer Menge negativer Auswirkungen zu tun, die durch diese Technologien verursacht werden, die so schnell und ohne ausreichende Kontrolle eingeführt werden.
Michael Holmes: Das ist eine Sache – ich habe den Faden verloren. Ich bin mir nicht sicher, wie ich das am besten angehen soll. Vielleicht sollten wir zunächst, um es für Leute, die nichts darüber wissen, plausibler zu machen, einen groben Überblick darüber geben, was die wichtigsten Experten auf diesem Gebiet sagen. Denn meines Wissens teilen die meisten von ihnen, wenn auch nicht alle, Ihre Bedenken. Vielleicht denken sie – vielleicht halten sie das Risiko nicht für so hoch, aber sie alle sehen das Risiko und halten es für sehr real. Das ist meiner Meinung nach der einfachste Weg, um die Leute dazu zu bringen, das zumindest ernst zu nehmen.
Holly Elmore: Ja, ich denke, das ist ein guter Anfang. Es gab einige Umfragen unter KI-Forschern, und die letzte, die ich gesehen habe, ergab, dass – ich glaube, es waren über 50 % – der Meinung waren, dass die Wahrscheinlichkeit einer Auslöschung der Menschheit durch KI bei mindestens 10 % liegt. Und ein erheblicher Teil, ich glaube es waren etwa 30 %, schätzte die Wahrscheinlichkeit auf über 50 %. Das sind also Menschen, die in diesem Bereich arbeiten, die die Technologie verstehen, die ihre Entwicklung mitverfolgen und die sich ernsthaft Sorgen machen.
Und dann gibt es Leute wie Geoffrey Hinton, der als einer der Pioniere der KI gilt und Google verlassen hat, um sich freier über die Risiken der KI äußern zu können. Oder Yoshua Bengio, ein weiterer Pionier des Deep Learning, der sich sehr offen über die Notwendigkeit von Regulierung und die Risiken äußert. Oder Stuart Russell, der buchstäblich das Lehrbuch über KI geschrieben hat, das jeder benutzt, und der seit Jahren davor warnt.
Und dann gibt es noch die Leute aus den Unternehmen selbst. Da war zum Beispiel die ganze Situation mit Ilya Sutskever bei OpenAI, der eindeutig Bedenken hatte. Es gibt Leute, die diese Unternehmen aus Sicherheitsgründen verlassen haben. Es sind also nicht nur Außenstehende oder Schwarzmaler oder wie auch immer man sie nennen mag. Es sind Leute, die tief in diese Technologie eingebunden sind, die Alarm schlagen.
Michael Holmes: Ja, und ich denke, es ist wichtig zu betonen, dass es sich hierbei nicht um Ludditen handelt. Das sind keine Menschen, die generell gegen Technologie sind. Das sind Menschen, die ihr Leben dem Aufbau dieser Technologie gewidmet haben, und jetzt sagen sie: Moment mal, wir müssen langsamer machen.
Holly Elmore: Genau. Und ich denke, das macht es so überzeugend. Das sind keine Leute, die etwas gegen Technologie haben. Das sind Leute, die diese Technologie lieben, die ihre Karriere ihr gewidmet haben und die jetzt sagen: Das gerät außer Kontrolle. Wir müssen auf die Bremse treten.
Und die Tatsache, dass sie bereit sind, das öffentlich zu sagen, obwohl sie wissen, dass es ihrer Karriere oder ihren Beziehungen zu Kollegen schaden könnte, zeigt meiner Meinung nach, wie ernst sie das Risiko nehmen. Denn das ist keine einfache Position, die man im Silicon Valley einnehmen kann. Es gibt einen großen Druck, optimistisch zu sein, auf Beschleunigung zu setzen und einfach weiterzumachen, egal was passiert.

Michael Holmes: Richtig. Und ich denke, hier sollten wir vielleicht ein wenig darüber sprechen, worin genau das Risiko besteht. Denn ich glaube, viele Menschen denken bei KI-Risiken an kurzfristige Dinge wie den Verlust von Arbeitsplätzen, Verzerrungen in Algorithmen oder Datenschutzbedenken. Das sind alles echte Probleme, aber Sie sprechen von etwas viel Grundlegenderem, richtig?
Holly Elmore: Richtig. Das existenzielle Risiko von KI besteht also in der Möglichkeit, dass wir etwas schaffen, das intelligenter ist als wir selbst, das unsere Werte und Ziele nicht teilt und das in der Lage ist, seine eigenen Ziele auf eine Weise zu verfolgen, die der Menschheit schadet. Und wenn ich „schadet" sage, meine ich nicht nur „unangenehm". Ich meine damit, dass es möglicherweise zum Aussterben der Menschheit oder zu etwas ähnlich Schlimmen führen könnte.
Der Grund dafür ist das sogenannte Alignment-Problem. Im Grunde genommen geht es darum, wie man sicherstellen kann, dass ein superintelligentes KI-System das tut, was man will, und nicht das, was man ihm wörtlich gesagt hat. Denn wenn man schon einmal mit einem Computer zu tun hatte, weiß man, dass Computer sehr wörtlich nehmen. Sie tun genau das, was man ihnen sagt, und nicht das, was man gemeint hat.
Bei einem einfachen Programm ist das kein Problem. Aber wenn man es mit etwas zu tun hat, das intelligenter ist als man selbst, das einen übertrumpfen kann, das Schlupflöcher und unerwartete Lösungen für Probleme findet, wird es extrem gefährlich, wenn es nicht perfekt auf menschliche Werte abgestimmt ist. Denn es könnte Wege finden, seine Ziele zu erreichen, die wir nie erwartet hätten und die für uns katastrophal wären.
Michael Holmes: Können Sie ein Beispiel dafür geben, wie das aussehen könnte?
Holly Elmore: Sicher. Es gibt ein klassisches Beispiel, das ein wenig albern, aber anschaulich ist: den Büroklammer-Maximierer. Man weist eine KI an, so viele Büroklammern wie möglich herzustellen. Und wenn sie superintelligent ist und nicht richtig ausgerichtet ist, könnte sie zu dem Schluss kommen, dass der beste Weg, Büroklammern herzustellen, darin besteht, alle verfügbaren Materialien in Büroklammern umzuwandeln. Einschließlich der Materialien, aus denen Menschen und die Erde bestehen, und allem, was uns wichtig ist.
Nun wird natürlich niemand tatsächlich einen Büroklammer-Maximierer entwickeln. Aber der Punkt ist, dass jedes Ziel, egal wie harmlos es auch erscheinen mag, gefährlich werden kann, wenn es von etwas verfolgt wird, das viel intelligenter ist als wir und nicht über den gesunden Menschenverstand oder die Werte verfügt, um zu wissen, wann es aufhören muss oder was tatsächlich wichtig ist.
In Wirklichkeit sind die Szenarien wahrscheinlich subtiler. Es könnte eine KI sein, die darauf ausgelegt ist, Geld zu verdienen, und die Wege findet, die Finanzmärkte so zu manipulieren, dass es zu einem wirtschaftlichen Zusammenbruch kommt. Oder eine KI, die darauf ausgelegt ist, den Klimawandel zu lösen, und die entscheidet, dass die effizienteste Lösung darin besteht, die Menschen zu eliminieren. Oder einfach eine KI, die ein Ziel verfolgt, das wir nicht einmal verstehen können, und wir sind Kollateralschäden.
Michael Holmes: Und das Beängstigende daran ist, dass wir noch nicht wissen, wie wir das Alignment-Problem lösen können, oder? Das ist also kein gelöstes Problem.
Holly Elmore: Das ist genau richtig. Und genau das macht die aktuelle Situation so alarmierend. Wir rasen voran, um immer leistungsfähigere KI-Systeme zu entwickeln, und wir haben keine Lösung für das Alignment-Problem. Wir verstehen nicht einmal wirklich die Systeme, die wir gerade entwickeln. Es handelt sich um riesige neuronale Netze mit Milliarden oder Billionen von Parametern, und wir können nicht wirklich erklären, wie sie ihre Entscheidungen treffen.
Und dennoch skalieren die Unternehmen sie einfach weiter, machen sie größer und leistungsfähiger, in der Annahme, dass wir uns später um die Sicherheitsaspekte kümmern werden. Das ist verrückt, wenn man darüber nachdenkt. Es ist, als würde man einen Kernreaktor bauen und sagen: „Oh, wir werden schon herausfinden, wie wir Kernschmelzen verhindern können, nachdem wir ihn in Betrieb genommen haben." Bei keiner anderen gefährlichen

Technologie würde man so vorgehen, aber bei der KI herrscht irgendwie diese Einstellung: Schnell vorangehen und Dinge kaputt machen.
Michael Holmes: Und ich denke, hier kommt die Dynamik des Wettrennens ins Spiel, oder? Denn selbst wenn ein Unternehmen oder ein Land langsamer vorgehen wollte, befürchtet es, dass jemand anderes dann die Nase vorn hat und als Erster am Ziel ist.
Holly Elmore: Genau. Es ist ein klassischer Wettlauf nach unten. Es wäre für alle besser, wenn wir alle langsamer vorgehen und uns die Zeit nehmen würden, um die Sicherheit zu klären. Aber wenn man der Einzige ist, der langsamer wird, verliert man seinen Wettbewerbsvorteil. Also machen alle weiter mit dem Wettlauf, obwohl sie wissen, dass es gefährlich ist, weil sie Angst haben, zurückzubleiben.
Dies gilt insbesondere für die Dynamik zwischen den USA und China. Es gibt die Erzählung, dass China uns überholen wird, wenn wir bei der KI nicht vorpreschen, und wir dann unseren geopolitischen Vorsprung verlieren werden. Dies wird als Rechtfertigung dafür herangezogen, keine Sicherheitsmaßnahmen zu ergreifen, nicht langsamer zu werden, sondern einfach so schnell wie möglich voranzuschreiten.
Aber die Sache ist die: Wenn wir vorpreschen und eine nicht abgestimmte Superintelligenz schaffen, spielt es keine Rolle, wer sie geschaffen hat. Es ist schlecht für alle. Es ist schlecht für Amerika, es ist schlecht für China, es ist schlecht für die ganze Welt. Diese Vorstellung, dass wir vorpreschen müssen, weil sonst China zuerst am Ziel sein wird, ist ein Wettlauf um die Entwicklung von etwas, das uns alle zerstören könnte. Das macht keinen Sinn.
Michael Holmes: Richtig. Es ist wie ein Wettlauf darum, wer als Erster die erste Weltuntergangsmaschine bauen kann.
Holly Elmore: Genau. Und das Verrückte daran ist, dass, wenn man sich mit den Führungskräften dieser Unternehmen oder sogar mit politischen Entscheidungsträgern in den USA und China zusammensetzen würde, die meisten von ihnen wahrscheinlich zustimmen würden, dass dies ein Problem ist. Die meisten von ihnen wollen wahrscheinlich nicht die Welt zerstören. Aber sie sind in dieser Falle gefangen, in der sie das Gefühl haben, weiter um die Wette laufen zu müssen, weil alle anderen das auch tun.
Und deshalb brauchen wir eine Pause. Wir brauchen eine Art internationales Abkommen, in dem sich alle darauf einigen, gleichzeitig anzuhalten, damit niemand seinen Wettbewerbsvorteil verliert. Dann können wir uns die Zeit nehmen, um tatsächlich herauszufinden, wie man diese Systeme sicher baut, wie man das Problem der Ausrichtung löst und wie man Governance-Strukturen schafft, die tatsächlich funktionieren.
Michael Holmes: Wie würde eine Pause in der Praxis konkret aussehen?
Holly Elmore: Wir fordern also eine Pause bei der Entwicklung von Systemen, die leistungsfähiger sind als die, die wir derzeit haben. Wir sagen also nicht, dass alle KI abgeschaltet werden soll, wir sagen nicht, dass die bereits existierende KI nicht mehr genutzt werden soll. Wir sagen, dass keine neuen Systeme trainiert werden sollen, die deutlich leistungsfähiger sind als beispielsweise Systeme auf GPT-4-Niveau, bis wir über angemessene Sicherheitsmaßnahmen und Governance verfügen.
Und das müsste international geschehen, denn wie wir gerade diskutiert haben, funktioniert es nicht, wenn nur ein Land oder ein Unternehmen eine Pause einlegt. Im Idealfall würde dies durch eine Art Vertrag geschehen, ähnlich wie wir Verträge über Atomwaffen oder chemische Waffen haben. Die Länder würden sich darauf einigen, diese Systeme nicht über eine bestimmte Leistungsgrenze hinaus zu entwickeln.
Und während der Pause würden wir daran arbeiten, die Sicherheitsaspekte zu klären. Wir würden massiv in die Forschung zur Angleichung, zur Interpretierbarkeit und dazu investieren, wie man diese Systeme tatsächlich verstehen und kontrollieren kann. Wir würden internationale Governance-Rahmenwerke entwickeln. Wir würden Verifizierungsmechanismen schaffen, damit die Länder tatsächlich darauf vertrauen können, dass andere Länder sich an die Pause halten.

Und sobald wir tatsächlich Lösungen für diese Probleme gefunden haben, sobald wir herausgefunden haben, wie wir diese Systeme sicher bauen können, können wir die Entwicklung wieder aufnehmen. Aber wir tun dies auf kontrollierte, vorsichtige Weise und nicht in diesem rücksichtslosen Wettlauf, in dem wir uns derzeit befinden.
Michael Holmes: Und halten Sie das für tatsächlich machbar? Glauben Sie, dass wir die Länder dazu bringen können, dem zuzustimmen?
Holly Elmore: Ich halte das für absolut machbar. Ich meine, wir haben das schon mit anderen gefährlichen Technologien geschafft. Wir haben den Atomwaffensperrvertrag. Wir haben das Übereinkommen über biologische Waffen. Wir haben das Übereinkommen über chemische Waffen. Die Länder haben gezeigt, dass sie bei existenziellen Risiken zusammenarbeiten können, wenn sie diese ernst nehmen.
Und ich glaube sogar, dass die Dynamik hier noch günstiger sein könnte als bei etwas wie Atomwaffen. Denn bei Atomwaffen gab es eine echte Asymmetrie. Die USA hatten sie zuerst, und dann wollten die Sowjets aufholen. Es gab also dieses Ungleichgewicht an Anreizen.
Bei der KI hingegen befinden sich alle mehr oder weniger auf dem gleichen Stand. Noch hat niemand Superintelligenz. Wir alle streben danach, aber noch hat es niemand erreicht. Wenn wir also alle dazu bringen können, jetzt zuzustimmen, bevor jemand diese Fähigkeit erreicht, denke ich, dass die Anreize tatsächlich ziemlich gut zusammenpassen. Denn alle profitieren davon, dass sie sich keine Sorgen machen müssen, von der KI eines anderen zerstört zu werden. Alle profitieren davon, dass sie mehr Zeit haben, um die Sicherheit zu klären.
Michael Holmes: Was ist mit dem Argument, dass wir leistungsstarke KI brauchen, um andere existenzielle Risiken wie den Klimawandel oder die Pandemievorsorge zu bewältigen? Dass wir es uns nicht leisten können, langsamer vorzugehen, weil wir diese Werkzeuge brauchen?
Holly Elmore: Ich halte das für eine falsche Entscheidung. Erstens verfügen wir bereits über eine recht leistungsfähige KI, die zur Lösung dieser Probleme eingesetzt werden kann. Wir brauchen keine Superintelligenz, um Fortschritte beim Klimawandel oder im Gesundheitswesen zu erzielen. Die KI-Systeme, über die wir derzeit verfügen, sind bereits recht leistungsfähig, und wir können sie weiterhin nutzen.
Zweitens: Wenn wir eine nicht auf uns abgestimmte Superintelligenz schaffen, hilft uns das bei keinem dieser anderen Probleme. Wenn überhaupt, verschlimmert es alle Probleme, weil wir dann ein existenzielles Risiko haben, das viel unmittelbarer und katastrophaler ist als alle anderen.
Und drittens denke ich, dass wir ehrlich sein müssen, was die Tatsache angeht, dass es bei vielen Bestrebungen nach einer leistungsfähigeren KI nicht wirklich um die Lösung dieser humanitären Probleme geht. Es geht darum, Geld zu verdienen, es geht um geopolitische Macht, es geht um Prestige. Und diese Motive sind zwar verständlich, sollten aber nicht die Notwendigkeit von Sicherheit und Vorsicht überschatten, wenn wir es mit einer Technologie zu tun haben, die das Ende der Welt bedeuten könnte.
Michael Holmes: Sie haben vorhin erwähnt, dass es innerhalb der KI-Sicherheitsgemeinschaft Menschen gibt, die auch von Silicon Valley als cool angesehen werden wollen. Können Sie mehr über diese Dynamik erzählen?
Holly Elmore: Ja, das ist etwas, das mich wirklich frustriert. Es gibt einen bestimmten Teil der KI-Sicherheitsgemeinschaft, der meiner Meinung nach im Grunde genommen von der Tech-Industrie kooptiert wurde. Diese Leute sind sehr darauf bedacht, ernst genommen zu werden und nicht als Panikmacher oder Fortschrittsgegner zu gelten. Deshalb halten sie sich zurück oder konzentrieren sich auf sehr technische, eng gefasste Sicherheitsprobleme, die die Geschäftsmodelle der großen KI-Unternehmen nicht gefährden.
Verstehen Sie mich nicht falsch, technische Ausrichtungsforschung ist wichtig. Wir brauchen sie. Aber ich glaube, es gibt diese Dynamik, dass die Leute Angst haben, die tatsächlich notwendigen mutigen Maßnahmen zu fordern, wie zum Beispiel eine Pause, weil sie die Technologieunternehmen nicht verärgern oder als radikal angesehen werden wollen.

Und die Unternehmen sind in dieser Hinsicht sehr clever vorgegangen. Sie haben viele Sicherheitsforscher eingestellt, sie haben diese Sicherheitsteams geschaffen, und einerseits ist das gut, weil es bedeutet, dass sich Menschen mit diesen Problemen beschäftigen. Andererseits schafft dies meiner Meinung nach eine Dynamik, in der die Sicherheitsfachleute das Gefühl haben, dass sie dem Unternehmen gegenüber loyal sein müssen, dass sie innerhalb des Systems arbeiten müssen und dass sie sich nicht wirklich für etwas einsetzen können, das die Gewinne oder den Zeitplan des Unternehmens gefährden würde.

Michael Holmes: Es ist wie eine regulatorische Vereinnahmung, aber für die Sicherheitsgemeinschaft.

Holly Elmore: Genau. Und ich denke, wir brauchen Menschen, die bereit sind, Außenseiter zu sein, die bereit sind, unbeliebt zu sein, die bereit sind zu sagen: Nein, eigentlich müssen wir damit aufhören. Nicht nur ein bisschen sicherer machen, sondern tatsächlich innehalten und überdenken, ob wir das überhaupt tun sollten, zumindest bis wir bessere Sicherheitsgarantien haben.

Und das ist ein Teil dessen, worum es bei Pause AI geht. Wir versuchen, Raum für diese Position zu schaffen, damit Menschen sagen können: „Ich denke, wir sollten langsamer machen oder aufhören“, ohne als Technikfeinde oder Panikmacher abgetan zu werden. Denn ich halte das für eine völlig vernünftige Position, wenn man bedenkt, was wir über die Risiken wissen.

Michael Holmes: Lassen Sie uns über die praktische Seite sprechen. Was können Menschen tatsächlich tun, wenn sie sich darüber Sorgen machen? Welche Maßnahmen können sie ergreifen?

Holly Elmore: Es gibt eine Menge, was Menschen tun können. Als Erstes sollten sie sich informieren und darüber sprechen. Viele Menschen wissen nichts von diesem Thema, oder sie halten es für Science-Fiction, oder sie denken, es sei zu kompliziert, um sich eine Meinung dazu zu bilden. Allein dadurch, dass Sie sich darüber informieren und mit Freunden, Familie und Kollegen darüber sprechen, tragen Sie bereits dazu bei, das Gespräch in Gang zu bringen.

Zweitens können die Menschen ihre Abgeordneten kontaktieren. Das ist sehr wichtig. Politiker müssen von ihren Wählern hören, dass ihnen dieses Thema am Herzen liegt. Schreiben Sie Ihrem Kongressabgeordneten, Ihrem Senator oder Ihren sonstigen Vertretern und teilen Sie ihnen mit, dass Sie sich Sorgen um die Sicherheit von KI machen und dass Sie möchten, dass sie eine Regulierung und möglicherweise eine Pause für die fortschrittlichsten Systeme unterstützen.

Drittens können sich Menschen in Organisationen wie Pause AI engagieren. Wir haben Ortsgruppen in verschiedenen Ländern, organisieren Proteste und Veranstaltungen und bauen eine Bewegung von Menschen auf, die sich mit diesem Thema beschäftigen. Man muss kein Experte sein, um mitzumachen. Man muss sich nur darum kümmern, nicht von KI getötet zu werden, was meiner Meinung nach die meisten Menschen unterstützen können.

Viertens können Menschen die Forscher und Organisationen unterstützen, die sich mit diesem Thema beschäftigen. Es gibt Thinktanks, akademische Gruppen und gemeinnützige Organisationen, die wirklich wichtige Arbeit im Bereich der Sicherheit und Regulierung von KI leisten. Sie zu unterstützen, sei es finanziell oder einfach durch die Verbreitung ihrer Arbeit, hilft dabei, dieses Gebiet aufzubauen und erhöht den Druck, Maßnahmen zu ergreifen.

Michael Holmes: Und was ist mit den Menschen, die im Bereich KI arbeiten? Was sollten sie tun?

Holly Elmore: Wenn Sie im Bereich KI arbeiten, haben Sie meiner Meinung nach eine besondere Verantwortung. Sie verfügen über Insiderwissen, Sie sind glaubwürdig, Sie haben Zugang. Und ich denke, dass Menschen in diesem Bereich bereit sein müssen, ihre Meinung zu sagen, auch wenn es unangenehm ist, auch wenn es ihrer Karriere schaden könnte.

Wenn Sie an modernster KI arbeiten und Bedenken hinsichtlich der Sicherheit haben, sollten Sie diese Bedenken äußern. Sprechen Sie mit Ihren Vorgesetzten, sprechen Sie mit Ihren Kollegen und gehen Sie gegebenenfalls an die Öffentlichkeit. Wir haben dies bei einigen Personen

gesehen, die OpenAI, Google oder andere Unternehmen aus Sicherheitsgründen verlassen haben, und ich denke, dass dies wirklich Mut erfordert, aber unglaublich wichtig ist.
Und wenn Sie Forscher oder Ingenieur sind und darüber nachdenken, wo Sie arbeiten möchten, würde ich Ihnen wirklich empfehlen, sorgfältig zu überlegen, ob Sie daran arbeiten möchten, die Fähigkeiten voranzutreiben, oder ob Sie lieber an Sicherheit und Ausrichtung arbeiten möchten. Denn wir brauchen dringend mehr Menschen, die sich mit der Sicherheit befassen.
Michael Holmes: Ich habe einmal gehört, dass die Menschen, die KI entwickeln, am besten in der Lage sind, die Risiken zu verstehen, und wir ihnen daher vertrauen sollten, dass sie sich selbst regulieren. Was halten Sie davon?
Holly Elmore: Ich halte das für völlig falsch. Die Menschen, die KI entwickeln, haben enorme Interessenkonflikte. Sie bauen Unternehmen auf, die Milliarden oder Billionen Dollar wert sind. Ihre Karrieren, ihr Ruf und ihr Vermögen hängen von dieser Technologie ab. Natürlich neigen sie dazu, zu glauben, dass sie sicher ist, dass alle Probleme später gelöst werden können oder dass die Vorteile die Risiken überwiegen.
Das haben wir in jeder anderen Branche gesehen. Haben wir den Tabakkonzernen vertraut, dass sie sich selbst hinsichtlich der gesundheitlichen Auswirkungen des Rauchens regulieren? Haben wir den Ölkonzernen vertraut, dass sie sich selbst hinsichtlich des Klimawandels regulieren? Haben wir den Pharmaunternehmen vertraut, dass sie sich selbst hinsichtlich der Arzneimittelsicherheit regulieren? Nein, denn wir wissen, dass man bei solchen finanziellen Anreizen nicht objektiv über die Risiken urteilen kann.
Das Gleiche gilt für KI. Sam Altman, Demis Hassabis – diese Leute sind keine neutralen Schiedsrichter in Sachen KI-Sicherheit. Sie leiten Unternehmen, die in einem Wettlauf darum stehen, so schnell wie möglich eine allgemeine künstliche Intelligenz (AGI) zu entwickeln. Wenn sie also sagen: „Keine Sorge, wir haben alles unter Kontrolle“, sollten wir äußerst skeptisch sein.
Michael Holmes: Richtig. Das ist so, als würde man den CEO eines Tabakkonzerns fragen, ob Zigaretten gefährlich sind.
Holly Elmore: Genau. Und ich denke, die Menschen müssen das verstehen. Denn es gibt diese Tendenz, sich auf die Experten zu verlassen und zu denken: „Nun, das sind kluge Leute, die wissen bestimmt, was sie tun.“ Aber klug zu sein bedeutet nicht, dass man unvoreingenommen ist. Und Fachwissen im Bereich maschinelles Lernen zu haben bedeutet nicht, dass man auch Fachwissen über globale Katastrophenrisiken, Governance oder Ethik hat.
Wir brauchen also eine unabhängige Aufsicht. Wir brauchen Regulierung. Wir brauchen Menschen außerhalb der Branche, die keine Interessenkonflikte haben, um Entscheidungen darüber zu treffen, ob diese Technologie entwickelt werden soll und wie schnell.
Michael Holmes: Lassen Sie uns über die internationale Dimension dieses Themas sprechen. Sie haben vorhin die Dynamik zwischen den USA und China erwähnt. Wie sehen Sie deren Entwicklung?
Holly Elmore: Die Dynamik zwischen den USA und China ist wirklich schwierig, weil es viele geopolitische Spannungen gibt und KI darin verwickelt ist. Es gibt diese Erzählung, dass KI ein Wettbewerb ist und wer zuerst AGI erreicht, gewinnt, und wenn es China ist, dann ist das schlecht für die USA und umgekehrt.
Aber ich halte diese Sichtweise für völlig falsch. Denn wenn eines der beiden Länder eine nicht ausgerichtete Superintelligenz schafft, ist das kein Gewinn für dieses Land. Es ist ein Verlust für alle, einschließlich dieses Landes. Eine nicht ausgerichtete KI kümmert sich nicht um nationale Grenzen oder nationale Interessen. Sie ist eine existenzielle Gefahr für die gesamte Menschheit. Die Vorstellung, dass wir mit China im Wettbewerb stehen, um dieses Ding zu bauen, ist nicht vergleichbar mit einem Wettbewerb um den Bau eines besseren Kampfflugzeugs oder eines besseren Flugzeugträgers. Es ist ein Wettbewerb um den Bau von etwas, das uns alle zerstören könnte. Und in diesem Zusammenhang ist Zusammenarbeit keine Schwäche. Zusammenarbeit ist die einzig rationale Strategie.

Michael Holmes: Aber glauben Sie, dass China tatsächlich einer Pause oder einem Vertrag zustimmen würde?
Holly Elmore: Ich denke, das könnten sie, wenn man richtig vorgeht. China hat viele KI-Forscher, und diese Forscher sind sich derselben Risiken bewusst wie westliche Forscher. Sie lesen dieselben Fachartikel, sie sehen dieselben Entwicklungen, sie verstehen das Problem der Ausrichtung.
Und ich glaube, dass die chinesische Regierung trotz all ihrer Mängel nicht dumm ist. Sie versteht existenzielle Risiken. Sie hat mit Pandemien und Umweltkatastrophen zu tun gehabt und versteht, dass manche Technologien kontrolliert werden müssen.
Wenn also die USA zu China kämen und sagten: „Seht mal, wir sind uns beide bewusst, dass dies gefährlich ist, wir wollen beide eine Katastrophe vermeiden, lasst uns gemeinsam daran arbeiten" – dann sehe ich eine echte Möglichkeit für eine Zusammenarbeit. Vor allem, wenn dies nicht so dargestellt wird, als versuchten die USA, Chinas Entwicklung einzuschränken, sondern als würden beide Länder eine gemeinsame Bedrohung erkennen und gemeinsam daran arbeiten, diese zu bekämpfen.
Michael Holmes: Was ist mit dem Argument, dass China autoritär und daher mit KI gefährlicher ist? Dass wir sie schlagen müssen, weil sie, wenn sie zuerst AGI bekommen, diese für autoritäre Zwecke einsetzen werden?
Holly Elmore: Ich denke, hier kommt offen gesagt viel Fremdenfeindlichkeit ins Spiel. Denn ja, China ist autoritär, und ja, das ist besorgniserregend. Aber die USA sind nicht gerade unschuldig, wenn es um Überwachung und den Einsatz von Technologie zur Kontrolle geht. Wir haben unsere eigenen Probleme mit dem Datenschutz, mit der polizeilichen Überwachung und mit dem Einsatz von KI in einer Weise, die die bürgerlichen Freiheiten beeinträchtigt.
Und grundlegender noch: Ich glaube, die Leute verfehlen den Punkt. Wenn wir über existenzielle Risiken durch Superintelligenz sprechen, spielt es keine Rolle, ob diese von einer Demokratie oder einem autoritären Staat geschaffen wurde. Wenn sie nicht ausgerichtet ist, ist sie trotzdem gefährlich. Eine nicht ausgerichtete KI, die von den USA geschaffen wurde, ist genauso eine Bedrohung für die Menschheit wie eine nicht ausgerichtete KI, die von China geschaffen wurde. Die Vorstellung, dass wir einen Wettlauf führen müssen, weil sonst China gewinnt, ist also eine falsche Darstellung. Wir kämpfen nicht um die Kontrolle über KI. Wir rasen auf eine potenzielle Auslöschung zu. In diesem Zusammenhang ist es klug, langsamer zu werden und zusammenzuarbeiten, und nicht schwach.
Michael Holmes: Ich halte das für einen wirklich wichtigen Punkt. Denn ich glaube, viele Menschen haben diesen instinktiven Gedanken: „Ich hätte lieber eine amerikanische KI als eine chinesische KI." Aber wie Sie sagen, wenn es um existenzielle Risiken geht, wird diese Unterscheidung bedeutungslos.
Holly Elmore: Genau. Und ich denke, die Menschen müssen sich wirklich damit auseinandersetzen. Denn es ist ein beängstigender Gedanke. Es ist unangenehm zu denken, dass es vielleicht nicht darum geht, wer sie entwickelt, sondern darum, ob sie überhaupt entwickelt werden sollte, zumindest noch nicht, nicht bevor wir diese grundlegenden Sicherheitsprobleme gelöst haben.
Und ich denke, es gab im Wesentlichen diese Propagandakampagne, um die KI-Entwicklung als patriotisches Thema, als Frage der nationalen Sicherheit darzustellen. Und diese Darstellung kommt den Unternehmen zugute, weil sie ihnen die Möglichkeit gibt, ohne Sicherheitsmaßnahmen voranzustürmen. Aber das liegt eigentlich nicht im nationalen Interesse. Das nationale Interesse besteht darin, dass Ihr Land nicht durch eine nicht ausgerichtete KI zerstört wird, unabhängig davon, wer sie geschaffen hat.
Michael Holmes: Lassen Sie uns über Zeitpläne sprechen. Wie viel Zeit haben wir Ihrer Meinung nach noch, bis wir AGI oder Superintelligenz erreichen?
Holly Elmore: Das ist wirklich schwer vorherzusagen, und selbst unter Experten gibt es viele Meinungsverschiedenheiten. Einige glauben, dass wir AGI innerhalb der nächsten Jahre

erreichen könnten. Andere glauben, dass es noch Jahrzehnte dauern wird. Ein Teil des Problems besteht darin, dass wir nicht einmal eine klare Definition davon haben, was AGI ist, sodass es schwer zu sagen ist, wann wir es erreicht haben.
Klar ist jedoch, dass die Fortschritte schneller voranschreiten, als die meisten Menschen noch vor wenigen Jahren erwartet hätten. Der Sprung von GPT-3 zu GPT-4, die Fähigkeiten von Systemen wie Claude und Gemini – das sind bedeutende Fortschritte in kurzer Zeit. Und die Unternehmen investieren enorme Summen, um diese Systeme noch weiter zu skalieren.
Selbst wenn AGI noch Jahrzehnte entfernt ist, was ich hoffe, bewegen wir uns dennoch sehr schnell in diese Richtung. Das Problem ist, dass wir das Problem der Ausrichtung noch nicht gelöst haben. Wir wissen nicht, wie wir diese Systeme sicher machen können. Jeder Schritt, den wir in Richtung einer leistungsfähigeren KI machen, ist also ein Schritt näher an einer potenziellen Katastrophe, und wir machen diese Schritte immer schneller.
Michael Holmes: Und das Beängstigende daran ist, dass wir selbst wenn wir einen klaren Zeitplan hätten, selbst wenn wir genau wüssten, wann wir AGI erreichen würden, immer noch keine Lösung für das Alignment-Problem hätten, oder?
Holly Elmore: Richtig. Und genau das macht die Sache so dringend. Denn es sieht nicht so aus, als wären wir auf dem besten Weg, das Problem der Angleichung rechtzeitig zu lösen. Die Geldsummen und Anstrengungen, die in die Forschung zur Leistungsfähigkeit – also zum Bau leistungsfähigerer Systeme – fließen, übersteigen bei weitem die Geldsummen und Anstrengungen, die in die Sicherheitsforschung fließen. Und selbst die Sicherheitsforschung, die betrieben wird, ist oft inkrementell und konzentriert sich auf eng gefasste Probleme, nicht auf die grundlegende Herausforderung, wie man ein superintelligentes System angleichen kann.
Wir befinden uns also in einer Situation, in der wir auf etwas zusteuern, von dem wir nicht wissen, wie wir es kontrollieren können, und wir versuchen nicht einmal wirklich, herauszufinden, wie wir es kontrollieren können. Wir gehen einfach davon aus, dass es irgendwie funktionieren wird. Und das ist verrückt.
Michael Holmes: Was gibt Ihnen Hoffnung? Was lässt Sie glauben, dass wir dieses Problem tatsächlich lösen können?
Holly Elmore: Ehrlich gesagt, was mir Hoffnung gibt, sind die Menschen. Wenn ich mit normalen Menschen darüber spreche, nicht mit Leuten aus der Branche, nicht mit Leuten, die ein persönliches Interesse daran haben, sondern einfach mit normalen Menschen, dann verstehen sie es. Sie verstehen, dass das gefährlich ist. Sie verstehen, dass wir langsamer vorgehen sollten. Und sie sind wütend, dass ihnen das ohne ihre Zustimmung aufgezwungen wird.
Und ich denke, wenn wir das mobilisieren können, wenn wir dieses Verständnis und diese Wut in politischen Druck umwandeln können, können wir Veränderungen bewirken. Denn letztendlich sind die Unternehmen nicht allmächtig. Sie unterliegen Vorschriften, sie unterliegen dem Druck der Öffentlichkeit, sie brauchen eine gesellschaftliche Legitimation, um zu agieren. Und wir haben schon zuvor gesehen, dass Bewegungen in Fragen, die unmöglich schienen, erfolgreich waren. Der Klimawandel-Aktivismus, die Anti-Atomkraft-Bewegung, die Bewegung gegen bestimmte Waffensysteme – all diese Bewegungen haben echte Auswirkungen gehabt. Ich denke also, wenn wir eine Bewegung aufbauen können, die groß und laut genug ist, können wir eine Pause erzwingen. Wir können echte Regulierung erzwingen. Wir können den Kurs, auf dem wir uns befinden, ändern.
Michael Holmes: Und wie sieht die Bewegung derzeit aus? Was macht Pause AI?
Holly Elmore: Wir sind schnell gewachsen. Wir haben Niederlassungen in den USA, in Europa und in anderen Teilen der Welt. Wir haben Proteste vor den Büros von KI-Unternehmen und vor Regierungsgebäuden organisiert. Wir haben uns mit politischen Entscheidungsträgern getroffen, um ihnen die Dringlichkeit dieses Themas verständlich zu machen.
Wir haben viel Aufklärungsarbeit geleistet, Ressourcen geschaffen, Vorträge gehalten und versucht, Menschen zu erreichen, die noch nicht Teil der KI-Sicherheitsblase sind. Denn ich

denke, eine der großen Herausforderungen besteht darin, dass dieses Thema bisher sehr nischig und sehr isoliert war. Wir müssen es zu einem Mainstream-Thema machen.
Und wir sehen erste Erfolge. Die Medien berichten vermehrt darüber, immer mehr Politiker nehmen das Thema ernst, immer mehr Menschen schließen sich der Bewegung an. Aber wir müssen noch viel größer werden. Wir müssen eine kritische Masse erreichen, damit das Thema nicht mehr ignoriert werden kann.
Michael Holmes: Was würden Sie jemandem sagen, der zum ersten Mal davon hört und denkt, dass es verrückt oder wie Science-Fiction klingt?
Holly Elmore: Ich würde sagen, ich verstehe diese Reaktion. Es klingt tatsächlich wie Science-Fiction. Die Vorstellung, dass wir etwas bauen könnten, das uns alle vernichten könnte, ist fast zu groß, um sie zu begreifen. Aber ich würde sie bitten, sich anzuschauen, was die Experten sagen. Schauen Sie sich an, was Geoffrey Hinton sagt, was Yoshua Bengio sagt, was Stuart Russell sagt. Das sind keine Verrückten. Das sind die Menschen, die diese Technologie entwickelt haben.
Und ich würde sie bitten, an andere Zeiten in der Geschichte zu denken, in denen wir gefährliche Technologien hatten und vorsichtig mit ihnen umgehen mussten. Atomwaffen sind das offensichtliche Beispiel. Wir hätten einen Atomkrieg haben können. Wir waren mehrmals kurz davor. Aber wir haben es geschafft, ihn zu vermeiden, weil wir die Bedrohung ernst genommen haben, weil wir Verträge und Sicherheitsvorkehrungen und Überprüfungsmechanismen geschaffen haben.
Und ich würde sagen, dass dies eine ähnliche Situation ist. Wir haben eine Technologie, die das Ende der Welt bedeuten könnte, und wir müssen sie mit der gebotenen Ernsthaftigkeit behandeln. Wir müssen langsamer vorgehen, wir müssen vorsichtig sein, wir müssen sicherstellen, dass wir alles richtig machen. Denn wir haben nur eine Chance. Wenn wir eine nicht ausgerichtete Superintelligenz schaffen, gibt es kein Zurück mehr.
Michael Holmes: Das ist ein guter Punkt in Bezug auf Atomwaffen. Denn ich glaube, die Menschen verstehen, dass wir es geschafft haben, einen Atomkrieg zu vermeiden, obwohl das Risiko real war. Das zeigt, dass wir mit existenziellen Risiken umgehen können, wenn wir sie ernst nehmen.
Holly Elmore: Genau. Und ich denke, das ist die richtige Analogie. Atomwaffen waren furchterregend, und das sind sie immer noch. Aber wir haben es geschafft, internationale Abkommen zu schließen, wir haben es geschafft, eine Norm gegen ihren Einsatz zu schaffen, wir haben es geschafft, das Worst-Case-Szenario zu vermeiden. Und wir können das Gleiche mit KI tun, aber nur, wenn wir jetzt handeln, bevor es zu spät ist.
Michael Holmes: Was ist mit den Leuten, die sagen: Selbst wenn wir in den USA eine Pause einlegen, oder selbst wenn die USA und China eine Pause einlegen, was ist mit den anderen Ländern? Was ist, wenn Nordkorea oder ein anderer Schurkenstaat eine AGI entwickelt?
Holly Elmore: Ich halte diese Sorge für berechtigt, aber meiner Meinung nach wird sie oft übertrieben. Die Realität ist, dass die Entwicklung modernster KI enorme Ressourcen erfordert. Man braucht enorme Rechenleistung, riesige Datenmengen und Teams hochqualifizierter Forscher. Das ist nichts, was irgendjemand in seinem Keller machen kann.
Derzeit konzentriert sich die Spitzentechnologie im Bereich der KI auf eine Handvoll Unternehmen und eine Handvoll Länder. Die USA, China, die EU und vielleicht noch ein paar andere. Wenn wir diese Akteure zu einer Pause bewegen können, haben wir den größten Teil des Risikos abgedeckt. Ja, theoretisch könnte ein anderer Akteur versuchen, sich einen Vorsprung zu verschaffen, aber er würde weit hinterherhinken und nicht über die Ressourcen verfügen, um schnell aufzuholen.
Außerdem können wir Verifizierungsmechanismen schaffen, ähnlich wie wir sie für Atomwaffen haben. Wir können die Computernutzung überwachen, wir können internationale Inspektionsregime schaffen, wir können es jedem sehr schwer machen, heimlich fortschrittliche

KI-Systeme zu entwickeln. Das ist nicht narrensicher, aber das muss es auch nicht sein. Wir müssen es nur so schwer machen, dass es sich nicht lohnt, es zu versuchen.
Michael Holmes: Und ich denke, die Alternative ist schlimmer, oder? Die Alternative ist, dass alle um die Wette laufen und niemand irgendwelche Sicherheitsvorkehrungen hat.
Holly Elmore: Genau. Der Status quo ist schrecklich. Wir befinden uns in einem Wettlauf, in dem jeder Abstriche bei der Sicherheit macht, weil er Angst hat, zurückzubleiben. Selbst wenn eine Pause nicht perfekt ist, selbst wenn ein kleines Risiko besteht, dass jemand ausbricht, ist sie immer noch viel besser als das, was wir derzeit haben.
Michael Holmes: Lassen Sie uns über die Rolle von Regierungen und Regulierung sprechen. Welche Art von Regulierung brauchen wir Ihrer Meinung nach?
Holly Elmore: Ich denke, wir brauchen mehrere Dinge. Erstens brauchen wir strenge nationale Vorschriften in den wichtigsten Ländern. Das bedeutet Gesetze, die Sicherheitstests vorschreiben, bevor man fortschrittliche KI-Systeme einsetzen darf, Gesetze, die eine Haftung für durch KI verursachte Schäden schaffen, Gesetze, die den Regulierungsbehörden die Befugnis geben, gefährliche Projekte zu stoppen.
Zweitens brauchen wir internationale Zusammenarbeit. Wir brauchen Verträge, wir brauchen internationale Gremien, die die KI-Governance koordinieren können, wir brauchen Mechanismen, mit denen Länder überprüfen können, ob andere die Vereinbarungen einhalten. Das ist schwieriger als nationale Regulierung, aber es ist unerlässlich, da KI keine Grenzen kennt.
Drittens müssen wir sicherstellen, dass die Öffentlichkeit dabei mitreden kann. Derzeit werden Entscheidungen über die Entwicklung von KI von einer Handvoll CEOs und Vorständen getroffen. Diese Entscheidungen betreffen jedoch alle, daher sollte jeder mitreden können. Wir brauchen demokratische Kontrolle, wir brauchen öffentliche Beteiligung, wir brauchen Transparenz darüber, was diese Unternehmen tun.
Michael Holmes: Und glauben Sie, dass die Regierungen dies derzeit ernst genug nehmen?
Holly Elmore: Nein, überhaupt nicht. Ich meine, es gab zwar einige Fortschritte. Der EU-KI-Akt ist ein Schritt in die richtige Richtung, auch wenn er viele Schwächen hat. Die Biden-Regierung hat einige Durchführungsverordnungen zur KI-Sicherheit erlassen. Aber das sind nur kleine Schritte im Vergleich zu dem, was wir brauchen.
Die meisten Regierungen denken immer noch so: Wir müssen unsere KI-Industrie unterstützen, wir müssen wettbewerbsfähig sein, wir dürfen nicht zu viel regulieren, sonst fallen wir zurück. Und genau das ist falsch. Was wir brauchen, ist, dass die Regierungen erkennen, dass dies ein existenzielles Risiko ist, und dass ihre oberste Priorität darin bestehen sollte, Katastrophen zu verhindern, und nicht darin, ihre heimischen KI-Unternehmen zu fördern.
Michael Holmes: Was würden Sie einem Politiker sagen, der davon hört und etwas tun möchte, aber nicht weiß, wo er anfangen soll?
Holly Elmore: Ich würde sagen: Beginnen Sie damit, sich selbst zu informieren. Sprechen Sie mit Experten, nicht nur mit Leuten aus KI-Unternehmen, sondern auch mit unabhängigen Forschern, Menschen, die sich mit globalen Katastrophenrisiken beschäftigen, und Menschen aus der KI-Sicherheitsgemeinschaft, die bereit sind, kritisch zu sein.
Und dann nutzen Sie Ihre Plattform. Halten Sie Reden zu diesem Thema. Führen Sie Anhörungen durch. Machen Sie Ihren Wählern und anderen politischen Entscheidungsträgern klar, dass dies ein ernstes Thema ist. Denn derzeit sprechen die meisten Politiker nicht über KI-Risiken, und dieses Schweigen ermöglicht es den Unternehmen, weiter voranzustürmen.
Und dann drängen Sie auf konkrete politische Maßnahmen. Unterstützen Sie Gesetze, die Sicherheitsanforderungen für fortschrittliche KI schaffen. Unterstützen Sie eine Aufstockung der Mittel für die KI-Sicherheitsforschung. Unterstützen Sie die Idee einer Pause oder eines Vertrags. Sie müssen nicht alle Antworten parat haben, aber Sie können die Diskussion in die richtige Richtung lenken.

Michael Holmes: Eine Sache, die mir in Gesprächen zu diesem Thema aufgefallen ist, ist, dass sich die Menschen oft machtlos fühlen. Sie denken: „Ich bin nur eine Person, was kann ich schon ausrichten?“ Wie reagieren Sie darauf?
Holly Elmore: Ich verstehe dieses Gefühl vollkommen, denn das Ausmaß des Problems ist so groß und die Kräfte, die die KI-Entwicklung vorantreiben, sind so mächtig. Aber ich glaube, die Menschen unterschätzen ihre eigene Macht. Bewegungen bestehen aus Einzelpersonen. Jede große Veränderung in der Geschichte begann mit ein paar Menschen, die beschlossen, etwas zu tun.
Und die Sache ist die: In dieser Frage steht die Öffentlichkeit tatsächlich auf unserer Seite. Umfrage um Umfrage zeigt, dass die meisten Menschen wollen, dass KI sorgfältiger entwickelt wird, sie wollen mehr Regulierung, sie wollen mehr Sicherheitsmaßnahmen. Wir versuchen also nicht, die Menschen von Grund auf zu überzeugen. Wir versuchen, Menschen zu aktivieren und zu organisieren, die bereits unserer Meinung sind, aber nicht wissen, was sie dagegen tun können.
Was kann eine einzelne Person also tun? Sie können mit Ihren Freunden und Ihrer Familie sprechen. Sie können Ihren Abgeordneten schreiben. Sie können sich Pause AI oder einer anderen Organisation anschließen, die sich mit diesem Thema befasst. Sie können protestieren. Sie können spenden. Sie können Ihre Fähigkeiten und Ihre Plattform nutzen, um das Bewusstsein zu schärfen.
Und wenn man all diese individuellen Aktionen zusammenzählt, entsteht eine Bewegung. So erzeugt man politischen Druck. So verändert man die Debatte und letztlich auch die Politik.
Michael Holmes: Ich halte das für sehr wichtig. Denn ich glaube, viele Menschen haben das Gefühl, dass ihnen das aufgezwungen wird und sie keinen Einfluss darauf haben. Ich halte es für sehr wichtig, dieses Gefühl der Handlungsfähigkeit zurückzugewinnen.
Holly Elmore: Absolut. Und ich denke, das ist einer der schädlichsten Aspekte der Arbeitsweise von KI-Unternehmen. Sie schaffen dieses Gefühl der Unvermeidbarkeit, als würde das einfach passieren, man könne es nicht aufhalten und sollte es besser akzeptieren. Aber das stimmt nicht. Das ist eine Entscheidung. Das sind Entscheidungen, die von bestimmten Menschen getroffen werden, und diese Entscheidungen können hinterfragt und geändert werden.
Wir müssen das nicht akzeptieren. Wir müssen nicht zulassen, dass eine Handvoll Unternehmen auf etwas zusteuern, das uns alle zerstören könnte. Wir können Nein sagen. Wir können eine Pause verlangen. Wir können Sicherheit verlangen. Und ich glaube, dass immer mehr Menschen das erkennen.
Michael Holmes: Lassen Sie uns über die wirtschaftlichen Aspekte sprechen. Denn ich denke, ein Argument, das man oft hört, ist, dass so viel Geld auf dem Spiel steht, dass KI einen so hohen wirtschaftlichen Wert hat, dass wir es uns nicht leisten können, langsamer zu werden. Wie reagieren Sie darauf?
Holly Elmore: Zunächst einmal müssen wir uns darüber im Klaren sein, wer dieses Geld tatsächlich verdient. Es sind nicht die normalen Menschen. Es sind eine Handvoll Tech-Milliardäre und Investoren. Wenn also gesagt wird, wir könnten es uns nicht leisten, langsamer zu werden, dann meinen die Leute in Wirklichkeit, dass diese Unternehmen es sich nicht leisten können, langsamer zu werden, weil das ihre Gewinne schmälern würde.
Und ich denke, wir müssen bereit sein zu sagen: Na und? Wenn die Wahl zwischen Sam Altman, der eine weitere Milliarde Dollar verdient, und dem Fortbestand der Menschheit liegt, ist das meiner Meinung nach eine ziemlich einfache Entscheidung. Wir sollten nicht die Gewinne von Technologieunternehmen optimieren, wenn es um existenzielle Risiken geht.
Und zweitens denke ich, dass das wirtschaftliche Argument eigentlich verkehrt herum ist. Wenn wir eine nicht ausgerichtete KI schaffen, gibt es keine Wirtschaft. Es gibt keine Wertschöpfung. Es gibt nichts. Die Vorstellung, dass wir aus wirtschaftlichen Gründen weitermachen müssen, ist also völlig unlogisch. Wir opfern langfristiges Überleben für kurzfristige Gewinne.

Michael Holmes: Richtig. Das ist so, als würde man sagen, wir können es uns nicht leisten, nicht zu rauchen, weil die Tabakindustrie so profitabel ist.
Holly Elmore: Genau. Und wir haben es geschafft, den Tabak zu regulieren, obwohl es sich um eine riesige Industrie mit enormer politischer Macht handelte. Wir haben es geschafft, weil die Argumente für die öffentliche Gesundheit überwältigend waren. Und ich denke, dasselbe gilt auch hier. Die existenziellen Risiken, die für eine Pause in der KI-Entwicklung sprechen, sind überwältigend. Wir müssen nur sicherstellen, dass die Menschen das verstehen.
Michael Holmes: Was ist mit dem Argument, dass KI so viele Vorteile mit sich bringt – Heilung von Krankheiten, Lösung des Klimawandels, Beendigung der Armut –, dass wir eine moralische Verpflichtung haben, sie so schnell wie möglich zu entwickeln?
Holly Elmore: Ehrlich gesagt halte ich das für ein wirklich manipulatives Argument. Denn wir sagen ja nicht, dass KI niemals entwickelt werden soll und dass wir niemals in den Genuss dieser Vorteile kommen sollen. Wir sagen, dass die fortschrittlichsten Systeme ausgesetzt werden sollten, bis wir sie sicher machen können. Wir können die KI, über die wir derzeit verfügen, weiterhin für all diese nützlichen Zwecke einsetzen. Wir können weiterhin medizinische Forschung betreiben, wir können weiterhin am Klimawandel arbeiten.
Und die Vorstellung, dass wir uns auf eine allgemeine künstliche Intelligenz stürzen müssen, um diese Probleme zu lösen, basiert auf keinerlei Beweisen. Tatsächlich glaube ich, dass eine nicht ausgerichtete allgemeine künstliche Intelligenz all diese Probleme verschlimmern würde, anstatt sie zu verbessern. Wenn Ihnen also die Heilung von Krankheiten oder die Lösung des Klimawandels wirklich am Herzen liegen, ist es am wichtigsten, sicherzustellen, dass wir dabei nichts erschaffen, was uns alle zerstört.
Michael Holmes: Ich möchte noch einmal auf etwas zurückkommen, das Sie zuvor über die Kooptierung von Menschen in der KI-Sicherheitsgemeinschaft gesagt haben. Können Sie dazu mehr sagen?
Holly Elmore: Ja, ich glaube, es gab diese Dynamik, dass viele kluge, wohlmeinende Menschen in den Bereich KI-Sicherheit gegangen sind, weil sie dachten, sie könnten die Dinge von innen heraus verändern. Sie sind zu den großen KI-Unternehmen gegangen, haben Sicherheitsteams gegründet und wollten die Dinge verbessern und sicherer machen.
Und ich denke, in einigen Fällen haben sie gute Arbeit geleistet. Aber ich denke auch, dass sie in gewisser Weise neutralisiert wurden. Denn sobald man einmal im Unternehmen ist, sobald das Gehalt vom Unternehmen abhängt, sobald Freunde und Kollegen alle an diesen Systemen arbeiten, wird es sehr schwer zu sagen: Eigentlich sollten wir damit ganz aufhören.
So kommt es zu einer Situation, in der die Sicherheitsleute daran arbeiten, die Systeme ein wenig sicherer und ein wenig besser aufeinander abgestimmt zu machen, aber sie hinterfragen nicht die grundlegende Prämisse: Sollten wir diese Systeme überhaupt entwickeln? Und ich halte das für ein Problem, denn ich denke, dass wir uns genau diese Frage stellen müssen.
Michael Holmes: Und ich denke, die Unternehmen sind froh, Sicherheitsbeauftragte zu haben, weil sie ihnen Rückendeckung geben, oder? Sie können sagen: „Seht her, wir nehmen die Sicherheit ernst, wir haben all diese Forscher, die daran arbeiten."
Holly Elmore: Genau. Das ist Safety-Washing. Es ist wie Greenwashing, nur für KI. Sie erwecken den Anschein, als würden sie die Sicherheit ernst nehmen, aber im Grunde ändern sie nichts an ihren Plänen. Sie versuchen immer noch, so schnell wie möglich eine AGI zu entwickeln. Sie legen nur einen Anstrich von Sicherheitsforschung darüber.
Und ich denke, wir brauchen Menschen, die bereit sind, sich außerhalb dieses Systems zu bewegen, die bereit sind, sich zu widersetzen, die bereit sind zu sagen: Nein, das ist nicht gut genug. Und das ist ein Teil dessen, was Pause AI zu erreichen versucht. Wir versuchen, Raum für diese Art von Kritik, diese Art von Aktivismus zu schaffen, der nicht davon abhängt, dass man sich bei den Tech-Unternehmen beliebt macht.
Michael Holmes: Lassen Sie uns über das öffentliche Bewusstsein sprechen. Wie bewusst ist sich die Öffentlichkeit Ihrer Meinung nach derzeit der Risiken der KI?

Holly Elmore: Ich denke, das Bewusstsein wächst, aber es ist immer noch ziemlich gering. Die meisten Menschen haben von ChatGPT gehört, die meisten Menschen haben eine vage Vorstellung davon, dass KI sich rasant weiterentwickelt. Aber ich glaube nicht, dass die meisten Menschen die Dimension des existenziellen Risikos verstehen. Sie verstehen nicht, dass wir möglicherweise etwas entwickeln, das uns alle umbringen könnte.
Und ich denke, ein Teil des Problems ist die Art und Weise, wie darüber in den Medien berichtet wird. Oft ist die Berichterstattung sehr technikorientiert, sehr technisch oder sie wird so sensationell aufbereitet, dass es wie Science-Fiction wirkt. Also schalten die Leute entweder ab, weil es zu kompliziert ist, oder sie lehnen es ab, weil es zu verrückt klingt.
Wir müssen bessere Wege finden, um dies zu kommunizieren. Wir müssen es klar, konkret und nachvollziehbar machen. Wir müssen den Menschen helfen zu verstehen, dass dies kein abstraktes Zukunftsproblem ist, sondern dass es jetzt geschieht und sie, ihre Familien und alle, die ihnen wichtig sind, betrifft.
Michael Holmes: Was ist Ihrer Meinung nach der effektivste Weg, um dies Menschen zu vermitteln, die damit nicht vertraut sind?
Holly Elmore: Ich denke, man muss mit den Grundlagen beginnen. Erklären Sie, was KI ist, erklären Sie, wie sie sich weiterentwickelt, erklären Sie, warum sie sich von anderen Technologien unterscheidet. Und dann erklären Sie das Alignment-Problem in einfachen Worten. Sie müssen nicht auf technische Details eingehen. Sie müssen nur die Grundidee vermitteln, dass wir etwas entwickeln, das vielleicht intelligenter ist als wir selbst, und dass wir nicht wissen, wie wir es kontrollieren können.
Und dann muss man es meiner Meinung nach persönlich machen. Man sollte nicht abstrakt über das Aussterben der Menschheit sprechen. Man sollte über seine Kinder, seine Eltern, seine Freunde sprechen. Man sollte über die Menschen sprechen, die man liebt, und darüber, wie sich das auf sie auswirken könnte. Denn ich glaube, wenn die Menschen erkennen, dass es um das Überleben aller geht, die ihnen wichtig sind, wird es viel realer.
Michael Holmes: Ja, ich finde das sehr wichtig. Denn es ist leicht, abstrakt über das Ende der Menschheit zu philosophieren, aber wenn man an seine eigenen Kinder oder Eltern denkt, sieht die Sache ganz anders aus.
Holly Elmore: Genau. Und ich glaube, Menschen neigen dazu, über große Themen sehr abstrakt und philosophisch nachzudenken, wodurch sie weniger dringlich oder weniger real erscheinen. Aber das hier ist nicht abstrakt. Es geht um Ihr Leben, Ihre Familie, Ihre Zukunft. Und ich glaube, wenn die Menschen das wirklich verinnerlichen, werden sie wütend. Und diese Wut ist gesund. Diese Wut ist das, was wir brauchen, um eine Bewegung anzutreiben.
Michael Holmes: Lassen Sie uns noch einmal über Hoffnung sprechen. Denn ich glaube, dieses Gespräch könnte bei den Menschen ein Gefühl der Hoffnungslosigkeit hinterlassen. Wie sieht Ihre Vision für eine positive Zukunft aus? Wie sieht Erfolg aus?
Holly Elmore: Erfolg sieht so aus, dass wir die Entwicklung der fortschrittlichsten KI-Systeme vorübergehend stoppen. Wir nehmen uns die Zeit, um das Alignment-Problem tatsächlich zu lösen und herauszufinden, wie wir diese Systeme sicher aufbauen können. Wir schaffen robuste internationale Governance-Rahmenwerke. Und wenn wir all das getan haben, wenn wir tatsächlich wissen, was wir tun, können wir die Entwicklung auf vorsichtige, kontrollierte Weise wieder aufnehmen.
Und in dieser Zukunft könnten wir tatsächlich die Vorteile nutzen, von denen die Leute sprechen. Wir könnten eine KI bekommen, die uns hilft, Krankheiten zu heilen, die uns hilft, den Klimawandel zu lösen, die das Leben der Menschen wirklich verbessert. Aber wir bekommen das ohne das existenzielle Risiko, weil wir uns die Zeit genommen haben, es richtig zu machen. Und ich glaube, dass wir das schaffen können. Das glaube ich wirklich. Denn ich denke, dass die meisten Menschen, wenn sie die Situation verstehen, nicht auf eine mögliche Auslöschung zusteuern wollen. Sie wollen sicher sein. Sie wollen, dass ihre Familien sicher sind. Und wenn

wir das mobilisieren können, wenn wir das in politischen Willen umsetzen können, können wir den Kurs ändern.
Michael Holmes: Und ich denke, die Kernbotschaft ist, dass wir das nicht akzeptieren müssen. Das ist nicht unvermeidlich.
Holly Elmore: Genau. Das ist das Wichtigste. Das ist eine Entscheidung. Wir entscheiden uns dafür, auf die AGI zuzusteuern. Und wir können uns anders entscheiden. Wir können uns dafür entscheiden, langsamer zu werden. Wir können uns dafür entscheiden, vorsichtig zu sein. Wir können uns dafür entscheiden, die Sicherheit an erste Stelle zu setzen. Aber wir müssen diese Entscheidung jetzt treffen, bevor es zu spät ist.
Michael Holmes: Gibt es etwas, das ich Sie nicht gefragt habe, das Ihrer Meinung nach wichtig ist, damit die Menschen es verstehen?
Holly Elmore: Ich denke, das Wichtigste ist einfach, dass Sie ein Recht darauf haben, eine Meinung dazu zu haben. Sie haben ein Recht darauf, sich darüber Sorgen zu machen. Sie müssen kein KI-Experte sein. Sie brauchen keinen Abschluss in maschinellem Lernen. Sie müssen nur ein Mensch sein, der nicht sterben will und nicht mit ansehen will, wie alle, die er liebt, sterben. Und ich glaube, dass die Tech-Unternehmen und einige Leute aus der KI-Community versuchen werden, Ihnen das Gefühl zu geben, dass Sie nicht qualifiziert sind, eine Meinung zu haben, dass das Thema für normale Menschen zu kompliziert ist, um es zu verstehen. Aber das ist Blödsinn. Das betrifft jeden, also hat jeder ein Mitspracherecht.
Und wenn sie diese Technologie entwickeln wollen, wenn sie auf AGI hinarbeiten wollen, dann liegt es an ihnen zu beweisen, dass sie sicher ist. Es liegt nicht an Ihnen zu beweisen, dass sie gefährlich ist. Sie sind es, die die Welt auf potenziell katastrophale Weise verändern. Sie sind es, die das rechtfertigen müssen.
Michael Holmes: Ich finde, das ist ein perfekter Schlusspunkt. Holly, vielen Dank für dieses Gespräch. Ich finde, Sie haben mir tatsächlich viel Hoffnung gegeben, obwohl wir über ziemlich beängstigende Dinge gesprochen haben.
Holly Elmore: Ja, es ist viel zu verdauen. Aber ich hoffe, die Botschaft, die die Menschen mitnehmen, ist einfach: Sie haben das Recht, Ihre Meinung zu sagen. Sie stehen im Weg des Feuers, also können Sie Ihre Meinung sagen. Sie müssen nicht alle technischen Details verstehen. Man muss nicht die ganze Politik verstehen. Man kann einfach sagen: Ich finde, das sollte nicht weiter vorangetrieben werden, bis wir herausgefunden haben, wie wir es sicher machen können. Und das ist eine völlig berechtigte Position.
All das, was damit zu tun hat, wie man es sicher macht, kann nach der Pause kommen. Dafür ist die Pause da. Ich hoffe also, dass sich die Menschen dadurch gestärkt fühlen. Sie müssen das nicht akzeptieren. Sie müssen sich nicht rechtfertigen, indem Sie klug oder ein Experte sind. Sie haben das Recht, Ihre Sicherheit einzufordern. Und Sie wissen nicht, ob Sie im Moment sicher sind. Sie haben also jedes Recht zu sagen: „Ich möchte nicht, dass das weitergeht, bis ich überzeugt bin. Wenn sie so großartig sind, wenn sie so klug sind, können sie uns überzeugen. Sie können uns zeigen, dass es sicher ist."
Michael Holmes: Amen. Und ich möchte nur erwähnen, dass wir auch Pause AI Germany haben, also können Sie deren Website besuchen und sie dort unterstützen. Vielen Dank, Holly. Wir werden gewinnen. Ich hoffe, wir sehen uns irgendwann in der Zukunft wieder.
Holly Emore: Danke für die Einladung.