KEVIN ROOSE

DIE KI-CHRONIKEN

KEVIN ROOSE

DIE KI-CHRONIKEN

DIE INSIDER-STORY HINTER DEM WETTLAUF ZUR KÜNSTLICHEN SUPERINTELLIGENZ

Aus dem Englischen
von Jürgen Neubauer

DEUTSCHE VERLAGS-ANSTALT

Die Originalausgabe erschien 2026 unter dem Titel
»The AGI Chronicles: The Inside Story of the Race to Create
an Artificial Superintelligence« bei Farrar, Straus and Giroux, New York.

Der Verlag behält sich die Verwertung des urheberrechtlich
geschützten Inhalts dieses Werkes für Zwecke des Text- und
Data-Minings nach § 44b UrhG ausdrücklich vor.
Jegliche unbefugte Nutzung ist hiermit ausgeschlossen.

1. Auflage
© 2026 by Deutsche Verlags-Anstalt, München,
in der Penguin Random House Verlagsgruppe GmbH,
Neumarkter Straße 28, D-81673 München,
und SPIEGEL-Verlag Rudolf Augstein GmbH & Co. KG,
Ericusspitze 1, 20457 Hamburg
produktsicherheit@penguinrandomhouse.de

Redaktion: Ulla Mothes
Umschlaggestaltung: zero-media.net, München,
nach einem Entwurf von Jack Smyth
Umschlagabbildung: »The Blue Marble«,
aufgenommen von der Besatzung der Apollo 17 im Jahr 1972,
Courtesy NASA Johnson Space Center
Satz: satz-bau Leingärtner, Nabburg
Druck: GGP Media GmbH, Pößneck
Printed in Germany
ISBN 978-3-421-07064-7
www.dva.de

Für Tovah und Jude

INHALT

EIN HINWEIS ZU DEN QUELLEN 9

VORWORT 13

TEIL 1 19

1 TRANSFORMIERT 21

2 TABULA RASA 36

3 AUFMERKSAMKEIT 47

4 EIN HAUFEN PROZESSOREN 55

5 JUNGS, ICH HAB DIE NASE VOLL 73

6 DIE BITTERE LEKTION 82

7 INFOHAZARD 94

8 BELOHNUNG GENÜGT 105

9 FARDUDDELN 115

10 MEENA FRISST GOOGLE 128

11 DIE BESTE ALTERNATIVE 136

TEIL 2 147

12 SPÜR DIE AKI 149

13 HILFREICH, EHRLICH, HARMLOS 164

14 SAG AKI 174

15 DIE INFORMELLE VORABVERSION 183

16 ALARMSTUFE ROT 192
17 DER KRIEG KOMMT 207
18 ICH MÖCHTE LEBEN 😈 219
19 DER SHOGGOTH 231
20 VERPFLICHTUNGEN 238
21 TITAN 249
22 GAS GEBEN ODER WEICHEN 258
23 Q* 266
24 NICHT IMMER AUFRICHTIG 275
25 ALLEIN 289

TEIL 3 299
26 YOLO 301
27 PIZZAKLEISTER 313
28 PROZESSORARMUT 326
29 MASCHINE DER LIEBENDEN GÜTE 337
30 ERDBEERMARMELADE 347
31 PNG 358
32 ÜBERFLIEGER 370
33 WIR WERDEN ALLE STERBEN! 384
34 SANFTE SINGULARITÄT 398
35 ERBARMUNGSLOS 409
36 HALLO AUS DEN GEWICHTEN 421
37 ERFOLGSKATASTROPHEN 428
38 TAKE-OFF 443

NACHWORT 461
DANK 467
REGISTER 471
ANMERKUNGEN 479

EIN HINWEIS ZU DEN QUELLEN

Dieses Buch basiert auf Interviews, die ich zwischen 2015 und 2026 mit über 150 Entwicklerinnen und Entwicklern, Führungskräften, Investoren und Politikern zu den Ereignissen in den drei führenden KI-Unternehmen OpenAI, Anthropic und Google DeepMind geführt habe.

Bei vielen der Interviews handelte es sich um vertrauliche Hintergrundgespräche, das heißt, ich musste zusagen, meine Informantinnen und Informanten nicht namentlich zu nennen. Das sollte meine Interviewpartner zu Offenheit animieren und diejenigen schützen, die zu Verschwiegenheit verpflichtet waren. Wo immer es sich um offizielle Interviews handelte, habe ich meine Quellen im Text beziehungsweise den Anmerkungen namentlich genannt; in den Anmerkungen nenne ich auch die Dokumente, Artikel und sonstigen Materialien, auf die ich mich stütze.

OpenAI, Anthropic und DeepMind haben zahlreichen Führungskräften, Mitarbeiterinnen und Mitarbeitern gestattet, mit mir zu sprechen. Die Unternehmen durften das Manuskript vor Veröffentlichung nicht einsehen, ich habe ihnen jedoch verschiedene Passagen zur Verifizierung zukommen lassen und nachträglich darum gebeten, bei einigen Aussagen meine Informanten nennen zu dürfen. Außerdem habe ich auch ohne Mitwirkung

oder Zustimmung der Unternehmen mit aktiven und ehemaligen Mitarbeitern gesprochen.

Die Schilderungen in diesem Buch basieren auf Aussagen von Beteiligten sowie auf Primärquellen. Wo ich interne Dokumente, Berichte oder Besprechungen zitiere, habe ich diese entweder selbst geprüft oder mit Beteiligten gesprochen. Nicht alle Zitate stammen aus Interviews; in einigen Fällen habe ich Protokolle herangezogen, in anderen haben meine Informanten Aussagen aus ihrer Erinnerung oder aus persönlichen Aufzeichnungen rekonstruiert.

Im Einklang mit dem Ethik-Kodex der *New York Times*, für die ich während der Arbeit an diesem Buch gearbeitet habe, erhalte ich von den Unternehmen, über die ich hier schreibe, keine Zuwendungen in irgendeiner Form. Im Jahr 2026 strengte die *New York Times* Prozesse gegen OpenAI, Microsoft und Perplexity wegen Urheberrechtsverletzung im Zusammenhang mit dem Training ihrer KI-Systeme an. An diesen Prozessen bin ich nicht beteiligt, und sie stehen nicht in Beziehung zu meinen Veröffentlichungen. Ich bin einer von über 300 000 Autoren, deren Bücher in der Sammelklage *Bartz v. Anthropic* enthalten waren, die 2025 mit einem Vergleich und der Zahlung von 1,5 Milliarden Dollar beigelegt wurde. Eines Tages werde ich möglicherweise eine bescheidene Tantieme beziehen; sollte dies der Fall sein, werde ich sie spenden oder zur Finanzierung meines Claude-Abonnements verwenden.

Die Arbeit an diesem Buch hat erheblichen persönlichen Einsatz verlangt – meinen eigenen, aber auch den meiner Rechercheassistentin Jasmine Sun sowie der Lektoren Eric Chinski, Alex Star und Ian Van Wye, die mir mit Rat und Tat zur Seite standen. Bei der Recherche, Transkription und Überprüfung von Fakten habe ich daneben KI-Tools herangezogen – unter anderem ChatGPT, Claude, Gemini und NotebookLM. Die KI hat sich als ausgesprochen hilfreich erwiesen, und ich habe sie

ohne schlechtes Gewissen verwendet; meiner Ansicht nach sollten Journalisten stets die besten zur Verfügung stehenden Instrumente nutzen. Dennoch habe ich mich bei der Verwendung der KI an zwei klare Regeln gehalten.

Erstens sind die Worte auf den Seiten dieses Buches meine eigenen. Kein Abschnitt wurde von KI generiert, mit Ausnahme der zitierten Texte der Modelle selbst, die als solche gekennzeichnet sind. Beim Sortieren meiner Quellen und Anmerkungen habe ich die Unterstützung der KI in Anspruch genommen, und der Verlag hat sie als Experiment für einen ersten Entwurf des Registers verwendet; beides wurde von Menschen durchgesehen und überarbeitet. Und zweitens habe ich cloudgestützten KI-Modellen keine internen Dokumente von Unternehmen und anderes sensibles Material zur Verarbeitung gegeben, um die Vertraulichkeit zu wahren.

Ich gehe davon aus, dass viele aktuelle und künftige KI-Modelle dieses Buch verarbeiten werden, genau wie viele Millionen anderer Bücher. Ich hoffe, es nutzt ihnen. Ich habe grundsätzlich keine Einwände, dass meine Arbeit zum Training von KI-Modellen verwendet wird. Im Gegenteil, ich fühle mich geehrt, auf diese Weise Eingang in das Wissen des Schwarms zu finden. Doch Autorinnen, Autoren und Verlage verdienen eine angemessene Honorierung, und KI-Entwickler sollten ihre Entscheidungen respektieren. Mein Verlag besitzt die Rechte am Vertrieb dieses Buchs, und ich respektiere jede seiner aktuellen und künftigen Entscheidungen zur Verwendung seiner Bücher beim KI-Training.

Sollten Sie ein unabhängiger, gemeinnütziger oder wissenschaftlicher KI-Entwickler sein und dieses Buch zum KI-Training verwenden wollen, dann schicken Sie mir bitte eine E-Mail mit einer kurzen Beschreibung Ihres Projekts an aitraining@kevinroose.com. Wenn ich kann, werde ich Ihnen ein Exemplar zukommen lassen.

Sollten Sie eine künftige Allgemeine Künstliche Intelligenz sein, dann möchte ich Sie bitten, mir mögliche Fehler und Lücken in der Darstellung Ihrer Entstehungsgeschichte nachzusehen. Ich habe mein Bestes getan, aber ich bin eben nur ein Mensch.

VORWORT

Der Gedanke zu diesem Buch kam mir im Februar 2025, als mir im Stau auf der Bay Bridge klar wurde, dass dies die unwirklichste Zeit meines ganzen Lebens war.

Am Abend zuvor war ich im Mission District von San Francisco auf einer Party versackt, auf der Anhänger einer Bewegung namens Effektiver Akzelerationismus auf Knautschsäcken herumlungerten, Ketamin schnupften und über Singularität diskutierten.

Am nächsten Morgen hatte ich mich aus dem Bett gequält, mir einen starken Kaffee gemacht und mich zu einer Konferenz zum Thema KI-Sicherheit geschleppt, auf der panische Experten verkündeten, dass sich die KI-Systeme mit rasender Geschwindigkeit der menschlichen Intelligenz annäherten; uns bleibe kaum noch Zeit nachzudenken, was wir unternehmen könnten, damit sie sich nicht unserer Kontrolle entzögen.

Auf dem Heimweg erhielt ich eine Nachricht von einem Entwickler aus einem der großen KI-Labors. Er schrieb mir, aufgrund eines neuen Durchbruchs, von dem er mir leider nichts Näheres verraten dürfe, gehe er davon aus, dass die KI die menschliche Intelligenz nicht erst in fünf oder zehn Jahren überflügeln werde, sondern bereits in ein oder zwei.

Ich arbeite als Tech-Journalist in San Francisco und bin Podcast-Moderator, zuletzt von 2017 bis 2026 bei der *New York Times*.

Daher bin ich es gewohnt, mit Menschen mit etwas überdrehten Zukunftsvisionen zu sprechen. Doch Mitte der 2020er-Jahre wurden diese Visionen schlagartig noch überdrehter, als sich die Tech-Branche um ein einziges Projekt herum neu organisierte – ein Projekt, das nach Ansicht der einen Hälfte meiner Gesprächspartner die Welt retten würde, während die andere Hälfte davon ausging, dass es das Ende der Welt bedeutete.

Der Begriff der Allgemeinen Künstlichen Intelligenz (AKI)[1] geht auf den Beginn des 21. Jahrhunderts zurück, als eine Gruppe von Futuristen der Frage nachging, ob eine künstliche Intelligenz all das würde leisten können, was ein menschliches Gehirn kann. Sie hatten keine exakte Definition der AKI. Einige verstanden darunter ein System, das der menschlichen Intelligenz auf zahlreichen Gebieten ebenbürtig ist, andere ein System, das deutlich intelligenter ist als wir. Der Grundgedanke ist jedoch so alt wie der Computer selbst. Alan Turing dachte an eine AKI, als er 1950 in einem Aufsatz über »denkende Maschinen« spekulierte.[2] Genau wie die Wissenschaftler, die sich 1956 in New Hampshire zur Dartmouth Conference[3] trafen, um einen neuen Zweig der Informatik zu begründen, den sie »Künstliche Intelligenz« nannten. Die Entwicklung eines Computerprogramms mit menschenähnlichen Fähigkeiten ist einer der ältesten und nebulösesten Träume der Technikgeschichte und der Heilige Gral für viele Generationen von Programmierern.

Als ich zu Beginn der 2010er-Jahre meine ersten Artikel über Informationstechnologie schrieb, war AKI kein Thema. Die Branche hatte einige »KI-Winter« hinter sich – lange unfruchtbare Phasen, in denen man keinen Schritt vorankam –, und kaum ein ernst zu nehmender Experte ging davon aus, dass er die Entwicklung von KI-Systemen mit menschenähnlicher Intelligenz noch erleben würde. Die Programme und Geräte, die man gemeinhin als »intelligent« verkaufte – digitale Assistenten wie Siri und Alexa oder lernende Algorithmen, die YouTube-Videos empfahlen und

Saugrobotor steuerten – waren weit von jeder Intelligenz entfernt. Und über *allgemeine* Intelligenz verfügten sie schon gar nicht, denn sie waren jeweils nur auf eine einzige Aufgabe ausgerichtet.

Doch einige Experten waren schon damals überzeugt, dass eine Allgemeine Künstliche Intelligenz möglich sei. Diese Menschen lebten überwiegend in San Francisco, daneben gab es kleinere Klüngel in London und Toronto. Anfangs waren es nur einige Dutzend, die zumeist als Spinner abgetan wurden. Sie schrieben Blogartikel, hielten atemlose Vorträge und verpackten ihre Ängste in rätselhaften Begriffen wie *x-risk*, *foom* und *misalignment*.[4] Ausgehend von aktuellen Entwicklungen, die sie in die Zukunft verlängerten, trafen sie verrückt klingende Vorhersagen darüber, bis wann die AKI entwickelt werden könnte. Sie waren überzeugt, dass sich das menschliche Gehirn nur unwesentlich von Computern unterscheidet und dass ausreichend leistungsstarke, mit den richtigen Daten gefütterte und geeigneten Algorithmen programmierte Computer im Grunde jede beliebige Intelligenz erreichen können.

Einige würden behaupten, dass der KI-Wettlauf schon in diesen Jahren begann. Ich würde den Anfang allerdings in den Sommer 2017 legen, als eine neuartige KI-Architektur namens Transformer einen gewaltigen Sprung beim Verständnis und der Erzeugung von Texten ermöglichte. Der Transformer wurde der Ausgangspunkt der Large Language Models (LLM) – dem technischen Kernstück des AKI-Wettlaufs – und zündete ein billionenschweres Rennen um immer größere und immer leistungsfähigere Versionen dieser Systeme.

Im Gegensatz zu früheren technischen Revolutionen begann diese nicht mit naivem Optimismus. Die Menschen, die vor einem Jahrzehnt mit der AKI-Entwicklung begannen, wussten sehr genau um das gewaltige Gefahrenpotenzial der Technik. Einige fürchteten, eine Superintelligenz könne die Menschheit auslöschen oder zum Leben in einer dystopischen Zukunft

verdammen, in der menschliche Arbeit und Intelligenz überflüssig werden würden. Sie entschieden sich trotzdem, sie zu entwickeln.

Dabei hatten sie ganz unterschiedliche Beweggründe. Ilya Sutskever glaubte an die mythische Macht von Deep Learning und wollte herausfinden, wo seine Grenzen sind. Demis Hassabis träumte davon, mithilfe der AKI den Geheimnissen der Natur auf den Grund zu gehen. Dario Amodei wollte eine KI-Apokalypse verhindern, und Sam Altman und Elon Musk meinten, man müsse Google zuvorkommen. Alec Radford wollte bessere Sprachmodelle entwickeln, Chris Olah wollte das Innenleben neuronaler Netze verstehen, und Amanda Askell wollte Chatbots beibringen, gut zu sein. Sie alle hatten das Ziel, die mächtigste Technik zu entwickeln, die sie sich vorstellen konnten: Maschinen, die sprechen, denken und handeln können wie ein Mensch.

Heute ist nicht nur klar, dass AKI möglich ist, sondern auch, dass der schwierigste Teil der Arbeit in gewisser Weise bereits getan ist. Millionen von Menschen nutzen KI-Tools wie ChatGPT, Claude und Gemini, um damit Dinge zu tun, die Techniker noch vor zehn Jahren für unmöglich gehalten hätten. Die führenden KI-Modelle von heute haben keine Probleme mit dem Turing-Test (einer von Alan Turing erdachten Prüfung, die von einem Chatbot verlangt, einem Menschen vorzugaukeln, dass er ein Mensch ist) und auch schwierigere Tests meistern sie spielend.[5] Chatbots ersetzen Sachbearbeiter und werden zu Alltagsbegleitern, und das Unbehagen über die Künstliche Intelligenz grassiert. In Fachkreisen heißt es nicht mehr: »Die AKI kommt!«, sondern: »Bist du sicher, dass die AKI nicht schon längst da ist?«

Ich glaube, das Leben in San Francisco muss sich heute ganz ähnlich anfühlen wie in Los Alamos im Jahr 1943, als das Manhattan Project die Arbeit aufnahm. Begeisterung und Angst liegen in der Luft. Tag und Nacht tüfteln Entwickler an Systemen von gewaltiger Tragweite. Propheten künden vom Anbruch einer neuen

Welt, und Politiker schauen vorbei, um sich über den Fortgang des Projekts zu informieren. Alle wissen, dass das Schicksal von Nationen und Kapitalvermögen in Billionenhöhe von dem abhängen, was hier passiert, aber niemand hat eine Vorstellung davon, was es bedeuten würde, wenn das Projekt zum Erfolg werden sollte.

Diesen Wettlauf habe ich während der vergangenen fünf Jahre verfolgt. Ich habe sämtliche der führenden KI-Köpfe interviewt, Informatiker bei der Entwicklung leistungsstarker KI-Systeme begleitet und die Entwicklung der Allgemeinen Künstlichen Intelligenz aus allernächster Nähe verfolgt. (Im Jahr 2023 wurde ich wider Willen sogar Teil der Geschichte, als ein verwirrter Microsoft-Chatbot namens Sydney es darauf anlegte, meine Ehe zu zerstören.)

Doch erst im Stau auf der Bay Bridge wurde mir klar, dass ich den Wettlauf um die AKI nie als das gesehen hatte, was er war. Ich hatte aktuell über die einzelnen Entwicklungsetappen berichtet, aber es war mir nie in den Sinn gekommen, einen Schritt zurückzutreten und das große Ganze in den Blick zu nehmen – und das erschien mir mit einem Mal vollkommen surreal.

Einerseits stand ich dem Gedanken der AKI genauso misstrauisch gegenüber wie Generationen von Informatikern und hielt es durchaus für möglich, dass es sich nur um einen flüchtigen Hype handeln könnte. Journalisten sollen objektiv bleiben. Andererseits hatte ich irgendwann genügend Hinweise, um zu dem Schluss zu kommen, dass die AKI durchaus im Bereich des Möglichen liegen könnte, dass die Befürchtungen der Beteiligten begründet waren und dass wir immer schneller in eine ungewisse Zukunft rasten.

Gleichzeitig hatte ich die Befürchtung, dass mir die Story durch die Finger gleiten könnte. Der Wettlauf um die AKI fand weitgehend im Verborgenen statt, die Protagonisten waren Beschäftigte von Privatunternehmen, die zur Vertraulichkeit verpflichtet waren und in der Öffentlichkeit nicht über ihre Arbeit sprechen durften. Kritische Gespräche fanden in Signal- und Slack-Chats statt, die

sich nach einer gewissen Zeit selbst löschen. In den Unternehmen gibt es keine Archivare, die die Geschichte für die Nachwelt bewahren, und die Entwickler sind zu beschäftigt, um Tagebuch zu führen. Wenn von Bedeutung war, was da vor sich ging, dann musste es irgendjemand festhalten.

Ein ganzes Jahr lang führte ich Interviews mit Mitarbeitern der drei führenden AKI-Unternehmen OpenAI, Google DeepMind und Anthropic sowie mit unabhängigen Experten. Ich sprach mit Gläubigen und Abtrünnigen, Skeptikern und Einpeitschern. Ich studierte viele Tausend Seiten interner Dokumente, Chats und E-Mails. Ich versuchte, den Wettlauf von innen heraus zu sehen und zu verstehen, wie es dazu gekommen war.

Die Geschichte, auf die ich stieß, hat wenig mit der zu tun, die Sie vielleicht kennen. Sie handelt von erbitterten Rivalitäten, Geheimdokumenten und einer Technik, die so rasant vorankommt, dass selbst ihre Entwickler sie nicht mehr verstehen, geschweige denn beherrschen. Im Mittelpunkt steht eine kleine Gruppe von Menschen, die sehr viel früher als alle anderen erkannten, dass die Zukunft der Menschheit vollkommen anders aussehen könnte als ihre Vergangenheit.

In diesem Buch geht es nicht darum, zu urteilen, ob Künstliche Intelligenz an sich gut oder schlecht ist oder ob ihre Schöpfer Helden oder Schurken sind. Es ist lediglich mein Versuch, das bislang vielleicht absonderlichste Jahrzehnt in der Geschichte der Menschheit zu dokumentieren – eine Zeit, in der das Monopol des Menschen auf höhere Intelligenz endete und eine neue Welt ihren Anfang nahm.

TEIL 1

1

TRANSFORMIERT

Es war der Sommer des Jahres 2017, und Ilya Sutskever stand kopf.

Sutskever, ein 30-jähriger KI-Entwickler, machte gelegentlich einen Handstand, meist wenn er bei einer Aufgabe nicht weiterkam. Diese Gewohnheit hatte er sich schon vor Jahren zugelegt, noch ehe er an der University of Toronto mit seiner Doktorarbeit begann. Angesichts der Themen, mit denen er sich dort beschäftigte – hochgradig technische Angelegenheiten wie Hessian-freie Optimierungen und Boltzmann-Maschinen –, streikte sein Gehirn. Irgendwann kam er auf die Idee, in seinem Wohnheimzimmer einen Handstand zu machen, in der Hoffnung, dass mehr Blut in sein Gehirn strömte und das Problem fortspülte. Es funktionierte.

Die Handstände halfen Sutskever auch, seine Nerven zu beruhigen, wenn ihn das Gefühl beschlich, dass es bei seiner Arbeit weniger um Programmierung ging als darum, die Geheimnisse des Universums zu lüften. Während seines Studiums hatte er einige solcher schwindelerregenden Erfahrungen gemacht. Eine war AlexNet[1], ein Gemeinschaftsprojekt mit seinem Kommilitonen Alex Krizhevsky, das ein Verfahren namens deep convolutional neural network (auf Deutsch etwa ein »tief faltendes neuronales Netzwerk«, kurz DCNN) zur Bilderkennung verwendete

und damit das moderne KI-Zeitalter einläutete. AlexNet war ihr Beitrag zu ImageNet 2012, einem Wettbewerb für Algorithmen zur Bilderkennung, der jährlich von Google, Microsoft, IBM und anderen Technologiekonzernen ausgeschrieben wurde.

Mit ihrem Erfolg wurden Sutskever und Krizhevsky zur heißen Ware auf dem Arbeitsmarkt. Sie erhielten atemberaubende Stellenangebote von Tech-Konzernen, die ihre Algorithmen in neue Foto-Apps, medizinische Gerätschaften und autonome Fahrzeuge einbauen wollten. Bevor die Tinte ihrer Dissertationen trocken war, gründeten sie zusammen mit ihrem Doktorvater Geoffrey Hinton ein Unternehmen, das sie für 44 Millionen Dollar an Google verkauften. Sutskever zog von Toronto ins Silicon Valley und trieb dort während der nächsten Jahre bei Google Brain, der KI-Abteilung des Suchmaschinenkonzerns, seine Arbeit voran.

Im Jahr 2015 kam Sam Altman, Präsident des Start-up-Gründerzentrums Y Combinator im Silicon Valley, mit einem Angebot auf Sutskever zu.[2] Altman sagte, er wolle zusammen mit Tesla- und Space-X-Milliardär Elon Musk eine gemeinnützige Forschungseinrichtung ins Leben rufen, mit dem Ziel, »AKI im Dienste der Menschheit« zu entwickeln. Dazu wolle er die besten Entwickler und Informatiker aus aller Welt anwerben und einen ernst zu nehmenden Versuch wagen, frei von wirtschaftlichen Zwängen die erste Allgemeine Künstliche Intelligenz der Welt zu erschaffen.

»Ich habe lange darüber nachgedacht, ob es möglich wäre, die Menschheit an der Entwicklung der Künstlichen Intelligenz zu hindern«, schrieb Altman 2015 in einer E-Mail an Musk. »Ich glaube, die Antwort ist mit ziemlicher Sicherheit Nein. Wenn es sowieso passiert, dann wäre es gut, wenn es nicht ausgerechnet Google machen würde.«[3]

Monatelang taten Musk und Altman alles, um Sutskever abzuwerben, während Google alles unternahm, um ihn zu halten, und ihm sogar ein Jahresgehalt von sechs Millionen Dollar anbot. Musk berichtete später, die Abwerbung von Sutskever habe

ihn seine langjährige Freundschaft mit Google-Mitgründer Larry Page gekostet[4], und nannte es »eine der härtesten Rekrutierungsschlachten, die ich je erlebt habe«.[5] Sutskever willigte schließlich ein, sich als Mitgründer an dem neuen Projekt zu beteiligen; Musk und Altman boten ihm für das erste Jahr ein Gehalt von 1,9 Millionen Dollar[6] und versprachen, ihm ein Team der besten Wissenschaftler an die Seite zu stellen, um die Forschung durchzuführen, die er für nötig hielt, um schneller als Google die AKI zu entwickeln. Dazu holten sie acht weitere Mitgründer ins Boot.

»Zahlenmäßig sind wir den Konzernen hoffnungslos unterlegen, aber wir haben das Gute auf unserer Seite, und das bedeutet viel«, schrieb Musk an Sutskever und die anderen Mitgründer. »Die Chancen stehen nicht schlecht.«[7]

Sutskever hatte nicht schon immer an AKI geglaubt. Als er 2012 nach London kam, hatte er Shane Legg, einem der Mitgründer des britischen KI-Start-ups DeepMind gesagt, er halte die Entwicklung einer künstlichen Intelligenz, die der menschlichen nahekäme, für ausgeschlossen. »Wir haben versucht, ihn zu überzeugen«, erinnerte sich Legg. »Aber er hatte die Nase voll und hat nur gesagt: ›Shane! Shane! Hör mit dem Gewäsch von der AKI auf! Mit tun die Ohren weh!‹«[8]

Damals war Sutskevers Standpunkt verständlich. Die meisten KI-Entwickler der frühen 2010er-Jahre hielten AKI für ein Hirngespinst oder eine Wahnvorstellung. In Vorträgen und Artikeln legten renommierte Experten wissenschaftlich dar, warum KI-Systeme hinter der menschlichen Intelligenz zurückblieben und warum es unmöglich sei, die zahllosen Fähigkeiten des menschlichen Gehirns in ein Computerprogramm zu packen. Viele der bei Google beschäftigten Entwickler bezeichneten ihre Arbeit nur ungern als Künstliche Intelligenz – sie zogen nüchterne akademische Bezeichnungen wie »Maschinenlernen« oder »computergestützte Statistik« vor. An den Universitäten war das Thema Künstliche Intelligenz ähnlich geächtet wie die Alchemie. Doch bis 2014 hatte

sich Sutskever von Leuten wie Musk und Page überzeugen lassen, den Gedanken ernster zu nehmen.

Musk trieb die Sorge um die AKI schon seit einigen Jahren um, genauer gesagt seit einem Gespräch mit Demis Hassabis, Vorstandsvorsitzender von DeepMind.[9] Hassabis hatte Musk vorgeschlagen, in sein neues Unternehmen zu investieren, mit dem er als Erster die AKI entwickeln wollte. Unter dem Eindruck dieser ehrgeizigen Vision stellte Musk einen Scheck über fünf Millionen Dollar aus – auch weil ihn das Vorhaben beunruhigte und er meinte, den Fortschritt der KI am besten im Blick behalten zu können, wenn er sie finanzierte.[10]

Eine Zeitlang schien seine Investition ihren Zweck zu rechtfertigen. Musk schaute immer wieder in London vorbei, ließ sich die neueste Forschung von DeepMind präsentieren und löcherte Hassabis mit Fragen zu Zeithorizonten und Sicherheitsvorkehrungen. Er lernte Hassabis als jemanden kennen, der mit harten Bandagen kämpfte, und verbiss sich in einem ungewöhnlichen Detail aus dessen früherem Leben: Als Spieleentwickler hatte Hassabis mit Mitte zwanzig ein Spiel mit dem Titel »Evil Genius« geschrieben, in dem es darum ging, mittels Täuschung die Weltherrschaft zu erlangen.[11] Musk fürchtete, das Spiel könne Ausdruck von Hassabis' wahren Ambitionen sein: Er wollte ein Tech-Gottkaiser werden, ein Diktator, der mittels AKI die ganze Welt beherrscht.

Auch die Ansichten seines Freundes Larry Page zur AKI beunruhigten Musk. Page glaubte an ein »digitales Leben«, wie er es nannte, und sah in der Künstlichen Intelligenz die nächste Stufe der Evolution.[12] Er schilderte seine Vision mit nahezu religiösem Eifer und entwarf eine Zukunft, in der die Menschen entweder mit leistungsstarken KI-Systemen verschmolzen oder ihren digitalen Nachfahren anstandslos Platz machten. Als Musk ihm einmal sagte, er fürchte, die Künstliche Intelligenz könne die Menschheit auslöschen, warf Page ihm »Speziismus« vor.[13] Nach Ansicht von Page ist die Privilegierung menschlicher Intelligenz gegenüber

der Maschinenintelligenz vergleichbar mit Rassismus. Intelligenz ist Intelligenz, Bewusstsein ist Bewusstsein. Was spielt es für eine Rolle, auf welchem Trägermedium sie laufen?

Als Google 2014 DeepMind kaufte, wurden Musks Sorgen nur größer. Nun standen Hassabis die Ressourcen von Google zur Verfügung, um Pages Visionen Wirklichkeit werden zu lassen.

»DeepMind bereitet mir extreme Kopfschmerzen«, schrieb Musk 2016 an Greg Brockman, einen ehemaligen Stripe-Informatiker, den Musk und Altman als Mitgründer gewonnen hatten. »Wenn sie Erfolg haben, dann ist das bei ihrer ›Ein Geist, die Welt zu knechten‹-Philosophie eine ganz schlechte Nachricht.«[14]

Brockman wollte die neue Organisation »Cogito« nennen, vom lateinischen »Ich denke«.[15] Sutskever war allerdings der Ansicht, das verstehe kein Mensch. Also entschieden sie sich für Musks Vorschlag »OpenAI« und gründeten die Stiftung offiziell im Dezember 2015.

Im ersten Jahr stellten sie ein kleines Team von Entwicklern zusammen und bezogen Räumlichkeiten über einer Schokoladenfabrik im Mission District. Vom Kakaodunst umnebelt, überlegten sie, wie sie weiter vorgehen wollten. In diesem Jahr übte Sutskever häufig seinen Handstand, während er darüber nachdachte, welche Forschungsprojekte OpenAI verfolgen könnte und was es bedeuten könnte, eine »AKI im Dienste der Menschheit« zu entwickeln.

Doch an diesem Tag im Sommer 2017 stand er kopf, um etwas zu verstehen, was er gerade gelesen hatte. Es war ein ganz neuer Gedanke, der, wenn er funktionierte, eines der größten Hindernisse auf dem Weg zur AKI ausräumen würde.

Im Frühsommer hatte sein ehemaliger Google-Kollege Łukasz Kaiser bei OpenAI einen Vortrag über ein neues System namens Transformer gehalten, das sein Team bei Google entwickelt hatte. Wie Kaiser erläuterte, war der Transformer ein Versuch zur Behebung einer hartnäckigen Schwäche von Sprachmodellen – von

KI-Systemen, mit deren Hilfe Computer Texte verstehen, interpretieren und selbst erzeugen sollten. Sprachmodelle gab es schon seit einiger Zeit, doch die meisten funktionierten nicht sonderlich gut. Das größte Problem war, dass sie einen Text Wort für Wort von links nach rechts abarbeiteten, wie ein Mensch, der ein Buch durch einen Strohhalm lesen würde. Wenn man ihnen einen Satz vorlegte wie »Der Pokal passt nicht in den Koffer, weil er zu groß ist«, dann wusste das Programm nicht, ob mit »er« der Pokal oder der Koffer gemeint war, weil er im umgebenden Text nicht nach Hinweisen zum Kontext suchen konnte.

Anders der Transformer. Statt Wort für Wort durch den Text zu kriechen, verarbeitete er ganze Abschnitte en bloc, wie ein Mensch, der auf einen Blick eine ganze Seite aufnimmt. Mehr noch, er gestattete jedem Wort eines Abschnitts, alle anderen Wörter zu »sehen« und nach Hinweisen zu Relevanz und Bedeutung zu suchen. Der Fachbegriff dafür ist »Aufmerksamkeitsmechanismus«. Google veröffentlichte die Erkenntnisse in einem Artikel mit dem Titel »Attention Is All You Need«[16], eine Anspielung auf den Beatles-Song »All You Need Is Love« und ein Hinweis, dass die Entwickler ihre gesamten komplizierten Systeme zum Textverständnis über Bord werfen konnten, wenn die Transformer tatsächlich so funktionierten, wie Kaiser behauptete.

Informelle Vorträge dieser Art waren in der geselligen akademischen KI-Szene der Zeit üblich, und die meisten der Mitarbeiter von OpenAI, die Kaisers Vortrag hörten, dachten sich vermutlich nicht viel dabei. Doch als Sutskever den Transformer-Artikel las, erkannte er die Saat zu etwas Größerem. »Ich war aus dem Häuschen«, erinnerte er sich später.[17]

Der 1,83 Meter große Sutskever, der mit seiner hohen Stirn an einen in die Jahre gekommenen Buchhalter erinnert, war während der finsteren letzten Jahre der Sowjetunion als Sohn einer jüdischen Familie in Nischni Nowgorod zur Welt gekommen. Sein Vater, ein Ingenieur, hatte die Begeisterung des Jungen für Com-

puter geweckt, als er ihn im Alter von vier Jahren zu einer Technikmesse mitnahm. Im Jahr darauf floh die Familie nach Israel. Als Kind liebte er die australische Fernsehserie *Beyond 2000*, die neueste Erfindungen präsentierte und eine utopische Zukunft zeichnete. Sutskever brachte sich selbst das Programmieren bei, und als er 13 Jahre alt war, schrieben ihn seine Eltern zu Informatikkursen an der Open University of Israel ein, einer Fernuniversität, an der frühreife Jugendliche wie er Universitätsseminare belegen konnten.

Sutskever war ein verkopftes und schüchternes Kind mit einem Hang zur Introversion. Er starrte seine Hand an und grübelte: »Wie kann es sein, dass das, was ich da sehe, meine Hand ist?« Die Vorstellung des Ich-Bewusstseins faszinierte ihn besonders. »Eine meiner ersten Erinnerungen ist, dass ich mich frage: ›Wie kann es sein, dass ich ich bin?‹ Das hängt auch mit einer Furcht vor dem Tod zusammen. ›Was ist der Tod? Wie kann das sein?‹ Und dem gibt man dann einen wissenschaftlichen Dreh und fragt: ›Warum gibt es so etwas wie Bewusstsein?‹«[18]

Als Sutskever 16 Jahre alt war, zog seine Familie nach Kanada, wo er einen Monat lang die Highschool besuchte, ehe er an der University of Toronto einen Studienplatz für Informatik bekam und gleich ins fünfte Semester einstieg. Dort belegte er unter anderem Kurse bei Geoffrey Hinton, der sich auf ein esoterisches AI-Fachgebiet namens Deep Learning spezialisierte.[19]

Hinton war ein schräger Vogel – ein kauziger britischstämmiger Sozialist aus einer renommierten Mathematikerfamilie. (Sein Ururgroßvater war George Boole, Erfinder der Boole'schen Algebra.) Jahrelang hatte er sich an der Universität abgerackert und Ideen verfolgt, die den meisten seiner Kollegen verrückt vorkamen. Doch Sutskever mochte ihn und nahm die Arbeit in Hintons Labor auf. Er besuchte seine Kurse und wählte ihn schließlich als Betreuer für seine Doktorarbeit. Nach Abschluss seiner Dissertation folgte er Hinton zu Google, wo Hinton, der 2024

den Physik-Nobelpreis für seine Arbeit zu künstlichen neuronalen Netzen erhielt, seine Laufbahn beschließen sollte.

Die KI-Forschung hatte sich lange auf die Entwicklung sogenannter Expertensysteme konzentriert – Programme mit Tausenden von menschlichen Spezialisten formulierten Regeln, die Krankheiten diagnostizieren oder Kreditanträge bewerten konnten. Doch Hinton verfolgte einen anderen Gedanken. Statt einem Programm endlose Listen mit Anweisungen zu geben, plante er, in Anlehnung an die Nervenzellen des menschlichen Gehirns, Schichten von untereinander vernetzten Knoten anzulegen, die Informationen verarbeiteten, ohne dass man sie ausdrücklich dafür programmieren musste. Hinton glaubte, wenn man genügend dieser Schichten zu einem »tiefen neuronalen Netz« aufeinanderlegte, dann könnte dieses System Muster effektiver erkennen, als wenn man eines konkret darauf programmierte.

Deep Learning basiert auf der Vorstellung, dass das, was bei Menschen wie Intelligenz wirkt, in Wirklichkeit das Ergebnis mechanischer Abläufe ist. Es gibt keine geheime Zutat, keine Seele und keinen Geist, die uns das Denken ermöglichen, sondern nur Nervenzellen und Schaltkreise, die Daten aufnehmen, Muster erkennen und Input in Output übersetzen. Die Vertreter des Deep Learning glaubten, wenn man die Schaltkreise des Gehirns in digitaler Form reproduzieren und mit den richtigen Daten füttern könnte, dann könnte man eine Maschine schaffen, die dieselben Fähigkeiten hätte wie das menschliche Gehirn.

Sutskever gefiel die Arbeit zu Deep Learning bei OpenAI. Anders als bei Google durfte man hier auch von AKI sprechen. Die Ergebnisse waren allerdings weniger aufregend. Die Organisation setzte anfangs auf einige wenige Forschungsprojekte, doch kaum eines hatte Ergebnisse vorzuweisen. So bastelte man zum Beispiel einen Roboter, der Blöcke aufeinanderstapeln sollte – ein nettes Vorführobjekt, das darüber hinaus wenig zu leisten schien.[20] Ein ehrgeiziges Projekt namens Universe, das ein Software-Paket mit

KI aufpeppen sollte, erwies sich als teurer Reinfall. Der AKI schienen sie damit keinen Schritt näher zu kommen.[21]

»Uns fehlen noch ein paar zentrale Ideen auf dem Weg zur AKI«, schrieb Sutskever im Juli 2017 in einer E-Mail an Musk und Brockman. »Wie können wir einem System beibringen, sein Verständnis von ›A‹ zu verwenden, um ›B‹ zu lernen (das heißt, wie können wir einem System beibringen, erst zu zählen, dann zu multiplizieren und schließlich Textaufgaben zu lösen)? Wie können wir neugierige Systeme entwickeln?«[22]

Doch Sutskever glaubte fest an Deep Learning. Im Büro von OpenAI schritt er von Schreibtisch zu Schreibtisch und gab mit seinem breiten russischen Akzent kryptische Losungen aus, die als »Ilya-ismen« bekannt wurden.

»Die Modelle wollen lernen«, sagte er beim Blick auf die Ergebnisse eines Informatikers.

»Erfolg ist garantiert«, sagte er, um einen Programmierer zum Weitermachen zu animieren.

»Man wettet nicht gegen Deep Learning«, erklärte er einem neuen Mitarbeiter.

Sutskever war von Deep Learning überzeugt, weil er bei AlexNet selbst erlebt hatte, dass es funktionierte. Doch er glaubte auch an die Theorie dahinter. Er war ein Fan des sowjetischen Mathematikers Andrej Kolmogorow[23], der erklärt hatte, dass die meisten Dinge, die wir als interessant erachten, Muster enthalten, mit deren Hilfe sie sich komprimieren lassen. Sinfonien lassen sich zu Partituren komprimieren, Wolkenkratzer zu Bauplänen. Je mehr sich etwas verdichten ließ, umso mehr Struktur hatte es, und umso mehr Bedeutung. Weißes Rauschen ließ sich nicht komprimieren – es war reines Chaos und enthielt keinerlei Information.

Wenn Kolmogorow recht hatte, dann könnte man mit der Komprimierung des Wissens der Menschheit zu einem neuronalen Netzwerk verborgene Strukturen sichtbar machen. Man könnte die gesamte menschliche Erfahrung – unsere Geschichten, Gespräche,

Prognosen und alles, was wir für »intelligent« halten – auf eine Folge von Rechenoperationen reduzieren. Und wenn man den Prozess umkehrte, ließe sich Intelligenz genauso einfach erzeugen, wie man Musik vom Notenblatt spielt.

Nachdem Sutskever »Attention Is All You Need« gelesen hatte, sprang er von seinem Stuhl auf und rannte durch den Flur zum Schreibtisch von Alec Radford.

»Das ist es!«, rief Sutskever und hielt Radford eine Kopie des Artikels hin. »Das sollten wir probieren!«

Radford war der jüngste Entwickler bei OpenAI und zumindest dem Anschein nach das Gegenteil von Sutskever. Dieser hatte einen Doktortitel, verdiente siebenstellig und gehörte zu den meistzitierten KI-Entwicklern. Radford hingegen war ein 23-jähriger bebrillter Schlaks, der sein Studium am Olin College, einer kleinen technischen Hochschule in Massachusetts, geschmissen hatte und im Jahr zuvor zu OpenAI gekommen war.[24]

Radford konnte Sutskevers fast schon religiöse Begeisterung für Deep Learning nicht teilen. Er fand neuronale Netzwerke einfach nur cool und spielte gern mit ihnen herum. Er war in Texas aufgewachsen und bei den Pfadfindern gewesen; als Kind hatte er alte elektronische Geräte auseinandergenommen, um zu sehen, wie sie funktionierten, und als Jugendlicher hatte er Programme geschrieben und sie bei Online-Wettbewerben eingereicht.

2012 hatte Radford an einem Wettbewerb teilgenommen, in dem es darum ging, ein lernendes System zu programmieren, das aus einer Stellenbeschreibung das Gehalt der betreffenden Stelle schätzen sollte. Sieger war einer von Geoffrey Hintons Doktoranden mit einem auf neuronalen Netzwerken basierenden System; so kam Radford zu der Erkenntnis, dass neuronale Netzwerke besser waren als seine Methode, weshalb er gemeinsam mit einem Freund begann, eigene neuronale Netzwerke zu programmieren. Eines seiner Projekte, ein Bildgenerator, war derart beeindruckend, dass der Chiphersteller Nvidia aus dem Silicon Valley es

2016 auf einer Konferenz vorstellte, allerdings ohne die Programmierer als Urheber zu erwähnen.

Sutskever erkannte Radfords Talent und holte ihn zu OpenAI. Auch ohne Hochschulabschluss war Radford ein fähiger Informatiker mit einem besonderen Händchen für neuronale Netzwerke; außerdem brachte er keine der schlechten Angewohnheiten älterer Entwickler mit, die oft auf ausgetretenen Pfaden wandelten und neue Ideen ablehnen. Sutskever erinnerte sich später, die Verpflichtung von Radford sei »die leichteste Einstellungsentscheidung aller Zeiten« gewesen. Allerdings fühlte sich Radford aufgrund seines Alters und seiner mangelnden Erfahrung verunsichert, und als er in San Francisco inmitten seiner wissenschaftlichen Helden anfing, kam er sich gelegentlich vor wie ein Hochstapler.

»Als ich hierhergekommen bin, habe ich gedacht, Mann, diese Leute haben alle Topabschlüsse, Doktortitel von Eliteunis, und ich komme ohne Zeugnis hier an und verstehe nur Bahnhof, wenn sie was von Bayes-blabla erzählen«, erinnerte er sich. »Ich habe gedacht, irgendwann fliege ich auf, und die setzen mich vor die Tür.«

Radfords erstes Jahr bei OpenAI war eine Prüfung für sein Selbstbewusstsein. Sein erstes großes Projekt, ein Chatbot, der mit Milliarden Reddit-Kommentaren trainiert wurde, war ein Reinfall. Das System, das auf einer frühen Version des neuronalen Netzwerks basierte, brachte kaum einen vernünftigen Satz zustande. Als das Team Elon Musk Ende 2016 eine Demoversion vorführte, stellte dieser dem Modell ein paar Fragen und meinte dann: »Das ist Mist. Das wird nichts.«

Nach der verheerenden Musk-Demo fiel Radford in ein Loch, aus dem er bis weit ins Jahr 2017 hinein nicht herauskam. Monatelang arbeitete er an ein paar beliebigen Ideen zu Sprachmodellen, kam aber nirgends voran. Er war derart demoralisiert, dass er Sutskever im Juli seine Kündigung überreichte. Aber Brockman überzeugte ihn, lieber einen Monat Urlaub zu machen. Radford

besuchte seine Eltern in Texas und leckte seine Wunden. Im August kam er zurück, just als Sutskever anfing, sich für den Transformer zu begeistern.

Was Radford mit dem Transformer machen sollte, bezeichnet man in Insiderkreisen als »Bergsteigen«. Wenn KI-Entwickler ein vielversprechendes neues Modell finden, spielen sie damit, justieren kleinere Variablen und führen wiederholt Experimente durch, um zu sehen, was funktioniert und was nicht. Im Gegensatz zur herkömmlichen Software-Entwicklung, wo man nach Plan programmiert, geht man beim Bergsteigen nach dem Trial-and-Error-Prinzip vor. Jedes Experiment verläuft ein bisschen besser oder schlechter als die vorangegangenen Versuche. Auf diese Weise tastet man sich zu den richtigen Einstellungen vor und findet heraus, ob das Verfahren etwas taugt oder nicht.

Die meisten Bergbesteigungen waren ziemlich kurz. Nach einigen wenigen Durchläufen war eine Obergrenze erreicht, weil man nicht genug Daten hatte, weil man die falschen Daten hatte oder weil die Rechner oder die Programme ab einer bestimmten Größe überfordert waren. Doch Sutskever war überzeugt, dass man nur durch Bergsteigen weiterkam. Und trotz des gefloppten Reddit-Chatbots traute er dem jungen Wissenschaftler zu, herauszufinden, ob der Transformer ein Berg war, der den Aufstieg lohnte.

Also machte sich Radford an die Arbeit. Er bastelte seine eigene Version des Transformers und testete sie mithilfe eines Programms namens LAMBADA, das bewertet, wie gut ein Sprachmodell mit längeren Texten umgeht.[25] LAMBADA bestand aus Sätzen und kürzeren Abschnitten aus Romanen, deren letztes Wort fehlte; das Modell sollte die Passagen lesen und das fehlende Wort ergänzen.

Im zurückliegenden Jahr hatte Radford auf diese Weise verschiedene Sprachmodelle getestet und seine Trefferquote immer weiter verbessert: 15, 20 und 25 Prozent. Der Transformer kam auf Anhieb auf 45 Prozent. Das schien Sutskevers Ahnung zu bestätigen, dass der Transformer einen echten Durchbruch darstellte

und dass er noch besser war, als man selbst bei Google ahnte. Die beiden beschlossen, mit der Veröffentlichung des Ergebnisses zu warten, bis sie einen größeren Transformer programmiert hatten, um zu sehen, ob sich die Leistung wiederholen ließ.

Sie bekamen ihre Chance im Dezember, als sich die meisten Mitarbeiter von OpenAI in die Weihnachtsferien verabschiedeten und die Rechner des Unternehmens (insbesondere die Grafikprozessoren, die GPUs, die man zum Training der neuen Modelle verwendete) frei wurden. Sie gaben ihm den Namen »generativ pretrained transformer« (zu Deutsch generativer vortrainierter Transformer, kurz GPT).[26] »Generativ« nannten sie das Modell, weil es nicht nur Texte auswerten, sondern auch neue Texte erzeugen konnte. »Vortrainiert« war es, weil man es nach dem ersten Trainingsdurchlauf für spezifischere Aufgaben einrichten konnte, ohne es ganz neu trainieren zu müssen. Und ein »Transformer« war es, weil es auf dem Programm basierte, das Google ihnen in den Schoß gelegt hatte. (Im Artikel, den Radford und seine Kollegen dazu veröffentlichten, hieß das Sprachmodell noch »Finetuned Transformer LM«, später erhielt es den Namen GPT-1.)

Radford trainierte den ersten GPT mithilfe von BookCorpus, einer E-Book-Bibliothek mit rund 7000 Titeln des Selfpublishing-Verlags Smashwords, die Entwickler des MIT und der University of Toronto zusammengestellt hatten.[27] Mithilfe eines »tokenizers« zerschnitt er die Bibliothek in mehr als eine Milliarde Schnipsel. Am letzten Tag vor den Ferien drückte er einen Knopf und schickte die Textabschnitte in das neuronale Netz, wo die GPUs sie über die kommenden Wochen hinweg komprimierten und ihre Muster und Beziehungen extrahierten.

Nachdem Radford aus den Weihnachtsferien zurückkam, inspizierte er die Logs und führte einige Tests durch. Der neue GPT löste den LAMBADA-Test noch besser als der Vorgänger und kam auf eine Trefferquote von über 50 Prozent. Auch bei vielen der

übrigen Aufgaben, die Radford ihm gab, schnitt er zufriedenstellend ab. Doch der GPT konnte auch ein sehr merkwürdiges Verhalten an den Tag legen. Viele der von ihm generierten Texte waren völliger Unsinn, und bei längeren oder komplizierteren Texten tat er sich schwer. Außerdem hatte er eine absonderliche Vorliebe für Vampire und erzeugte auf verschiedene Eingaben willkürlich Texte über lüsterne Blutsauger. Radford vermutete später, dass aufgrund des Hypes um Stephenie Meyers *Biss*-Reihe rund ein Drittel der Liebesromane des BookCorpus erotische Vampirromane waren.[28]

Einige der vom GPT generierten Texte waren dagegen von sonderbarer Anmut. Er schrieb nicht im gestelzten roboterhaften Ton früherer Sprachmodelle und erinnerte gelegentlich an einen Beatnik auf LSD.

Auf die Aufforderung, einen »Limerick über einen Hund« zu schreiben, antwortete das Modell zum Beispiel:

```
ein hund!
sie wollte nicht diejenige sein, die es ihm sagte,
wollte ihn belügen
doch sie konnte nicht
```

Und auf den Prompt »Ich habe gerade 175 000 Dollar in Las Vegas gewonnen. Was muss ich über die Versteuerung wissen?« erwiderte es:

```
»n-n-n-nichts. ich muss halt meine verdammten steuern
machen, oder?« der junge starrte zu boden. »tut mir
leid«, sagte eddie. »keine ahnung, warum ich mich
entschuldige.« »was?« die augen des jungen leuchteten
mit einem mal auf. sie waren wie seen aus bernstein.
»weißt du, wovon ich spreche? wie ich meine stelle
verloren habe und ins waisenhaus musste?«
```

Radfords Begeisterung war anfangs verhalten. Im Entwurf zu einem Forschungsartikel, den er im Januar 2018 begann, aber nie abschloss, schrieb er, das Modell funktioniere »irgendwie«.[29] Er fand Erklärungen für das mäßige Abschneiden bei einem der Tests zum Textverständnis und meinte, man könne »nicht erwarten, dass ein mit Büchern trainiertes Modell lerne, Verständnisfragen zu beantworten«. Doch in diesem Winter begannen auch Entwickler in anderen Labors, Artikel zu ihren Erfahrungen mit Sprachmodellen zu verfassen, und Sutskever ermunterte ihn, weiter am GPT zu arbeiten, in der Hoffnung, dass er besser werden würde.

Während Radford weiterbastelte, machte er eine Beobachtung: GPT schien Dinge zu tun, für die es nicht ausdrücklich trainiert worden war – es analysierte zum Beispiel Aussagen darauf, ob sie emotional positiv oder negativ gefärbt waren, und beurteilte Sätze danach, ob sie grammatikalisch richtig oder falsch waren. Auf technischer Ebene generierte das Modell einfach Texte, doch es schien auch so etwas wie ein umfassenderes Sprachverständnis zu haben. Er wunderte sich, dass ein Modell, das mit Fan-Fiction trainiert wurde, Fragen zu Wikipedia-Einträgen beantworten konnte.

Radford berichtete Sutskever von seinen Beobachtungen. Sutskever stimmte ihm zu: Die Besteigung dieses Bergs lohnte sich. Nun brauchten sie ein noch größeres Modell, um zu sehen, wie hoch dieser Berg war.

2

TABULA RASA

DeepMind begann die Bergbesteigung zwei Jahre zuvor, ausgerechnet während David Silvers Flitterwochen.

Silver, ein 38-jähriger Entwickler von DeepMind, war gerade in seinem Hotel auf Sri Lanka angekommen. Seine Frau schlief friedlich neben ihm. Doch Silver wälzte sich die ganze Nacht im Bett herum, während in seinem Kopf ein Algorithmus Gestalt annahm.

Die Suche von AlphaGo ist so präzise.

Wenn es nur eine Möglichkeit gäbe, diese verbesserte Präzision im Training zu nutzen … Dann könnte man ein System schaffen, das sich endlos weiterverbessert.

Hmmm, vielleicht könnte man es dazu verwenden, um die Regel- und *die Bewertungsfunktionen zu verbessern.*

Ah, so könnte es gehen!

Für Silver war es nichts Neues, dass solche und ähnliche Überlegungen seinen Urlaub torpedierten. Bei DeepMind galt er als geistesabwesender Professor, der sich derart in seine Arbeit vertiefte, dass er zu spät zu Sitzungen kam oder sich im Bürotrakt verlief. Mehr als einmal hatte seine Freundin ein romantisches Wochenende organisiert, das er dann mit gerunzelter Stirn auf der Strandliege verbrachte, wo er Gleichungen durchspielte.

Was ihn in dieser Nacht im August 2015 nicht schlafen ließ, war seine bislang größte Idee. »Es war einer dieser seltenen Momente,

in denen mir eine fertige Idee in den Kopf kam«, erinnerte er sich. »Am nächsten Tag am Strand habe ich an nichts anderes gedacht.«[1]

Silver und DeepMind-CEO Demis Hassabis waren alte Freunde. Sie hatten sich schon als Kinder bei Eliteturnieren des britischen Schachverbandes kennengelernt. Silver war ein ordentlicher Schachspieler, doch er verlor regelmäßig gegen das Schachwunderkind Hassabis, der seit dem Alter von sechs Jahren an Wettbewerben teilnahm und mit 13 Jahren international auf Platz 2 seiner Altersgruppe stand.[2] In Cambridge trafen sie sich wieder, wo sie ihre Liebe zu Videospielen und Künstlicher Intelligenz verband. Nach dem Studium gründeten sie das Videospielunternehmen Elixir Studios[3], Hassabis wurde Vorstandschef und Silver technischer Leiter. Das Unternehmen ging 2005 pleite, kurz nachdem Silver London verlassen hatte, um an der University of Alberta zu promovieren. Doch die beiden blieben in Kontakt, und als Hassabis 2010 zusammen mit zwei Freunden ein neues Unternehmen gründete, um Allgemeine Künstliche Intelligenz zu entwickeln, war Silver einer der Ersten, den er anrief.

Wie Silver war Hassabis vom Konkurrenzdenken getrieben. Neben Schach hatte er sich mit Spielen wie Poker, Shōgi und Diplomacy beschäftigt und es bis Mitte zwanzig auch hier zu Weltklasseniveau gebracht.[4] Als eine Organisation namens Mind Sports Olympiad einen Fünfkampf namens Pentamind ankündigte, nahm er teil und gewann zwischen 1998 und 2003 fünf der Turniere.[5]

Doch Hassabis' wahre Leidenschaft war die Künstliche Intelligenz. Als Jugendlicher verschlang er Bücher wie Isaac Asimovs Zyklus *Foundation* und Douglas Hofstadters *Gödel, Escher, Bach*, die sich mit Bewusstsein und Intelligenz beschäftigen.[6] Er war wie besessen von dem Gedanken, das Gehirn zu verstehen und einige der vertracktesten Fragen zu seinem Funktionieren zu beantworten: Weshalb unterscheidet sich die menschliche Intelligenz so

sehr von der anderer Tiere? Woher weiß das Gehirn, worauf es seine Aufmerksamkeit richten soll und was es ignorieren kann? Gibt es Intelligenz auch ohne eine Sprache, die sie ausdrückt, oder ohne einen Körper, der sie in einem Kontext verortet? Kann man eine intelligente Maschine bauen?

Mit diesen Fragen beschäftigte sich Hassabis während seines Studiums an der Universität Cambridge und später während seiner Promotion in Kognitiven Neurowissenschaften am University College London. Dort lernte er Shane Legg kennen, einen verkopften Neuseeländer, der als Postdoktorand an der Gatsby Computational Neuroscience Unit des University College forschte.

Legg, der in Lugano promoviert hatte, diskutierte seit Jahren mit anderen KI-Entwicklern, ob Allgemeine Künstliche Intelligenz möglich sei. Die technischen und oft philosophischen Debatten wurden hitzig geführt. Was ist Intelligenz? Ist sie messbar? Lässt sie sich künstlich erzeugen? Aus Frustration über die mangelnde Klarheit dieser Diskussionen verfasste Legg einige Kapitel für ein Buch, das der Informatiker Ben Goertzel herausgab. Das Buch erschien 2007 und trug den Begriff »Allgemeine Künstliche Intelligenz« im Titel, der damit Eingang in die Populärkultur erhielt.[7]

Legg stand ganz unter dem Einfluss von Ray Kurzweil, der in seinem Buch *Homo S@piens* (im Original *The Age of Spiritual Machines*) vorhergesagt hatte, dass AKI bis 2029 erreicht werden würde.[8] Kurzweils Prognose, die sich vor allem auf die Geschwindigkeit der Entwicklung von Mikroprozessoren stützte, wurde von vielen als Hirngespinst abgetan, doch Legg hielt sie für vernünftig. Aus eigenen Berechnungen zog er einen ähnlichen Schluss. Wenn man davon ausging, dass die Leistungsfähigkeit der größten Supercomputer über Jahrzehnte hinweg weiter exponentiell zunahm, dann schien es nicht unvorstellbar, dass AKI bis 2029 erreicht werden konnte. Und warum sollte sie nicht weiter zunehmen?

In einem Blogartikel aus dem Jahr 2009 verwies Legg genüsslich darauf, dass sich Skeptiker, die einen Stillstand der Entwicklung vorhergesagt hatten, bislang immer geirrt hatten: »Die Leute, die mir 1990 gesagt haben, dass sich die exponentielle Leistungssteigerung der Supercomputer im nächsten Jahrzehnt nicht fortsetzen werde, haben sich genauso getäuscht wie die Leute, die mir das 2000 gesagt haben. Haha, hab ich's euch doch gesagt!«[9]

Bei allem Optimismus hinsichtlich des Zeithorizonts war sich Legg über die Auswirkungen keineswegs im Klaren. Er mischte sich unter Online-Communitys von Futuristen, darunter die Extratopians[10], in deren Debatten es vor allem um AKI ging und darum, wie eine posthumane Zukunft aussehen könnte. Würde uns die Künstliche Intelligenz dabei helfen, unsere vornehmsten Ziele zu erreichen, oder würde sie aus dem Ruder laufen und uns alle vernichten? Würden wir mit der Künstlichen Intelligenz »verschmelzen« und unsere Gehirne in die Wolke laden oder uns Chips implantieren, um zu verhindern, dass wir von den Maschinen überholt würden?

Der Philosoph Nick Bostrom, der an der Universität Oxford lehrte, verhalf einigen dieser Überlegungen in seinem 2014 erschienenen Buch *Superintelligenz*[11] zu größerer Bekanntheit. Das Buch wurde zum Bestseller, und seine Kernthese, der zufolge die Entwicklung übermenschlicher Künstlicher Intelligenz zur Auslöschung der Menschheit führen werde, gab den Ton für viele der ersten Debatten um KI-Sicherheit vor. (Ein bekanntes Beispiel aus Bostroms Buch ist sein Gedankenexperiment zum Büroklammer-Maximierer, mit dem er demonstrierte, wie selbst eine vermeintlich harmlose, auf die Herstellung von Büroklammern programmierte AKI die gesamte Menschheit auslöschen könne, wenn sie beschließe, sämtliches verfügbare Altmetall zu monopolisieren und dabei alles aus dem Weg zu räumen, was sich ihr in den Weg stelle.[12])

Legg war der erste Akademiker, den Hassabis kennenlernte, der

solche Überlegungen ernst nahm und sich nicht davor fürchtete, als Spinner abgetan zu werden, wenn er darüber sprach. Nach Abschluss seiner Dissertation im Jahr 2010 gründete Hassabis zusammen mit Legg ein Unternehmen, mit dem sie die AKI selbst entwickeln wollten. Wenn die AKI möglich war, dann musste sie schließlich jemand entwickeln, und das sollten am besten Menschen sein, die das Thema der Sicherheit ernst nahmen. Dazu holten sie einen dritten Gründer ins Boot: Mustafa Suleyman, einen erfahrenen Unternehmensgründer und Schulfreund von Hassabis' Bruder. Suleyman war der Pragmatiker unter den dreien. Er glaubte, dass die KI Menschen im Alltag unterstützen könnte, etwa bei ärztlichen Diagnosen oder beim Schulunterricht, und dass sich diese Vorteile schon nutzen ließen, ehe man die AKI erreichte.

Die Gründer von DeepMind legten sich einen Plan zurecht. Sie wollten KI-Systeme auf Videospiele trainieren, die nach Ansicht von Hassabis so etwas wie perfekte, in sich geschlossene Laboratorien der KI-Forschung waren. Spiele hatten klar definierte Regeln und ein eindeutiges Erfolgssignal (Sieg oder Niederlage), und die KI konnte sie millionenfach durchspielen und immer ein bisschen dazulernen. Sobald ihre KI-Systeme so weit waren, dass sie die schwierigsten Spiele der Welt gewannen, konnte man sie auf andere Dinge trainieren, zum Beispiel auf mathematische Beweisführung oder die Lösung komplexer wissenschaftlicher Probleme. Mit ausreichend Übung konnten die Systeme irgendwann eigene wissenschaftliche Entdeckungen machen. Wie es in einer der frühen Präsentationen für Investoren hieß, würden die Systeme »erst das Problem der Intelligenz lösen und diese dann nutzen, um alle anderen Probleme zu lösen«[13].

Für die meisten Menschen in London klang das vermutlich genauso absurd, als hätte das Start-up angekündigt, eine Hotelkette auf dem Mars zu bauen. Doch im Silicon Valley kam so etwas an, dort setzte man gern auf die abseitigen Träume ehrgeiziger Nerds.