
Nick Srnicek
SILICON EMPIRES
Der Kampf um die Zukunft der KI
Hamburger Edition

Nick Srnicek

SILICON EMPIRES

Der Kampf um die Zukunft der KI

Aus dem Englischen
von Felix Kurz

Hamburger Edition

Hamburger Edition HIS Verlagsges. mbH
Verlag des Hamburger Instituts für Sozialforschung
Mittelweg 36
20148 Hamburg
verlag@hamburger-edition.de
www.hamburger-edition.de

ISBN 978-3-98722-121-7
E-Book Umsetzung: Dörlemann Satz, Lemförde

ISBN 978-3-98722-012-8

Titel der Originalausgabe: »Silicon Empires. The Fight for the Future of AI«

Umschlaggestaltung: Lisa Neuhalfen, Berlin

Inhalt

1

Wo ist der Wert generativer KI?

Seit der Veröffentlichung von ChatGPT im November 2022 ist generative Künstliche Intelligenz (KI) ins Zentrum der öffentlichen Aufmerksamkeit gerückt und hat einen regelrechten Hype ausgelöst. Die Entwicklung von Technologien, die scheinbar wie Menschen Texte, Bilder, Audio- und Videodateien erzeugen können, wird von immensen Versprechen über ihre Potenziale begleitet. Ihre euphorischsten Verfechter prophezeien, generative KI werde nahezu alle Branchen umkrempeln, die Produktivität von Unternehmen gewaltig steigern und das Wirtschaftswachstum auf Höhenflüge schicken.[1] Manche Stimmen sehen in ihr eine technologische Revolution vom Rang des Internet, der Elektrizität – oder sogar ohne historischen Präzedenzfall. Während eine Investmentfirma erklärt, KI werde sich fünfmal so stark auf das Bruttoinlandsprodukt auswirken wie die Dampfmaschine, meinen andere, sie sei »möglicherweise das Wichtigste – und Beste –, was unsere Zivilisation jemals hervorgebracht hat«.[2]

Solche überschwänglichen Behauptungen haben den KI-Diskurs immer wieder geprägt. Sie sind teilweise Ausdruck der finanziellen Zwänge, denen Risikokapitalgeber und die Geschäftsmodelle des Silicon Valley unterliegen.[3] Ebenso zeugt der Hype von der ungewissen Zukunft der neuen Technologie – ihr Potenzial ist gewaltig, aber noch so unbestimmt, dass sich verschiedenste wilde Szenarien um sie ranken können. Rückblickend war 2023 vielleicht das Jahr, in dem der Hype vollends abhob und seinen Höhepunkt erreichte. Die Grenzen der KI blieben noch im Dunkeln, ihre Fähigkeiten dagegen erstaunlich. Investoren und Unternehmen pumpten im Lauf des Jahres zig Milliarden

Dollar in das neue Geschäftsfeld.[4] Hunderte von KI-Modellen wurden lanciert und Tausende von Varianten auf ihnen aufgebaut.[5] Allein in ihren Geschäftsberichten für das dritte Quartal erwähnten die Unternehmen 35.000 Mal »KI«.[6] Die KI-Lieferketten wurden unterdessen zu einem Thema von erheblichem geopolitischem Interesse.

Im Licht dieser dramatischen Ereignisse ist das vorliegende Buch als Einführung, Überblick und Intervention in unser Verständnis der gegenwärtigen Entwicklungen konzipiert. Es soll auch für Leserinnen und Leser zugänglich sein, die sich mit diesen Fragen bislang nicht befasst haben oder eine systematische Perspektive auf die derzeitige KI-Landschaft gewinnen wollen. Mit einem neuartigen Blick auf Geschäftspraktiken, Klassenbündnisse und geopolitische Strategien versucht es aber auch in aktuelle Debatten einzugreifen. Die hier vorgelegte Analyse lässt sich als deflationär betrachten, insofern sie die Übertreibungen des Hypes wie auch der Schwarzmalerei zu vermeiden sucht, an denen Diskussionen über KI bislang kranken. Dabei stehen Künstliche Allgemeine Intelligenz und existenzielle Gefahren nicht im Mittelpunkt.[7] Ebenso wenig geht es vorrangig um unmittelbare praktische Bedenken mit Blick auf Automatisierung, verzerrte Technikentwicklung oder die Ausbeutung von Arbeiterinnen und des gemeinsamen menschlichen Wissenskorpus.[8] Vielmehr befasst sich das Buch mit der geopolitischen Ökonomie der gegenwärtigen KI – mit den Versuchen von Großunternehmen und machtvollen Staaten, ihre Entwicklung zu steuern, den dabei generierten Wert abzuschöpfen und sie als Machtmittel einzusetzen.[9] Kurz gesagt: Es geht um die »Silicon Empires«, die gegenwärtig die Zukunft der KI bestimmen. Auf den Kommandohöhen der Konzern- und Staatsmacht werden weitreichende Entscheidungen darüber getroffen, wie sich die Technologie entwickelt, welche Werte sie verkörpert und wie sie eingesetzt werden soll.[10] Diese Entscheidungen liegen in den Händen einer winzigen Elite, haben aber Folgen für Milliarden von Menschen. Und je mehr die Fähigkeiten von KI zunehmen, umso größer wird die Kluft zwischen der geringen Zahl der Entscheidungsträger und den gewaltigen Auswirkungen ihres Handelns ausfallen. Da KI weitreichende Veränderungen erzeugen wird, ist es wichtig zu verstehen, wer ihre Zukunft bestimmt, wie dies geschieht und welchen Zielen es dient.

Jenseits von ChatGPT

Doch was versteht man unter Künstlicher Intelligenz überhaupt? Immerhin hat der Ausdruck mittlerweile eine lange Geschichte. 1956 zu Marketingzwecken geprägt – er sollte ein Forschungsprojekt vom bereits damals bekannten Feld der Kybernetik abgrenzen, um Fördermittel einzuwerben –, bezeichnet er heute eine wesentlich breitere Palette an Technologien. In der gegenwärtigen Phase ist Deep Learning zum vorherrschenden Ansatz geworden. Entsprechende Systeme bauen auf künstlichen neuronalen Netzen auf, die in sehr grober Weise die Neuronen des menschlichen Gehirns nachbilden; von einem »tiefen« Lernen spricht man, weil sie viele Schichten umfassen. Der »Lernprozess« wird auf vielfältige Weise gestaltet, zumeist aber als sogenanntes überwachtes Lernen, bei dem das KI-Modell aus einem bestimmten Input einen bestimmten Output erzeugt, der anschließend mit dem richtigen Ergebnis abgeglichen wird. Inwieweit ist ein Modell beispielsweise imstande, Katzenbilder korrekt zu erkennen? Abweichungen vom richtigen Ergebnis werden dazu verwendet, mit einem »Backpropagation« oder »Fehlerrückführung« genannten Algorithmus die Gewichte und Parameter einzelner Neuronen so zu korrigieren, dass das Modell dem Zielwert näher kommt. Dieser Prozess wird zahllose Male wiederholt, um die Treffsicherheit der KI Schritt für Schritt zu verbessern. Ein Modell, das ein solches »Training« absolviert hat, kann danach neue Inputs verarbeiten. Der Rechenprozess beim Betrieb einer KI, gewöhnlich »Inferenz« genannt, hat nicht nur technische, sondern, wie wir sehen werden, auch politische und wirtschaftliche Implikationen.

Dieses Verfahren zum Aufbau von Künstlicher Intelligenz war bereits seit vielen Jahren durchaus bekannt, hatte aber überwiegend nicht die erhofften Resultate erzielt. Das änderte sich 2012, als ein Team beim Bilderkennungswettbewerb ImageNet einem neuartigen Ansatz folgte. Während die Fehlerquote in den Vorjahren bei 25 bis 30 Prozent gelegen hatte, gelang es ihm mit einem Deep-Learning-Modell, sie auf beeindruckende 15 Prozent zu senken.[11] Das Ergebnis war überraschend und bewies, dass die durch die Digitalisierung der Welt erzeugten riesigen Datenmengen zusammen mit der fortschreitenden Leistungskraft von Computern (insbesondere von Grafikprozessoren) eine effektive Anwendung der bereits bekannten Deep-Learning-Algorithmen ermöglichen.

Der von dem Team verfolgte Ansatz wird als »Convolutional Neural Network« bezeichnet und erwies sich als eine Architektur, die für die Bilderkennung und andere visuelle Aufgaben besonders gut geeignet ist. Sie verfährt ähnlich, wie das menschliche Gehirn es mutmaßlich tut, wenn es Bilder verarbeitet: durch mehrere Schichten von Neuronen, die immer detailliertere Elemente erfassen. Seit 2012 wurde allerdings auch die Forschung zu anderen Ansätzen stark vorangetrieben und hat in manchen Fällen zu beträchtlichen Fortschritten geführt.

Der mit Abstand wichtigste Ansatz ist dabei die Transformer-Architektur gewesen.[12] Sie wurde 2017 vorgestellt, um einige der Grenzen zu überwinden, die andere Architekturen – vor allem Recurrent Neural Networks (RNN) und Long Short-Term Memory Networks (LSTM) – bei Aufgaben der Sprachverarbeitung und Übersetzung gezeigt hatten.[13] Einer der bedeutendsten Durchbrüche der Transformer besteht darin, dass sie weitgehend eine parallele statt einer sequenziellen Verarbeitung ermöglichen. Dadurch konnten beim Training deutlich größere Datensätze verwendet werden, und so entstanden die riesigen Modelle, die für diese Phase charakteristisch waren.[14] Die Transformer-Architektur hat ihrerseits die sich überschneidenden Ideen der generativen KI, Large Language Models (LLM) und Foundation Models hervorgebracht. Und sie war die Basis für die jüngsten beeindruckenden Fortschritte des maschinellen Lernens. ChatGPT ist das bekannteste dieser Systeme, doch die Transformer-Architektur hat sich als erstaunlich vielseitig erwiesen und wird auch fernab der Sprachaufgaben eingesetzt, für die sie ursprünglich konstruiert wurde. Ihre Verwendung ist heute vorherrschend in der Welt der KI.[15]

Zumindest im öffentlichen Bewusstsein ist ChatGPT heute zum Paradigma der aktuellen KI geworden – ein Produkt, auf das Befürworter wie Kritikerinnen ihre Hoffnungen und Ängste projizieren können. An der Oberfläche handelt es sich um eine schlichte Schnittstelle: Die Benutzerin kann in einem Eingabefeld eine beliebige Frage stellen und das System gibt eine häufig erstaunlich solide Antwort. Ungeachtet seiner Bekanntheit führt es allerdings in die Irre, ChatGPT als repräsentativ für die heutige KI zu betrachten. Seine Probleme und Erfolge gelten dann als direkter Indikator dafür, wie es um das Feld insgesamt bestellt ist. Auch Kritiken und Verteidigungen des Systems erhalten dadurch ein größeres Gewicht, als sie vielleicht verdienen. Generative KI – und erst recht

die breitere Kategorie des modernen Deep Learning – umfasst weitaus mehr als die Form des Chatbots, für die ChatGPT emblematisch steht. Während das Programm auf den allgemeinen Verbraucher zugeschnitten ist (auch wenn sich das gerade ändert, wie wir sehen werden), wird KI heute überwiegend für Unternehmenszwecke entworfen. ChatGPT erweckt außerdem den Eindruck, Chatbots seien die wichtigsten KI-Schnittstellen, doch in Wirklichkeit werden aktuell auch viele andere Ansätze verfolgt. Transformer haben auch wesentlich breitere Einsatzfelder als lediglich die Produktion von Text oder selbst von Fotos, Videos und Audiodateien. Auf technischer Ebene erklärt sich das teilweise daraus, dass solche LLMs nicht das nächste Wort in einem Satz, sondern das nächste »Token« vorhersagen. Tokens können weitaus mehr sein als Sprachkomponenten, etwa Handlungen, Pixel oder Moleküle. Die Fähigkeit von Transformern, sie vorherzusagen, erschöpft sich daher keineswegs im Feld der Sprache, und entsprechend kommt generative KI in verschiedensten Branchen wie Werbung, Logistik, Finanzwesen, Vertrieb, Gesundheitswesen, Pharmaforschung, Bergbau, Industrie, Robotik und Hardware-Design zum Einsatz – mit Verfahrensweisen, die mit dem gängigen Bild von Chatbots wenig gemein haben.

Es hat außerdem eine Reihe von technischen Fortschritten gegeben, die die Fähigkeiten von KI weit über die ersten ChatGPT-Versionen hinaus gesteigert haben. Eine erste Weiterentwicklung bestand in der Einführung von Multimodalität in die Systeme. Während ChatGPT Text aufnehmen und ausgeben konnte, können neuere Modelle eine Vielfalt von Medien (etwa Text, Audio und Video) verarbeiten. Bis vor Kurzem wurde solche Multimodalität etwas umständlich auf textbasierte LLMs aufgesattelt: Bat man ChatGPT, ein Bild zu erstellen, gab das Programm den Auftrag an ein separates Modell (DALL-E) weiter. Heute dagegen sind die Modelle immer häufiger von vornherein multimodal, können also selbst mit diversen Medien arbeiten. Das verringert die Latenz (und steigert so die Benutzerfreundlichkeit), vermeidet Informationsengpässe, die bislang bei der Weitergabe von Anfragen zwischen verschiedenen Modellen entstehen konnten, und erhöht die Leistungsfähigkeit der Systeme, da sie nun mit ganz unterschiedlichen Modalitäten umgehen können. Multimodalität umfasst außerdem die Fähigkeit, visuelle Daten aus der Umgebung aufzunehmen – was für die Robotik und das menschliche Lernvermögen zentral ist.

Jüngere Modelle haben auch die bereits existierenden Formen von Gedächtnis verbessert und neue hinzugewonnen. In grober Analogie zum menschlichen Gehirn verfügen KI-Modelle heute über eine Art Langzeitgedächtnis in Form von »Retrieval-Augmented Generation« (RAG) und Vektordatenbanken sowie über ein Kurzzeitgedächtnis in Form eines »Kontextfensters« variabler Größe und der zunehmenden Personalisierung auf Grundlage des »Gedächtnisses« der Nutzer. Das Langzeitgedächtnis ermöglicht es der KI, auf bestehendes Wissen zurückzugreifen, um so ihre Neigung zum »Halluzinieren« zu dämpfen, das zu falschen Ausgaben führt. Ein Unternehmen kann sie zum Beispiel so einrichten, dass sie interne Dokumente zitiert, damit sie seine Geschäftspraktiken nicht falsch darstellt. Das Kurzzeitgedächtnis der Kontextfenster dagegen erlaubt es, mehr Informationen in einen Prompt einzuspeisen; wenn das System dann versucht, eine Antwort zu generieren, berücksichtigt es eine viel breitere Spanne an spezifischen Daten. Eine Studentin zum Beispiel kann das PDF eines 500 Seiten langen Lehrbuchs anhängen und das System auffordern, sich auf den gesamten Text zu beziehen; ein Gesundheits- und Arbeitsschutzinspektor kann das Video einer Baustellenbesichtigung einfügen, damit die KI nach Verstößen gegen die Sicherheitsauflagen sucht. Gegenwärtig entstehen auch neue Formen von personalisiertem Gedächtnis, die mit dem Versprechen antreten, den individuellen Kontext der Nutzerinnen stärker zu berücksichtigen – wobei sich allerdings bereits erhebliche Datenschutzprobleme abzeichnen. Fortschritte bei all diesen Arten von Gedächtnis haben neue Verwendungen hervorgebracht, die weit über die frühe Chatbot-Version von ChatGPT hinausgehen.

Ende 2024 erfolgte der bemerkenswerteste neueste Fortschritt, als OpenAI sein Modell o1 vorstellte und bereits einen Ausblick auf das Modell o3 gab.[16] Beiden gelangen bei einigen der größten Herausforderungen für KI signifikante Durchbrüche. Entscheidend dafür war, dass sie mehr Zeit auf die Inferenz verwenden durften: Anstatt so schnell wie möglich eine Antwort zu generieren, sollten sie über das gegebene Problem länger »nachdenken«, indem sie es in mehrere kleine Schritte zerlegen, verschiedene Antworten ausprobieren und ihre eigene Arbeit überprüfen. Diese Verschiebung zu mehr Rechenzeit – »Test-time Compute« genannt – ist Teil des allgemeinen Trends, dass Forscherin-

nen versuchen, ihren Modellen die Fähigkeit zu »Abwägungen« zu verleihen. Die dazu verwendeten Methoden scheinen recht einfach zu sein: Im Rahmen des »verstärkenden Lernens« werden die Modelle belohnt, wenn sie richtige Antworten generieren.[17] Durch einen Prozess von Versuch und Irrtum beginnen sie zu lernen, bei der Lösung von Problemen Schritte zu unternehmen, die einem Abwägen ähneln.

Auf einer bestimmten Ebene spiegeln diese Modelle menschliche Denkprozesse wider (was keine große Überraschung ist, da sie mit menschlicher Sprache trainiert worden sind). Wenn ihre Abwägungen sichtbar gemacht werden, wirken sie mitunter frappierend human. In einem Fall schien das KI-Modell einen Moment der Erkenntnis zu haben, als es gerade ein Problem durchdachte und erklärte: »Warte, warte. Warte. Ich habe gerade einen Aha-Moment, den ich hier anzeigen kann.«[18] Man sollte die Analogie aber nicht übertreiben, denn die für die Öffentlichkeit konzipierten Modelle werden in der Regel zusätzlich so trainiert, dass Menschen sie besser nachvollziehen können. Vor diesem Training sind sie zumeist deutlich schwieriger zu interpretieren, etwa weil sie während des Nachdenkens zwischen verschiedenen Sprachen wechseln. Und wie Forscher erklärt haben, wird der Denkprozess durch eben diese Verwandlung in Tokens, die ihn für Menschen lesbar macht, behindert, weshalb zukünftige Modelle wesentlich undurchsichtiger ausfallen könnten.[19] So besteht die Transformer-Architektur zwar im Kern in der Fähigkeit, das nächste Token vorherzusagen, doch dieser simple Mechanismus bringt bis heute immer wieder überraschende Ergebnisse hervor.

Der Trend zu mehr Rechenzeit für die Inferenz ist auch für einen anderen zunehmend wichtigen Strang der KI-Forschung entscheidend, der sich »Agenten« widmet. Während die paradigmatische Erfahrung mit generativer KI darin besteht, ChatGPT eine Frage zu stellen und eine Antwort zu erhalten, sind Agenten als wesentlich ausgeklügeltere Systeme in der Lage, relativ eigenständig zu handeln. Anstatt lediglich eine Antwort zu geben, erstellen sie ausgehend von einem benutzerdefinierten Ziel einen Plan und führen eine Reihe von Aufgaben durch.[20] Sie sind nicht wie ChatGPT darauf ausgerichtet, der Nutzerin so schnell wie möglich eine Antwort vorzulegen, sondern nehmen sich die notwendige Zeit, um dieses Ziel zu erreichen. Während der anfängliche Fokus bei generativer KI auf der Phase des Vortrainings lag, die

selbstüberwachtes Lernen einschließt, gewinnt mit den Agenten die Forschung zu verstärkendem Lernen und der Frage, wie innerhalb bestimmter Umgebungen geeignete Belohnungen für Agenten konzipiert werden können an Bedeutung. Entscheidend ist, dass Agenten potenziell weitaus mehr Nutzen für Unternehmen bieten als ein schlichter Chatbot. Sie sind nicht bloß ein Orakel, dem man Fragen stellen kann, sondern tatsächlich in der Lage, Schritte durchzuführen, um ein Ziel zu erreichen. Mit ihrer Hilfe könnten KI-Systeme zunehmend zu Vermittlern zwischen den Nutzern und dem übrigen digitalen Ökosystem werden. Vielleicht müssen wir eines Tages nur unserem personalisierten Agenten unsere Wünsche mitteilen, und schon macht er sich eigenständig ans Werk, um sie uns zu erfüllen (ein typisches Beispiel dafür ist schon heute die Urlaubsplanung, doch verglichen mit dem, was möglich werden könnte, ist dies eine triviale Aufgabe). Es gibt bereits Bemühungen, das digitale Ökosystem für Agenten umzugestalten, etwa indem man ihnen Zugriff auf Apps und Werkzeuge gibt. Das wirtschaftliche Potenzial einer solchen Entwicklung ist beträchtlich, was ein Grund dafür ist, dass sich alle führenden KI-Forschungsgruppen auf die Entwicklung von Agenten konzentrieren. Manche Firmen arbeiten auch an Systemen mit mehreren Agenten, um eine kollektive Dynamik für die Lösung von Problemen und die Umsetzung von Projekten zu erzeugen.[21]

Während im öffentlichen Diskurs einzelne Modelle im Mittelpunkt stehen, gewinnen integrierte Systeme immer mehr an Bedeutung, »die KI-Aufgaben mithilfe mehrerer interagierender Komponenten erfüllen, etwa indem verschiedene Modelle, Retriever und externe Werkzeuge einbezogen werden«.[22] Agenten sind ein gutes Beispiel dafür: Ein LLM mag hier zwar die Schnittstelle zwischen Nutzer und System sein, doch danach kann der Agent auf vielfältige Instrumente wie Suchmaschinen, Debugger, Rechner usw. zurückgreifen. Ebenso setzen viele Unternehmen aufgrund der hohen Betriebskosten von sehr großen Modellen auf Systeme, die verschiedene Modelle kombinieren und Anfragen automatisch an das geeignetste weiterleiten.[23] Einfache Anfragen landen bei den kleineren und günstigeren Modellen, komplexere bei den größeren und teureren. Microsoft zum Beispiel hat dies mit Bing umgesetzt.[24] Wie bemerkt kommt es hier nicht auf ein einzelnes Modell an, sondern auf das im Hintergrund operierende Gesamtsystem. Ein beträchtlicher

Teil der heutigen KI-Forschung zielt lediglich darauf, durch die Entwicklung solcher Systeme das Leistungsvermögen bereits bestehender Modelle auszuschöpfen und zu steigern. Wie Google Research in einem Artikel resümiert: »Modelle mit starken Werkzeugen auszustatten, erhöht ihre Leistungskraft.«[25] Die Medien berichten vor allem über die Steigerung der schieren Rechenleistung, also die Verbesserung von KI durch immer größere Modelle, doch das Konzept des Systems zeigt, dass auch das Gerüst, das um ein gegebenes Modell herum gebaut wird, erhebliche Verbesserungen bewirken kann.

Während die Öffentlichkeit gebannt auf ChatGPT schaut, wird das vorliegende Buch daher eine breitere Perspektive auf KI einnehmen. Um nur mit einigen Stichworten anzudeuten, was das heißt: Im Zentrum stehen Unternehmen, nicht Verbraucher; Token, nicht Text; Multimodalität, nicht Sprache; Nachdenken, nicht Nachplappern; Agenten, nicht Antworten; und Systeme, nicht Modelle. Im Blick behalten sollten wir auch die gegenwärtige Forschung an verschiedenen Architekturen, die einige Beschränkungen des Transformers überwinden sollen – unter anderem Zustandsraummodelle, Flüssige Neuronale Netzwerke sowie xLSTM- und Titan-Modelle.[26] Besonders vielversprechend scheinen dabei Diffusions- und Weltmodelle zu sein. Das alles bedeutet, dass Chatbots schwerlich wegweisend sind für die derzeitige KI-Entwicklung und sich daher Kritiker wie Verfechterinnen der Technologie vergewissern sollten, ob sie ihr Thema nicht verfehlen.

Die vier Ebenen der generativen KI

Wir haben nun eine Vorstellung davon, was KI heute bedeutet, doch wie stellt sich die Landschaft aus wirtschaftlicher Perspektive betrachtet dar? Schon ein rascher Blick in die Finanznachrichten führt uns eine verwirrend komplexe Fülle von Start-ups, Marktführern in rapidem Wandel, technischen Innovationen und beinahe täglich neuen Produkten vor Augen. Wie können wir diese vielfältigen Elemente vor dem Hintergrund schneller und tiefgreifender technologischer Veränderungen begreifen? Wenn wir uns die generative KI wie einen Stapel vorstellen, lassen sich darin vier Ebenen ausmachen. Von unten nach oben sind dies Hardware, Infrastruktur, Modelle und Anwendungen.

Manche Unternehmen beschränken sich auf eine davon, andere sind stärker vertikal integriert.

Hardware

Die erste der vier Ebenen besteht in der Hardware – besonders in sogenannten Beschleunigern, die beim Training und Betrieb von KI-Modellen zum Einsatz kommen. Im Unterschied zu Central Processing Units (CPUs), die alle möglichen Rechenoperationen durchführen sollen, sind diese Chips darauf spezialisiert, einige ausgewählte Funktionen besonders effizient zu erfüllen. Beispiele dafür reichen von Graphics Processing Units (GPUs) bis zu Tensor Processing Units (TPUs) und einer ganzen Reihe anderer Architekturen. Sie alle sind so konzipiert, dass sie die für die heutige KI nötigen Berechnungen – vor allem Matrizenmultiplikationen – möglichst schnell und effektiv leisten können. Auf dieser Ebene der Hardware dominiert zurzeit Nvidia mit einem Marktanteil von 92 Prozent.[27] Für mehr als 90 Prozent der KI-Forschungspapiere wurden zuletzt Nvidia-Chips verwendet.[28] Entsprechend stellen sich die Geldströme dar – Umsatz und Gewinn des Unternehmens sind seit der Einführung von ChatGPT steil gestiegen. Nvidia ist bislang zweifellos der größte Gewinner des Booms von generativer KI. Seine Marktkapitalisierung ist explosionsartig gewachsen, zeitweilig war es sogar das wertvollste Unternehmen der Welt. Diese marktbeherrschende Stellung ist den Kartellämtern nicht entgangen; sie haben inzwischen mehrere Ermittlungsverfahren wegen wettbewerbswidrigen Verhaltens gegen das Unternehmen eröffnet.[29]

Allerdings gibt es in dieser Schicht eine wachsende Konkurrenz, da etliche Start-ups alles daransetzen, einen Teil des expandierenden Marktes zu erobern. Um einen Wettbewerbsvorteil zu erzielen, setzen sie in der Regel auf die Entwicklung effizienterer Chips, die stärker auf die Transformer-Architektur spezialisiert sind als herkömmliche GPUs (deren Ursprung in den Berechnungen für Polygone in Videospielen liegt). Andere drängen vor allem auf den Markt für Inferenz-Chips, auf dem die Dominanz von Nvidia etwas weniger ausgeprägt ist.[30] Auch etablierte Firmen wie AMD versuchen Marktanteile mit Chips zurückzugewinnen, die teilweise so leistungsstark sind wie die von Nvidia, aber ohne deren üppige Preisaufschläge angeboten werden. Vielleicht am wichtigsten für das Marktgeschehen ist allerdings, dass alle großen

Cloud-Unternehmen – sogenannte Hyperscaler wie Amazon, Microsoft und Google – gerade ihre eigene Hardware entwickeln. Im Augenblick sind sie noch stark abhängig von Nvidia und folglich dessen Gutdünken und Preisen für die leistungsfähigsten Chips ausgeliefert – eine Situation, in der sich die Hyperscaler selten selbst befunden haben und der sie unbedingt entfliehen wollen. Deshalb finden zurzeit beträchtliche Investitionen in die Entwicklung von Chips für Training wie Inferenz statt. Einen großen Vorsprung dabei hat Google, da es bereits seit fast einem Jahrzehnt an seinen eigenen Chips – TPUs – arbeitet. Tatsächlich erzielen TPUs auf dem Markt für Trainingschips bereits einen erheblichen Anteil, der aus dem schlichten Grund unterschätzt wird, dass Google keine Angaben darüber macht, wie viele TPUs es einsetzt.[31]

Zusätzlich zu diesen Entwicklungen findet ein konzertierter Angriff auf eine tragende Säule der Vorherrschaft von Nvidia statt: die Plattform Compute Unified Device Architecture (CUDA). CUDA ist eine proprietäre Software, mit deren Hilfe GPUs und ihre parallelisierten Abläufe vielfältige Aufgaben jenseits der Grafikverarbeitung erfüllen können, für die sie ursprünglich konzipiert wurden. Im Lauf der Zeit wurde das um CUDA entstandene Ökosystem allgegenwärtig. In Reaktion auf diese Zentralität beteiligen sich Big-Tech-Unternehmen wie Microsoft, Google und Meta an der Entwicklung der Open-Source-Alternative Triton.[32] Gleichzeitig haben sie sich mit weiteren Firmen zur Ultra Accelerator Link Promoter Group zusammengeschlossen, um einen neuen Industriestandard für die wichtigen Komponenten, die KI-Chips miteinander verbinden, zu entwickeln – auch dies im Bemühen, ein proprietäres Produkt von Nvidia zurückzudrängen.[33]

Infrastruktur

Über der Hardware-Ebene liegt die der Infrastruktur: In immer größeren Gebäuden, die riesige Mengen an Energie und Wasser verbrauchen, werden die mit Chips bestückten Server-Racks miteinander vernetzt. Um heute eine KI zu trainieren, genügt eine einzelne GPU (oder auch das Achter-Set, das Nvidia in der Regel anbietet) bei Weitem nicht – Zehn- oder sogar Hunderttausende GPUs müssen zusammengeschlossen werden. Ein Training mit derart vielen Recheneinheiten durchzuführen, ist ungemein schwierig und erfordert hochkomplexe Verbindungssysteme.[34] Angesichts der Herausforderung, ein solches

Aggregat herzustellen, greifen die meisten Firmen auf spezialisierte Anbieter zurück und beschaffen sich die benötigten Rechenkapazitäten über Cloud-Dienste. Auf dem Markt für Cloud Computing (der selbstverständlich mehr umfasst als KI-Angebote) herrscht auf der Infrastruktur-Ebene mehr Wettbewerb als auf der der Hardware – allerdings nicht viel mehr. Statt eines einzigen Unternehmens sind es hier drei, die dominieren: Amazon Web Services (AWS), Microsoft Azure und Google Cloud kommen zusammen auf 64 Prozent des globalen Cloud-Markts.[35] Sie haben den Sektor im Lauf der letzten Dekade monopolisiert – und nahezu unüberwindliche Hürden errichtet, um potenzielle Herausforderer abzuschrecken.

Wie auf der Hardware-Ebene aber haben das schiere Tempo der Investitionen und der KI-Hype auch hier dazu geführt, dass einige Konkurrenten entstanden sind, die sich einen Teil der Milliardenbeträge sichern wollen. Insbesondere gibt es eine Reihe von Firmen, die sich fast ausschließlich darauf konzentrieren, GPUs für die KI-Inferenz bereitzustellen. Das tun zwar auch die Hyperscaler, aber ihre Infrastruktur ist etwa mit Blick auf Kühlung und Stromversorgung nicht im selben Maß für die enormen Leistungsanforderungen der KI optimiert. Die »Neo-Clouds« der Herausforderer, die teilweise bis vor Kurzem für das Bitcoin-Mining verwendet wurden, scheinen im Moment aus zwei Gründen tragfähig zu sein. Zum einen ist die Nachfrage nach Rechenleistung gegenwärtig schlicht größer als das Angebot. Alle großen KI-Unternehmen wollen mehr Kapazitäten, aber das gibt der Markt nicht her. Selbst Big-Tech-Unternehmen wie Microsoft mussten ihre Rechenkapazitäten unter den Mitarbeitern rationieren und mit Konkurrenten zusammenarbeiten, um eine ausreichende Versorgung sicherzustellen.[36] Der schiere Umfang der Nachfrage nach GPUs bedeutet, dass nicht einmal die Hyperscaler den Markt abdecken können.

Der zweite Grund dafür, dass sich die Neo-Clouds zurzeit behaupten können, ist ihr Bemühen, die großen Cloud-Anbieter mithilfe niedrigerer Fixkosten preislich zu unterbieten oder sich durch die Optimierung von Größen wie Latenz und Datendurchsatz von ihnen abzuheben.[37] Indem sie sich auf KI-Anforderungen spezialisieren, hoffen sie ihre Kosten zu senken und bessere Dienstleistungen anzubieten als die breiter aufgestellten traditionellen Cloud-Unternehmen. Ob dies mittelfristig tragfähig bleibt, ist keineswegs gewiss – die Halbleiter-

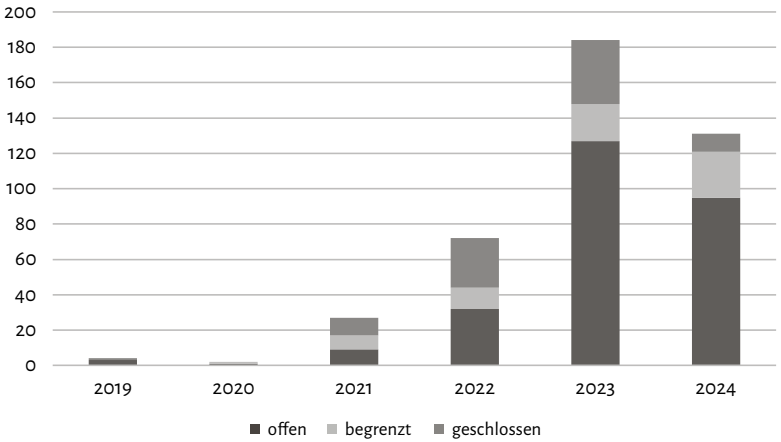
branche ist bekannt dafür, dass sie zyklische Schwankungen durchläuft. Vorerst aber gibt es einen Markt für diese kleineren Infrastrukturanbieter.

Modelle

Über der Ebene der Infrastruktur kommen wir zu der der Modelle, die für die jüngsten KI-Entwicklungen am prägendsten gewesen ist. Diese Modelle können eine Vielfalt von Formen annehmen – sie unterscheiden sich unter anderem in der Architektur, der Größe und den Trainingsdaten –, doch in den vergangenen Jahren haben vor allem solche Verbreitung gefunden, die auf der Transformer-Architektur beruhen. Sie laden offenbar dazu ein, ihre Komplexität zu überzeichnen und zu mystifizieren, doch in der simpelsten Variante bestehen sie oft nur aus zwei Dateien: eine ausführbare Code-Datei und eine, die die Gewichte der Parameter des neuronalen Netzwerks enthält.[38] Diese Parameterdatei allerdings ist mitunter das Ergebnis von Milliardeninvestitionen und immensen Rechenoperationen. Riesige Datenmengen – beinahe das gesamte Internet – werden in ihr auf ein relativ kleines Format komprimiert. In den prägnanten Worten von Ted Chiang ist generative KI wie ein unscharfes JPEG des Internet: Sie versucht, das Web nachzubilden, doch die Details verschwimmen dabei, und die Genauigkeit bleibt manchmal auf der Strecke.[39]

Auf der Ebene der Modelle finden sich die Unternehmen, die für die KI-Branche emblematisch sind: bekannte Namen wie OpenAI, Google, Meta und Anthropic. Zu dieser kleinen Gruppe von Marktführern gesellt sich allerdings eine überraschende Vielfalt an anderen Optionen. Im Jahr 2023 zum Beispiel veröffentlichten 96 Firmen 182 neue Basismodelle (siehe Abb. 1). Die meisten davon waren Open-Source-Modelle, wobei in der Regel die Parametergewichte offengelegt wurden. Während die Hardware-Ebene von einem einzigen Unternehmen und die der Infrastruktur von dreien dominiert wird, konnten auf der Ebene der Modelle tausend Blumen blühen – die Unternehmen nutzen sie zu Werbezwecken und um Kunden in ihre Ökosysteme zu lotsen.[40] Zumindest bis jetzt besteht unter den Modellentwicklern ein starker Wettbewerb: Ein »Krieg der hundert Modelle« hat begonnen, wie der Vizepräsident von Tencent formulierte.[41]

Abb. 1 **Basismodelle nach Zugangstyp (2019 – 2024)**



Quelle https://crfm.stanford.edu/ecosystem-graphs/index.html

Anwendungen

Über den Modellen erhebt sich die letzte Ebene: die der Anwendungen, auf der aus den Modellen Produkte hergestellt werden. ChatGPT ist das bekannteste, doch diese Ebene birgt innerhalb der heutigen KI am meisten Ungewissheit und ist am stärksten im Fluss – es werden viele verschiedene Ansätze verfolgt, und ein klares Vorbild, dem andere folgen könnten, gibt es nicht. Manche Unternehmen bieten einfach eine neue Schnittstelle für ein bereits existierendes Modell an, die den Nutzerinnen leichteren Zugang zu bestimmten Funktionen gibt und sie so von dem riesigen offenen Raum weglockt, der sich hinter dem leeren Eingabefeld eines Chatbots verbirgt. Andere setzen auf komplexere Systeme, die Nutzeranfragen je nach den Erfordernissen an verschiedene Modelle weiterleiten. Manche passen existierende Modelle an bestimmte Aufgaben an, andere konstruieren eigene Modelle, die ihre Anwendungen unterstützen. Adobe Firefly zum Beispiel bietet eine Reihe von Werkzeugen zur Bilderstellung und -bearbeitung, die auf einem Modell aufbauen, das (angeblich) mit frei zugänglichen Daten trainiert wurde, um urheberrechtliche Probleme zu umschiffen. Wie sich Künstliche Intelligenz weiterentwickeln wird, scheint allerdings nirgends so ungewiss wie auf der Ebene der Anwendungen.

Spannungen und Entwicklungen

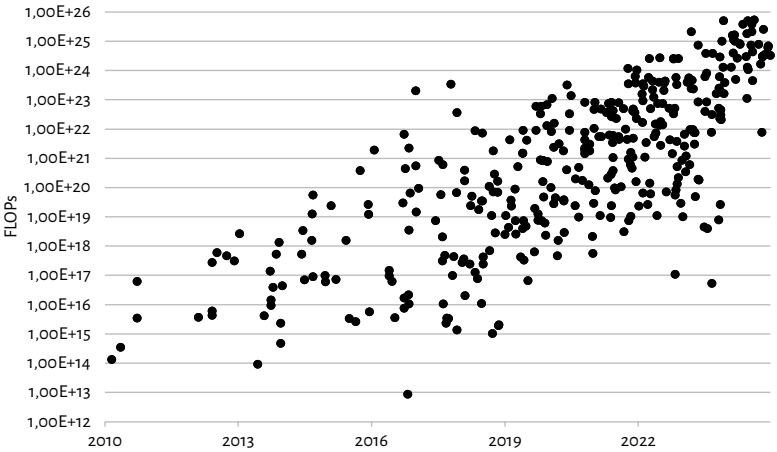
Die gegenwärtige KI-Landschaft sieht somit folgendermaßen aus: Während auf der Ebene der Anwendungen und der Modelle eine hektische Konkurrenz herrscht, besteht bei der Infrastruktur ein Oligopol, das allerdings zunehmend herausgefordert wird, und bei der Hardware ein solides Monopol, das sich jedoch ebenfalls neuen Wettbewerbern gegenübersieht. Das ist allerdings eine sehr statische Momentaufnahme des Ökosystems. Wenn wir verstehen wollen, wie sich die Landschaft in den kommenden Monaten und Jahren verändern wird, müssen wir zwei Elemente berücksichtigen, die die Dynamik von KI maßgeblich bestimmen: Profite und Macht.

Wo bleiben die Gewinne?

Der wichtigste Aspekt besteht darin, dass es für nahezu alle am KI-Boom beteiligten Unternehmen schwierig bleibt, Profite zu erzielen. Hauptgrund dafür sind die enormen und stetig wachsenden Kosten. Die Konkurrenz zwingt die Firmen dazu, immer bessere Modelle zu entwickeln – doch dafür müssen immer größere Mengen an Kapital investiert werden. Was gerade bei generativer KI das Deep Learning fördert und daher unverzichtbar ist, sind wachsende Rechenkapazitäten.[42] In einer viel beachteten Reflexion über die lange Geschichte der KI-Entwicklung heißt es, Forscher sollten die »bittere Lektion« akzeptieren, dass der menschliche Erfindungsgeist für sie zwar hilfreich sei, »Methoden zur Steigerung der Rechenleistung aber letztlich bei Weitem am effektivsten sind«.[43] Was schlussendlich zählt, ist der Einsatz von mehr Rechenkapazität und deren optimale Ausnutzung (siehe Abb. 2). Der Drang danach aber treibt die Trainingskosten von KI-Modellen deutlich in die Höhe, obwohl die Hardware billiger und effizienter wird: Die Nachfrage nach mehr Rechenleistung wächst schneller, als die Kosten der Hardware sinken.

Deep Learning war schon immer auf starke Computer angewiesen, doch die Ära der generativen KI hat diesen Bedarf auf eine neue Stufe gehoben. Da die Transformer-Architektur parallele Operationen stärker zulässt als frühere Ansätze, hat die erwähnte »bittere Lektion« noch an Bedeutung gewonnen. Das Ergebnis besteht darin, dass die enormen Kosten für das Training der fortgeschrittensten Modelle stetig weiter

Abb. 2 Rechenkapazität für das Training wichtiger Modelle (2010 – 2024)



Quelle https://epochai.org/data/notable-ai-models

wachsen. Während Anthropic für seine neuesten Modelle nur mehrere zehn Millionen Dollar ausgegeben haben soll, waren es Gerüchten zufolge bei GPT-4 rund 100 Millionen und bei Googles Gemini 1.0 fast 200 Millionen Dollar.[44] Diese Trainingskosten sind obendrein nur die Spitze des Eisbergs – hinzu kommen die Ausgaben für die Hardware, die Löhne der Beschäftigten, die Kosten gescheiterter Trainingsläufe und vieles andere mehr. Im Jahr 2025 wurden neue Ansätze für Training und Betrieb von KI-Modellen verfolgt, die den Bedarf an Rechenkapazitäten weiter in die Höhe treiben.

Der erste besteht im bestärkenden Lernen, das den Modellen nach dem eigentlichen Training die Fähigkeit zum »Abwägen« oder »Nachdenken« verleihen soll. Genaue Zahlen über Umfang und Kosten der dafür erforderlichen Rechenleistung liegen aktuell nicht vor, aber es wird allgemein angenommen, dass sie immens sind und »sich annähernd in der Größenordnung des Trainings bewegen«.[45] Danach kommt die Stufe der Inferenz, also des Betriebs der Modelle. Das Training ist zwar aufwendig und teuer, umfasst aber viel weniger Durchläufe als die spätere Nutzung. Anfang 2023 gab OpenAI Schätzungen zufolge tagtäglich rund 700.000 Dollar für die Rechenleistung aus, die der Betrieb seiner Modelle erfordert – was aufs Jahr gerechnet gut 255 Millionen Dollar entspricht.[46] Sein Konkurrent Anthropic soll etwa die Hälfte