# Die wahren Kosten der Künstlichen Intelligenz

KI hätte nie die Dampfmaschine erfunden. Und das ist nur ein Beispiel für zwei Risiken, die derzeit unterschätzt werden.

*Von Christian R. Ulbrich und Urs Gasser*

An kritischen Stimmen zum Thema KI herrscht derzeit wahrlich kein Mangel. Nimmt man die Flut von Beiträgen einmal genauer unter die Lupe, lassen sich zwei grundsätzliche Arten von Kritikern und Mahnerinnen unterscheiden: Die eine Fraktion nimmt die kurzfristigen Folgen ins Visier, die andere die langfristigen. Übersehen werden die mittelfristigen Folgen. Dabei werden sie die gravierendsten Probleme und die größten Kosten verursachen.

Die übergeordnete Aufgabe von KI – Fachleute sprechen häufig lieber von maschinellem Lernen – ist es, menschliche Tätigkeiten, Entscheidungen oder Bewertungen durch algorithmische Entscheidungen oder Urteile zu ersetzen. Dafür analysiert eine KI meist Trainingsdaten auf raffinierte Weise, um in ihnen Muster zu erkennen und das in den Daten implizite Wissen aufzudecken. Oft wendet die KI dieses Wissen direkt an, indem sie konkrete Empfehlungen gibt, denen die Menschen fast immer folgen. Dieses Phänomen wird von Psychologen als „Automation Bias" bezeichnet. Oder indem sie die menschliche Tätigkeit gleich ganz selbst ausführt. Menschliche Aufgaben werden also von einer KI digital automatisiert. Infolge dieses Transfers vom Menschen zur Maschine kann es zu verschiedenen unerwünschten Konsequenzen kommen.

Die möglichen kurzfristigen Folgen sind die offensichtlichsten: Aufgrund von statistischen Verzerrungen oder ungenügenden Modellen können etwa einzelne Personen oder Personengruppen diskriminiert werden. Fehlende Transparenz und mangelnde Nachvollziehbarkeit von KI-Entscheidungen können Vertrauen untergraben und zu unvorhergesehenen Fehlern führen. Im Zusammenhang mit den Trainingsdaten müssen oft Urheberrechtsfragen und fundamentale Verteilungsfragen geklärt werden. Gleiches gilt für ethische Dilemma oder etwa den Einsatz von KI im Krieg. Auf diese leicht zugänglichen und ganz konkreten Fragestellungen haben sich Politiker, Sozialwissenschaftler, Medien und NGOs primär gestürzt. Das ist auch wichtig und gut so.

Die ebenfalls viel diskutierten langfristigen Folgen sind eher abstrakter Natur. So werden etwa (auf falschen Annahmen basierende) Ängste vor Massenarbeitslosigkeit geschürt. Oder gleich die Überflüssigkeit von menschlichen Entscheidungen prognostiziert und eine Machtübernahme durch KI-Systeme unter dem Stichwort Singularität wahlweise als Teufel an die Wand gemalt oder als erstrebenswertes Ziel ausgegeben. Diese Art von Diskussionen ist die bevorzugte Domäne der Informatiker und der Technologieelite aus dem Silicon Valley.

Eine dritte Kategorie von Folgen des unüberlegten massenhaften Einsatzes von KI kommt indes im Diskurs zu kurz, weil sie weniger konkret und daher schwerer zu greifen sind. Doch die Chancen stehen gut, dass sie mittelfristig für die spürbarsten Konsequenzen für unsere gesellschaftliche Entwicklung und unsere Position im globalen Wettbewerb sorgen wird. Hier sind insbesondere zwei Risiken zu nennen.

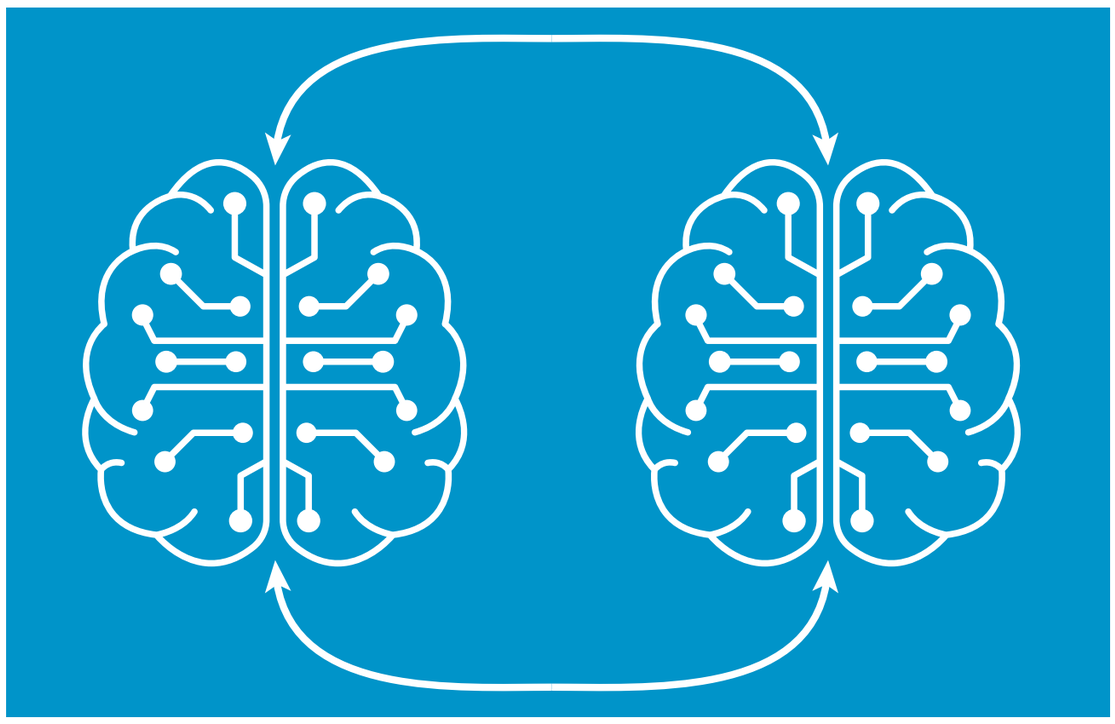

Illustration F.A.Z.

## Zunehmende Zentralisierung

Erstens kommt es durch den Transfer von Entscheidungen und Tätigkeiten auf Maschinen zwangsläufig zu einer weiteren Zentralisierung von Macht und Einfluss, die sich immer stärker auf eine Instanz konzentrieren. Entscheidungen, die bisher von vielen verschiedenen Mitarbeitenden von Tag zu Tag dezentral getroffen wurden, werden nun durch ein KI-System gebündelt. Diese Verschiebung ist fast immer auch dann bedeutsam, wenn die Mitarbeitenden bisher weisungsabhängig waren oder anhand Standards und Checklisten tätig geworden sind. Denn einen individuellen, von der Führung kaum kontrollierbaren Entscheidungsraum gibt es fast immer.

Ein KI-System hingegen wird nur von einigen wenigen kontrolliert und gesteuert, etwa durch die Auswahl der Trainingsdaten oder des Designs und der Art der Implementierung des jeweiligen Algorithmus. Das gilt im Übrigen auch für künstliche neuronale Netzwerke. Selbst bei ihnen kann „top down" jederzeit in das Modell, etwa über das Anpassen von Gewichtungen, steuernd eingegriffen werden. Das verschafft der obersten Instanz oder der zuständigen IT-Abteilung wesentlich mehr Einfluss. Mit der Zeit kann es zu einer unerwünschten Ballung von Macht kommen – wie die großen Internetunternehmen derzeit schon sehr plastisch vor Augen führen. Digitale KI-Systeme sind daher von Grund auf viel anfälliger für Machtmissbrauch jeglicher Art. Aber auch kleinere Manipulationen und die gezielte Beeinflussung der Nutzenden lassen sich deutlich leichter umsetzen.

## KI ist konservativ

Zweitens ist nur wenigen klar, dass eine KI im Herzen konservativ ist. Das ist bemerkenswert, weil es so kontraintuitiv ist – schließlich gelten Künstliche Intelligenzen derzeit als der Inbegriff von Fortschritt. Die Möglichkeit von automatisierten Anpassungen über neue Trainingsdaten, also etwa ein eingebauter Mechanismus zur Aktualisierung, stellt in der Tat eine merkliche Verbesserung gegenüber den starren, expliziten Regeln klassischer Algorithmen dar. Allerdings wird im Zuge der Dateneuphorie oft unterschätzt, dass Daten ausschließlich Informationen und Vorgänge aus der Vergangenheit beinhalten. Datenbasierte Ansätze erlauben es zwar, flexibel auf ungewohnte Situationen zu reagieren, allerdings tun sie das stets mit Mustern, die sie aus der Vergangenheit gelernt haben.

Natürlich lässt sich aus vergangenen Daten sehr viel lernen – jedoch nur, wenn die Welt sich nicht verändert, wenn also sowohl die Umstände, welche die Grundlage unserer Entscheidungen und Tätigkeiten bilden, als auch die Ziele gleich bleiben, die wir erreichen wollen. Eine primär datenbasierte digitale Transformation gibt dem Status quo sehr viel mehr Gewicht und könnte auf Dauer die Weiterentwicklung der Gesellschaft und die Innovationskraft der Wirtschaft hemmen. Bewusste Abweichungen von der Praxis der Vergangenheit, insbesondere solche Abweichungen, die vor allem auf menschlichen Werturteilen, menschlicher Vorstellungskraft oder auch auf dem Zufall basieren, werden im Zuge der digitalen Automatisierung erschwert.

KI-Systeme können nach aktuellem Stand der Technologie überzeugend den Eindruck vermitteln, etwas Neues zu schaffen, indem sie vorhandenes Wissen ungewohnt kombinieren. Eine Neukombination vorhandenen Wissens kann kurzfristig sehr nützlich und effizienzsteigernd sein. Aber sie ist eben nicht wirklich neu. Mittelfristig besteht das Hauptproblem darin, dass datenbasierte KI-Systeme nicht wirklich experimentieren. Sie produzieren Ergebnisse auf Grundlage früherer Entscheidungen und optimieren diese immer weiter. Suboptimale Entscheidungen werden direkt verworfen, auch wenn sie sich später – etwa im Falle geänderter Umstände oder Ziele – als deutlich besser herausstellen könnten. KI-Systeme werden immer den effizienteren Weg wählen, anstatt zu versuchen, einen neuen zu finden. Plakativ gesprochen, hätte eine KI nie eine Dampfmaschine erfunden, sondern die Nutzung der Kraft von Pferden, Rindern und Eseln immer weiter optimiert. Möchte man experimentieren und völlig neue Wege einschlagen, hat man gerade keine historischen Daten, auf die man seine Entscheidung basieren oder mit denen man sie rechtfertigen könnte.

Zudem verstärkt sich dieses Problem mit der Zeit. Denn je länger Entscheidungen von KI-Systemen getroffen werden, desto größer wird der Anteil an maschinellen Entscheidungen in den Trainingsdaten. Die Trainingsdaten werden mit der Zeit also immer homogener. Dadurch sinken die Verhaltensvielfalt und Variabilität in den Trainingsdaten konstant und verstärken fortwährend den konservativen Charakter des Outputs. Je mehr wir menschliche Entscheidungen mit Künstlichen Intelligenzen automatisieren, desto stärker schränken wir die Fähigkeit unserer sozialen Systeme zu echtem Fortschritt und fundamentalen Anpassungen in Form von völlig Neuem ein. Das wäre fatal. Denn gerade Wissen und Innovation sind die verbliebenen Vorteile des an Ressourcen und Energieträgern armen Europas im globalen Wettbewerb.

## Auswege aus dem Dilemma

Natürlich ist diese Entwicklung nicht zwangsläufig. Aus den Risiken zu schlussfolgern, auf KI ganz zu verzichten, ist offensichtlich der falsche Weg. Vielmehr gilt es, sich der Probleme bewusst zu sein und sie gezielt zu minimieren. Die unerwünschten Folgen einer zu starken Zentralisierung lassen sich etwa vermeiden, indem man bewusst auf Skalierung verzichtet und anstelle großer zentraler Systeme viele kleinere Systeme ausrollt. Damit das praktikabel wird, ist es aber unumgänglich, Standards, Protokolle und Schnittstellen zu etablieren, die den reibungslosen Austausch zwischen den verschiedenen Systemen ermöglichen. Darüber hinaus können begrenzte menschliche Hoheitsbereiche in strategisch wichtigen Gebieten definiert werden, die auch in Zukunft frei von digitaler Automatisierung bleiben. Auch die Variabilität in den Daten lässt sich erhalten. Das bedeutet aber, dass die KI-Systeme auch suboptimale Entscheidungen treffen müssen – wie es etwa die „Temperature" in ChatGPT vormacht. Das könnte über explizite Zufallskomponenten gelingen. Auch müssen die KI-Modelle lernen, zu weit zurückliegendes implizites Wissen wieder zu vergessen, was eine ganz besondere Herausforderung werden wird. Die dadurch entstehenden zusätzlichen Kosten, Ineffizienzen und den erhöhten Zeitaufwand gilt es bewusst in Kauf zu nehmen.

**Christian R. Ulbrich** ist Leiter der Forschungsstelle für Digitalisierung in Staat und Verwaltung (e-PIAF) an der Universität Basel und zusammen mit Bruno S. Frey Autor des Buches „Automated Democracy – Die Neuverteilung von Macht und Einfluss im digitalen Staat".

**Urs Gasser** ist Gründungsdekan der School of Social Sciences and Technology der TU München. Zuvor war er Leiter des Berkman Klein Centers for Internet & Society an der Harvard University. Zusammen mit Viktor Mayer-Schönberger ist er Autor des Buches „Guardrails – Guiding Human Decisions in the Age of AI".



# Wie Open AI das nächste ChatGPT auf ein neues Level heben kann

*Von Martin Wendiggensen*

Die Veröffentlichung von GPT-3 im November 2022 hat viel verändert. Generative Künstliche Intelligenz entwickelte sich von einem akademischen Feld zu einem eigenen Wirtschaftszweig – und brachte für Open AI den Durchbruch. Kein Wunder also, dass schon während der Ankündigung des Nachfolgers GPT-4 die Werbesprache primär aus Superlativen bestand. Noch größere Versprechen macht Open AI für GPT-5 – so kündigte das Unternehmen an, dass das Sprachmodell besser logisch denken könne, mit Videos interagieren und „die Intelligenz eines Promovierten" haben werde. Wie diese Fortschritte aber gemessen, geschweige denn erreicht werden könnten, dies bleibt offen und wurde bei vorherigen Modellen auch im Nachhinein nicht veröffentlicht.

Die Entwicklung fortgeschrittener KI-Modelle besteht wesentlich aus drei Komponenten – Daten, Algorithmen und Rechenzeit. Daten sind die Wissensbasis, auf der ein Modell durch das Anwenden von Algorithmen trainiert wird. Die Algorithmen sind die Architektur des neuronalen Netzes und optimieren das Modell während des Trainings. Die Menge an Rechenzeit bestimmt, wie viel Computerleistung auf das Training verwendet wird, und damit, wie nuanciert das fertige Modell ist. Die drei Komponenten – Daten, Algorithmen und Rechenzeit – können also zusammen herangezogen werden, um die Leistung eines KI-Modells zu bestimmen.

Potentiell kann Open AI an allen drei Stellschrauben in der KI-Entwicklung drehen. Realistisch sind Veränderungen der drei Komponenten jedoch unterschiedlich aufwendig umzusetzen. Aktuelle Forschung deutet darauf hin, dass die Verbesserung einzelner Komponenten deutlich positive Effekte auf die Leistung der KI-Modelle hat. Wird etwa auf Basis gleicher Daten und Algorithmen mehr Rechenzeit investiert, verbessert sich die Leistung messbar.

Eine Erhöhung der Rechenzeit scheint also absehbar. Diese hat Open AI schon über die Vorgängermodelle konsistent erhöht. Doch Rechenzeit allein wird für eine bahnbrechende Entwicklung wahrscheinlich nicht reichen. Viele Fachleute sind der Meinung, die jetzigen Modelle befänden sich an einem Sättigungspunkt, was die Verbesserung durch Rechenleistung betrifft. Denn die Effizienz der Algorithmen und Art der Daten bestimmen den Wert, den mehr Rechenzeit hat. Idealerweise geht Open AI also alle drei Komponenten an.

Die größten Veränderungen wird Open AI darum wahrscheinlich bei den Algorithmen und der Architektur von GPT-5 vornehmen. Ein häufig genannter Ansatz ist, sogenannte Expertennetze zu verwenden, die auf bestimmte Aufgaben oder Themen innerhalb des Modells spezialisiert sind. GPT-4 soll solche Netze nutzen, ein weiterer Ausbau könnte aber die Zuverlässigkeit und Präzision noch einmal erhöhen. Wie wirkungsvoll diese zunehmende Spezialisierung ist, zeigte kürzlich der Konkurrent Google: Zwei seiner Expertenmodelle nahmen an der Internationalen Mathematikolympiade teil, bei der die besten Schüler verschiedener Nationen um die Wette Mathematikklausuren schreiben. Die Neuheit bestand darin, dass Googles Modelle Zwischenschritte überprüfen können. Die Modelle beginnen damit, die Klausuraufgabe durch das Gemini-Sprachmodell zu verstehen und mit ihm Lösungsvorschläge zu entwickeln. Darin sind generative Sprachmodelle gut, sie neigen aber auch zu sogenannten Halluzinationen – Antworten, die logisch und überzeugend formuliert, aber faktisch falsch sind. Um diese zu identifizieren und zu beheben, übersetzt eines der Modelle die Lösungsvorschläge in eine Programmiersprache, die auf mathematische Beweise spezialisiert ist. Der Vorschlag kann nun Schritt für Schritt durchgerechnet werden und damit verlässlich aufzeigen, ob die vorgeschlagene Beweiskette logisch konsistent ist oder Fehler aufweist. Wird ein Fehler in einem Schritt gefunden, wird das Sprachmodell automatisch um einen neuen Vorschlag für diesen Beweisschritt gebeten, bis die gesamte Kette der Überprüfung standhält.

Diese Zerlegung der Aufgabe in logische Komponenten und die automatisierte Überprüfung einzelner Schritte sind eine vielversprechende Arbeitsweise. So schnitten die Modelle hinsichtlich der Qualität ihrer Lösungen unter den besten 30 Prozent der Schüler ab. Die Antwortzeiten sprengten allerdings teilweise die Begrenzung auf 90 Minuten und hätten die KI eher zum Schlusslicht gemacht. Die Ergebnisse sind trotzdem spannend, da sie nahelegen, dass KI-Modelle auch in Mathematik besser werden. In diesem Bereich hatten sie bislang wegen der hohen Anforderungen an logische Konsistenz und numerisches Verständnis Schwierigkeiten. Sollte Open AI auch den Weg des Wettbewerbers Google gehen, könnte es mit einer Kombination an Expertenmodellen und formalen Überprüfungsregeln einen neuen Standard setzen. Mit dem richtigen Training und Regeln für Arbeitswege stünden GPT-5 damit auch bisher verwehrte Bereiche wie die Buchhaltung oder Steuererklärungen offen.

Auch hinsichtlich der Daten gibt es für GPT-5 viel Raum für Verbesserungen. Open-AI-Chef Sam Altman hat in dieser Hinsicht ein besonders ambitioniertes Vorhaben angekündigt: Er möchte die Analyse und Erstellung von Videos direkt in das Hauptmodell einbinden. Wer dabei nur ein weiteres Werkzeug für Nutzer sieht, verpasst die grundlegenden Veränderungen, die dafür erforderlich sind. Für den Erfolg eines derartigen Systems muss das Modell verschiedene Datenarten verstehen und miteinander verknüpfen. Es muss mit einem Konzept sowohl in der Datenart Bild als auch in Text, Ton oder Video umgehen können. Konkret sollte es also mit dem Wort „Stuhl" sowohl ein konkretes Beispiel als auch die generelle Form und Funktion aller Stühle verbinden können.

Was für Menschen selbstverständlich ist, ist für Maschinen eine komplexe Herausforderung. Dass die aktuelle GPT-Version (statische) Bilder erstellen kann, ist beeindruckend. Aber jede weitere Form an Input- und Output-Medien bildet eine völlig neue Herausforderung. Zwischen einem Bild von einem Stuhl und einem Video von einem umkippenden Stuhl liegen für KI-Modelle Welten. Denn Videos sind eine neue Datenart für die Systeme. Wenn es gelingt, Modellen ein solches Verständnis für Objekte im Raum zu vermitteln, könnten sie auch logisches Verständnis erlangen, vermutet Geoffrey Hinton, einer der bekanntesten KI-Forscher der Welt. Das würde tatsächlich einen Evolutionssprung der KI-Modelle darstellen.

Die Hoffnung von Forschern wie Hinton liegt darin, dass ein logisches Verständnis sich in weniger Halluzinationen und einem verlässlicheren Umgang mit neuen Konzepten niederschlägt. Denkbar wäre, dass GPT-5 hier zumindest einige Fortschritte macht. Videos und bessere Bilder generieren zu können wäre für Kreative interessant. Der verlässliche Umgang mit neuen Konzepten hingegen könnte GPT-5 in sich schnell entwickelnden Feldern wie Forschung oder Programmieren nützlicher machen.

Das Potential der Entwicklungen in generativer KI scheint also noch lange nicht ausgeschöpft. Und obwohl Open AI wohl nur spärliche Informationen über Daten, Algorithmen und Rechenzeit von GPT-5 veröffentlichen wird, werden sich diese sofort in der Qualität des Modells zeigen.

**Martin Wendiggensen** promoviert an der School of Advanced International Studies der Johns Hopkins University in Washington, D.C. Sein Forschungsthema sind die Schnittpunkte zwischen Internationalen Beziehungen und Künstlicher Intelligenz.