Heft #4 / Oktober 2023

Error 404

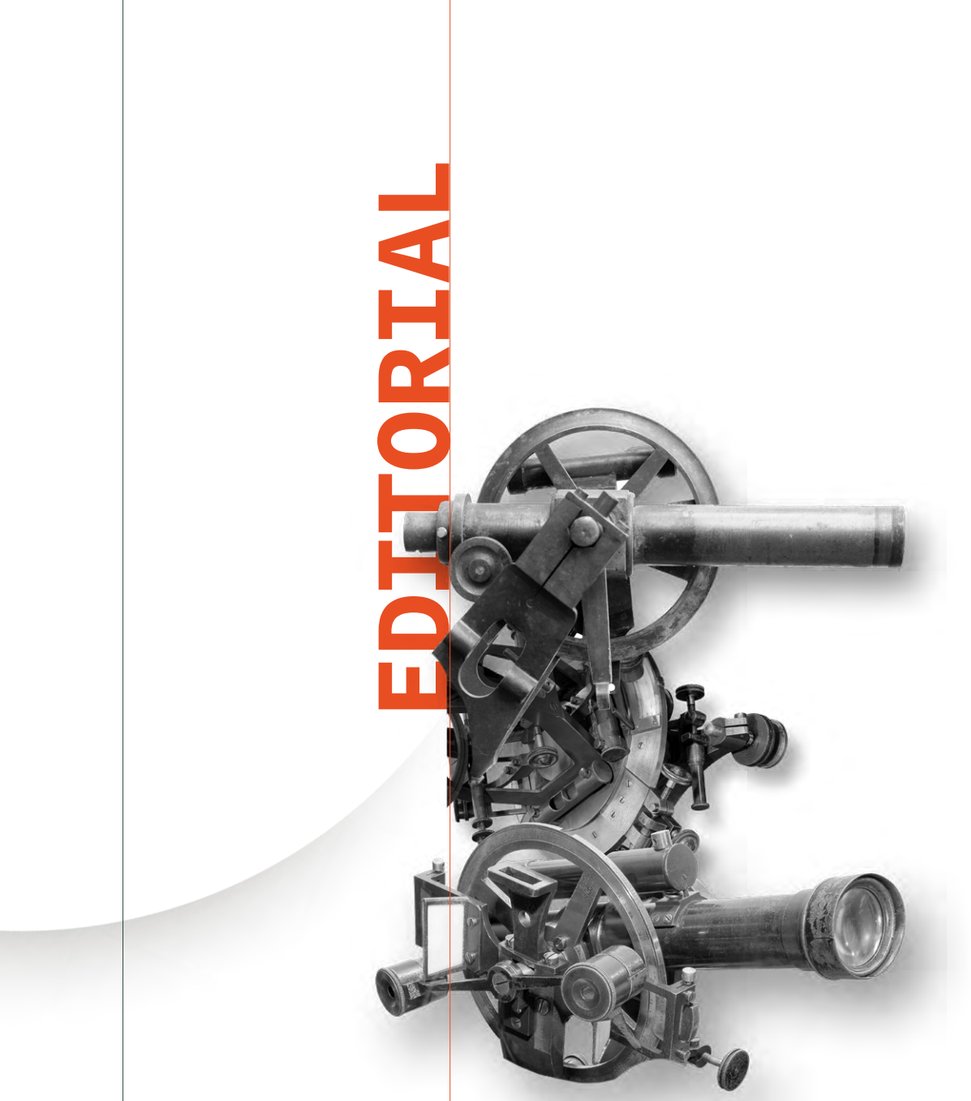
zv ki

Missing Link

Foundation Models – KI zwischen Wachstum und Nachhaltigkeit

Magazin für vertrauenswürdige Künstliche Intelligenz

EDITORIAL



REWRITE THE HYPE

Wir retten mit Künstlicher Intelligenz nicht die Welt. Wir müssen die Welt aber auch nicht vor ihr retten. Angesichts immer komplexerer KI-Modelle sollten wir uns vielmehr der damit verbundenen Probleme bewusst werden. Das Ziel ist, Gestaltungsoptionen zu erkennen. Dabei hilft uns die Frage: Wie nachhaltig ist die KI-Entwicklung derzeit und wie nachhaltig könnte sie sein?

Die Antwort fällt leicht, wenn wir auf den gegenwärtigen Trend blicken, immer größere KI-Modelle zu entwickeln und sie für unzählige Alltagsaufgaben einsetzen zu wollen. Es zeigt sich: Die Modelle werden derzeit kaum an Nachhaltigkeitskriterien gemessen. Das scheitert bereits am Erfassen des Energieverbrauchs oder daran, die Qualität der Datengrundlage komplexer KI-Modelle umfassend zu prüfen. Warum das so ist und welche Möglichkeiten bestehen, um das zu ändern, halten wir in dieser Ausgabe fest. Davor müssen wir allerdings noch einen kleinen, aber bedeutsamen Umweg gehen: hin zu den Wurzeln einer Perspektive auf KI, die Nachhaltigkeit wenig Raum lässt.

Die Vermenschlichung von KI-Systemen hat Tradition. Sie zeigt sich zum Beispiel im Paper „Sparks of Artificial General Intelligence: Early experiments with GPT-4“ von März 2023. Darin berichten Mitarbeiter*innen aus der Forschungsabteilung von *Microsoft* von ersten Anzeichen einer Künstlichen Allgemeinen Intelligenz, die sie dem Sprachmodell *GPT-4* zusprechen. Auch wenn sie verdeutlichen, dass dieses KI-Modell bislang nicht mit der menschlichen Intelligenz vergleichbar ist, wollen sie anhand von Beispielen zeigen, dass *GPT-4* verschiedene Aufgaben ähnlich gut lösen kann wie Menschen. Die Ausführungen der Autor*innen sind voll von Vergleichen mit menschlichen Fähigkeiten. Eines der Beispiele ist, dass *GPT-4* über wenige Wochen hinweg „lernt“, ein besser erkennbares Einhorn zu „zeichnen“ als zu Beginn der Testphase. Die Autor*innen schreiben *GPT-4* dabei nicht nur Intelligenz zu, sondern auch die Fähigkeit zu „sehen“.[1] Die Beschreibungen reihen sich in zahlreiche Medienberichte und fachliche Auseinandersetzungen ein, die KI-Modellen und -Verfahren menschliche Fähigkeiten zuweisen. Diese Tradition hat ihren Ursprung in der Erfindung des Begriffs „Künstliche Intelligenz“ in den 1950er-Jahren. Mit ihm schufen die Verfasser eines Antrags für Forschungsgelder rund um den US-amerikanischen Informatiker und Logiker John McCarthy einen Begriff, der ihre Arbeit für Geldgeber interessant machen sollte.[2]

Aber zurück zum Einhorn und damit auch zur Frage, ob die Vermenschlichung einerseits haltbar und andererseits sinnvoll ist. Was das Sprachmodell ausgibt, ist Programmiercode zur Darstellung eines Einhorns. Die Programmiersprache war Bestandteil der Trainingsdaten, aus denen das Modell entstanden ist. Daher kann das Modell auf Anfrage auch einen Code für ein Einhorn in dieser Programmiersprache erzeugen. Im Detail ist das weit komplexer als hier beschreiben. Dennoch „sieht“ oder „zeichnet“ das Modell aber keineswegs. Diese Skizze der Funktionsweise eröffnet einige Fragen: Können wir KI-Verfahren – also die Quantifizierung und Vermessung von Ausschnitten unserer Welt in Form von Daten sowie ihre statistische Gewichtung und Rekombination – tatsächlich mit menschlicher Intelligenz oder Kreativität vergleichen? Welches Menschenbild haben wir dann eigentlich? Spielen unsere Denk-, Reflexions- und Abwägungsprozesse, unsere Erfahrungen und die Fähigkeit zur Empathie keine Rolle? Der *Deutsche Ethikrat* spricht in seiner Stellungnahme, die er im März 2023 veröffentlichte, von einem Menschenbild, das er eindeutig von KI-Verfahren abgrenzt – auch wenn es sich um neue KI-Modelle handelt, die komplexer sind und für viele Anwendungszwecke infrage kommen.[3]

Vermenschlichungen erlauben es, für die Funktionsweise komplexer KI-Modelle scheinbar greifbare, zumindest aber bekannte Begriffe zu finden. Zugleich kann der Vergleich mit menschlichen Fähigkeiten eine gewisse Faszination auslösen. An dieser Stelle endet der Nutzen. Ganz pragmatisch stellen sich dann die Fragen: Was haben wir von diesen Vergleichen? Was können wir daraus ableiten? Die Antwort ist ernüchternd: wenig mehr als vage Hoffnungen und diffuse Ängste.

Um die Grenzen und Möglichkeiten von KI-Modellen zu erfassen und derzeitige Probleme bei der Entwicklung und im Einsatz zu erkennen, brauchen wir andere Perspektiven auf immer komplexer werdende KI-Modelle. Dabei kann uns der Abgleich mit Nachhaltigkeitskonzepten und -kriterien helfen. So werden beispielsweise Marktoligopole, Daten- und Wissenskonzentration, hohe Energie- und Ressourcenverbräuche oder Ungleichgewichte in den politischen und ökonomischen Teilhabechancen sichtbar. Diese umfassenden sozialen, kulturellen, ökologischen und ökonomischen Herausforderungen sind nicht erst mit komplexen und leistungsstarken KI-Systemen, sogenannten Foundation Models oder Basismodellen, in unser Leben getreten. Seitdem drängen sie sich allerdings noch mehr auf. Denn während die Notwendigkeit steigt, unseren Alltag, unsere Arbeits- und Infrastrukturen nachhaltiger zu gestalten, wird KI-Fortschritt hauptsächlich an einem Kriterium gemessen: bigger is better.[4]

Wenn wir in diesem Magazin über Nachhaltigkeit und komplexe KI-Modelle sprechen, geht es uns nicht um die Frage, wie wir mit KI die Welt retten können. Wir legen stattdessen den Finger in die Wunde und zeigen, wie weit die Entwicklung und der Einsatz von KI-Modellen derzeit von Nachhaltigkeitszielen entfernt sind. Das bedeutet aber nicht, dass wir die Welt vor KI retten müssen. Vielmehr geht es darum, die Gegenwart kritisch zu hinterfragen, um neue Ziele und mögliche Wege dorthin zu definieren – Wege, die eher im Einklang mit individuellem Wohlergehen, gesellschaftlichem Zusammenhalt, gesunden Märkten und unserer Umwelt stehen.

Trust Issues – der Podcast über vertrauenswürdige KI

In sechs Ausgaben findet unser Host gemeinsam mit seinen Gästen heraus, wodurch Künstliche Intelligenz vertrauenswürdig wird und welche Rolle wir Menschen dabei spielen. Jede Folge widmet sich einem Teilaspekt von Vertrauenswürdigkeit wie Transparenz, Fairness oder Sicherheit. Am Ende der Serie zieht er gemeinsam mit den Hörer*innen Bilanz: Wie steht’s um die Vertrauenswürdigkeit von KI?

Alle Episoden findet ihr in Kürze bei den gängigen Audio-Streaming-Diensten oder auf unserer Webseite www.zvki.de.

	Autorin
Jaana Müller-Brehm	

1 Vgl. Bubeck et al., S. 6 ff.
2 Vgl. Brooks, o. S.
3 Vgl. Deutscher Ethikrat, S. 88.
4 Vgl. Ananthaswamy.

X (Twitter): ZVKI_de
LinkedIn: ZVKI

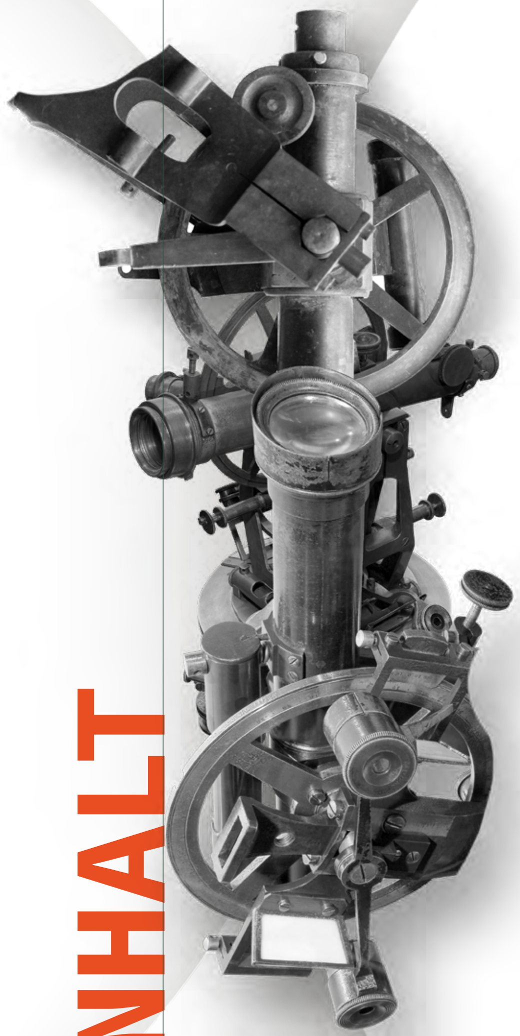
MEHR INHALTE:
Web: zvki.de
Instagram: zvki.de
X (Twitter): ZVKI_de
LinkedIn: ZVKI

MEHR INHALTE:
Web: zvki.de
Instagram: zvki.de
X (Twitter): ZVKI_de
LinkedIn: ZVKI

MEHR INHALTE:
Web: zvki.de
Instagram: zvki.de
X (Twitter): ZVKI_de
LinkedIn: ZVKI

EHR INHALTE:
eb: zvki.de
stagram: zvki.de
(Twitter): ZVKI_de
nkedIn: ZVKI

MEHR INHALTE:
Web: zvki.de
Instagram: zvki.de
X (Twitter): ZVKI_de
LinkedIn: ZVKI



INHALT

BENENNEN 6
Was bedeutet Nachhaltigkeit? Was sind Basismodelle?
Wichtige Begriffe dieser Ausgabe und was sie bedeuten

VERMESSEN 10
700.000 Liter Wasser für ein Sprachmodell
Zahlen und Fakten zur Nachhaltigkeit von Foundation Models

VERSTEHEN – Hinter den Glitzer blicken 20

Fortschritt auf wackligem Fundament
Der praktische Umgang mit Basismodellen im Abgleich mit Nachhaltigkeitskriterien

Warum die Regulierung bislang nicht genügt 24
Unzureichende Lösungsansätze in den KI-Verordnungsentwürfen und anderen Gesetzestexten

Nachhaltige Basis für KI 27
Strategien, die die Nachhaltigkeit von Foundation Models und ihres Einsatzes fördern

VERARBEITEN 30
Mit Open Source die ökonomische Teilhabe stärken
Praxisbeispiel für das Teilen von Daten, Modellen und Wissen

NACHFRAGEN 34
Wofür wollen wir unsere Ressourcen nutzen?
Friederike Rohde über gleichberechtigte Teilhabe, vielfältige Märkte und begrenzte Ressourcen

Welche Fragen bleiben, wenn der Hype abebbt? 38
Andreas Jungherr über gängige Bilder von KI und mögliche Folgen für die Demokratie

KOMBINIEREN 40
Fortschritt wohin?
Erkenntnis dieser Ausgabe: Nachhaltige Basismodelle gründen auf der Frage nach sinnvollen Einsatzzwecken.

VERBINDEN 44
Wozu all das?
Warum wir dieses Magazin schreiben

BELEGEN 46
Woher stammen die Informationen?
Die Quellen dieser Ausgabe

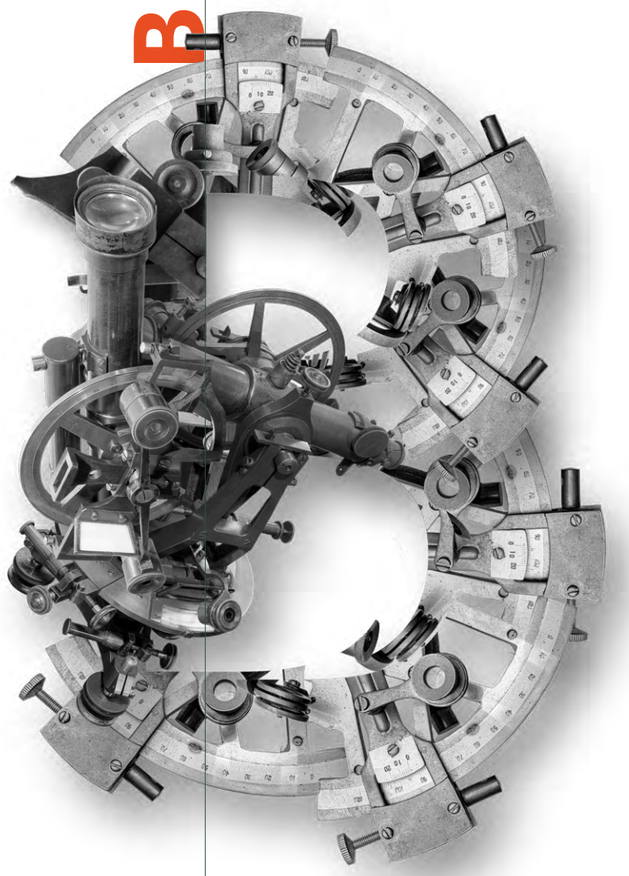
IMPRESSUM 50

EHR INHALTE:
eb: zvki.de
stagram: zvki.de
(Twitter): ZVKI_de
nkedIn: ZVKI

MEHR INHALTE:
Web: zvki.de
Instagram: zvki.de
X (Twitter): ZVKI_de
LinkedIn: ZVKI

MEHR INHALTE:
Web: zvki.de
Instagram: zvki.de
X (Twitter): ZVKI_de
LinkedIn: ZVKI

BENENNEN



WAS BEDEUTET NACHHALTIGKEIT? WAS SIND BASISMODELLE?

Was ist Nachhaltigkeit?

Der *Wissenschaftliche Beirat der Bundesregierung Globale Umweltveränderungen (WBGU)* bezeichnet die Umweltkrise in einem Gutachten von 2019 als „eine der vielen Nachhaltigkeitsherausforderungen, welche die industrielle Moderne hervorgebracht hat“.[1] Die Autor*innen weisen darauf hin, dass wir Bewältigungsstrategien nicht von Fragen nach sozialer Gerechtigkeit und gesellschaftlichem Zusammenhalt trennen können.[2] Zahlreiche fachliche Betrachtungen sehen Nachhaltigkeit als Querschnittsthema, das alle Bereiche menschlichen Lebens und gesellschaftlicher Systeme betrifft. Sie weisen dem Begriff Nachhaltigkeit eine ökologische, soziale und ökonomische Dimension zu.[3]

Diese Verbindungen zeichnen sich auch in den „Sustainable Development Goals“ der *Vereinten Nationen* (*United Nations – UN*) ab.[4] Im September 2015 verabschiedete die Generalversammlung der *UN* „Die Agenda 2030 für nachhaltige Entwicklung“ mit 17 Zielen und 169 Zielvorgaben. Sie sollen bis 2030 umgesetzt werden und dazu beitragen, Menschenrechte und Wohlstand für alle zu verwirklichen, Armut und Hunger zu begegnen, eine gesunde Umwelt sicherzustellen, den Planeten vor Schädigungen zu schützen sowie Frieden und globale Partnerschaften zu fördern. Die Agenda benennt dabei ebenfalls die drei Dimensionen von Nachhaltigkeit: ökologisch, sozial und ökonomisch.[5] Vielfach üben Studien von Forschungseinrichtungen und Analysen zivilgesellschaftlicher Organisationen Kritik an der Vereinbarkeit der Ziele: Sie stünden teilweise im Widerspruch zueinander.[6] Bildungskonzepte für nachhaltige Entwicklung betonen darüber hinaus eine kulturelle Dimension. Sie soll beispielsweise traditionelles Wissen oder unterschiedliche Werte und Nachhaltigkeitsverständnisse von Gemeinschaften auf der Welt berücksichtigen.[7]

Was hat KI mit Nachhaltigkeit zu tun?

Es gibt zahlreiche Möglichkeiten, KI-Systeme so zu gestalten und einzusetzen, dass sie dazu beitragen, die Ziele einer nachhaltigen Entwicklung zu erreichen. Das gilt vor allem für den Bereich Umwelt und dafür, Klimafolgen zu bekämpfen. Zum Beispiel können Anwendungen dabei unterstützen, Ölverschmutzungen in den Meeren zu erkennen oder Waldbrände gezielt zu bekämpfen.[8] Allerdings wirken sich die Entwicklung und der Einsatz von KI-Modellen auch problematisch auf unsere Umwelt, unser Zusammenleben und unser Wohlergehen als Einzelne aus: etwa im Zusammenhang mit zunehmenden Energie- und Ressourcenverbräuchen sowie vermehrten Treibhausgasemissionen. Auch soziale Nachhaltigkeitskriterien sind betroffen. So können beispielsweise die Arbeitsprozesse, in denen KI-Systeme entstehen, und die Ergebnisse, die sie erzielen, strukturelle Ungerechtigkeiten verstärken.[9]

Deshalb definiert die *High-Level Expert Group on Artificial Intelligence (AI HLEG)* der *Europäischen Kommission* in ihren Ethik-Leitlinien für vertrauenswürdige KI gesellschaftliches und ökologisches Wohlergehen als eines von sieben zentralen Kriterien, die beim Entwickeln und Nutzen von KI-Modellen

1 Vgl. WGBU, S. 2.
2 Vgl. ebd.
3 Vgl. Henkel et al., S. 12.
4 Vgl. Vereinte Nationen, o. S.
5 Vgl. Generalversammlung Vereinte Nationen, S. 1 ff.
6 Vgl. z. B. Reiner, o. S.; Potsdam-Institut für Klimafolgenforschung, o. S.
7 Vgl. Stoltenberg, o. S.
8 Vgl. Vinuesa, S. 2 ff.
9 Vgl. van Wynsberghe, S. 213 f.

zu beachten sind. Dabei betont sie vor allem die Bedeutung umweltfreundlicher und -verträglicher KI sowie die Notwendigkeit, die sozialen Auswirkungen für den*die Einzelne*n, die Gesellschaft und die Demokratie zu beachten.[1]

Daran anschlussfähig ist die Definition von nachhaltiger KI eines Autor*innenteams des *Instituts für ökologische Wirtschaftsforschung* und des Projekts *SustAIn*. Das Team arbeitet an einem Kriterienset zur Nachhaltigkeitsbewertung von KI. Die Autor*innen schreiben: „Eine nachhaltige KI ist aus unserer Perspektive vorhanden, wenn Entwicklung und Einsatz dieser Systeme die planetaren Grenzen respektier[en], keine problematischen ökonomischen Dynamiken verstärk[en] und den gesellschaftlichen Zusammenhalt nicht gefährde[en].“[2] Kriterien, die hierbei eine Rolle spielen, betreffen also beispielsweise den Ressourcen- und Energieverbrauch, Fairnessziele sowie die Nachvollziehbarkeit der Funktionsweise und Auswirkungen von KI-Systemen.[3]

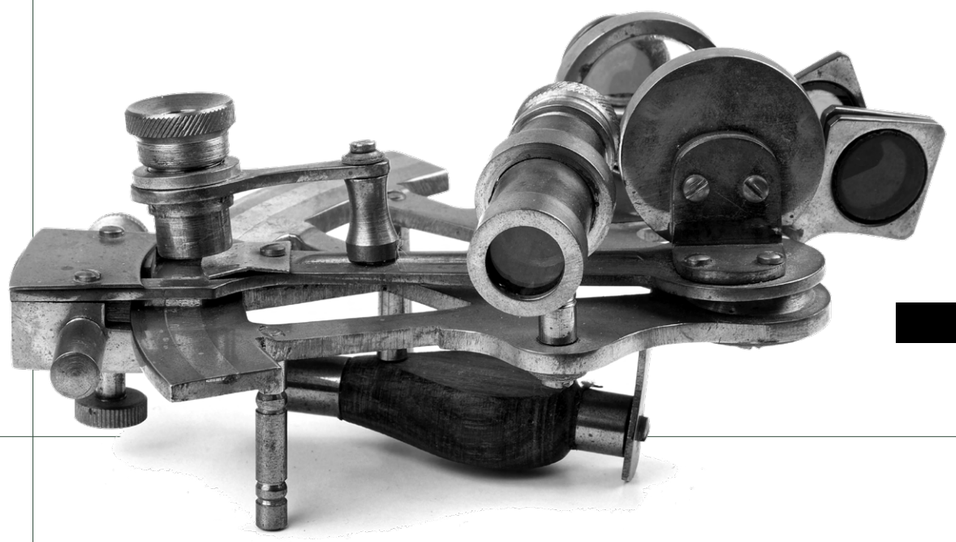
In Bezug auf Basismodelle und KI-Anwendungen, die darauf aufbauen, ergeben sich besondere Herausforderungen, um solche Kriterien ausreichend zu beachten.

FAQ: Funktioniert KI wie ein menschliches Gehirn?

Nein. Es gibt Methoden in der KI-Forschung, die sich an der Bauweise unseres Gehirns orientieren und versuchen, es mathematisch nachzubilden. Eine dieser Methoden sind künstliche neuronale Netze. Daten werden dabei von verschiedenen Knotenpunkten nacheinander verarbeitet. Die Knotenpunkte sind wie ein Netz miteinander verbunden, was an den Aufbau des menschlichen Gehirns erinnert. Die Daten durchlaufen in diesem Netz einen vorgegebenen Pfad, der bestimmt, welches Ergebnis das KI-System ausgibt.

Weitere häufig gestellte Fragen finden Sie hier: www.zvki.de > KI-Navigator > Unsere Inhalte > FAQs

1 Vgl. High-Level Expert Group on AI, S. 18 und S. 23.
2 Vgl. Rohde et al., S. 30.
3 Vgl. ebd., S. 32.



Was sind Basismodelle?

Foundation Models[1] oder „General Purpose AI“[2] bezeichnen KI-Modelle oder auf ihnen basierende Anwendungen, die für viele verschiedene Aufgaben infrage kommen, große Datenmengen verarbeiten und über eine komplexe Modellarchitektur verfügen.[3] Der *Deutsche Ethikrat* bezeichnet sie in seiner Stellungnahme „Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz“ von März 2023 als breite KI. Im Vergleich dazu erfüllen bisher gängige KI-Systeme eine bestimmte und klar abgesteckte Aufgabe.[4] Der *Deutsche Ethikrat* betont, dass breite KI sowie denkbare Zukunftsszenarien solcher Modelle und Anwendungen keine Form sogenannter starker KI sind. Letztere würde „jenseits der möglicherweise perfekten Simulation menschlicher Kognition auch über mentale Zustände, Einsichtsfähigkeit und Emotionen verfügen“.[5]

Die Autor*innen von „On the Opportunities and Risks of Foundation Models” des *Stanford Institute for Human-Centered Artificial Intelligence (HAI)* prägen den Begriff Foundation Models. Demnach ist ein Merkmal dieser Modelle das unüberwachte Training mithilfe umfangreicher Datenmengen. Als Beispiele verweisen sie auf *BERT* (*Google*), *GPT-3* (*OpenAI*) und *CLIP* (*OpenAI*).[6] Der Begriff „foundation“ bezieht sich in diesem Zusammenhang darauf, dass die Modelle eine Grundlage für viele verschiedene Anwendungen darstellen können.[7]

Vor allem Weiterentwicklungen in drei Bereichen gelten als Voraussetzung für solche Basismodelle: Sie betreffen die Hardware, beispielweise die Prozessoren, die Modell-Architektur und die Verfügbarkeit von sehr umfangreichen Trainingsdaten.[8] Modell-Architekturen wie die sogenannte Transformer-Architektur sind leistungsfähiger. Sie wurde 2017 von *Google* entwickelt und ermöglicht kürzere Berechnungswege, sodass mehr Daten und Knotenpunkte einbezogen werden können. Diese Knotenpunkte stehen für Gewichtungen einzelner Datenpunkte und werden auch als Parameter bezeichnet. Eine solche Architektur erlaubt Berechnungen, die von den konkreten Eingabedaten unabhängig sind. Dadurch funktionieren Basismodelle über bestimmte Aufgabenfelder hinweg.[9] Sie sind effizienter und flexibler als bisher gängige Systeme, die in den Bereich enger KI fallen.[10]

Zu den Basismodellen zählen unter anderem Large Language Models (LLM), die Sprache verarbeiten und ausgeben wie die *GPT*-Modelle. So basiert etwa *ChatGPT* auf den Modellen *GPT-3* und *GPT-4*.[11] Ein Large Language Model wie *GPT-4* stellt als Foundation Model demnach nur eine Komponente eines KI-Systems dar. Foundation Models werden deshalb auch als „vortrainiert“ („pre-trained“) bezeichnet.[12] Damit Systeme wie *ChatGPT* für die vorgesehenen Aufgaben funktionieren und die gewünschte Leistungsfähigkeit erreichen, müssen Menschen an ihrer Feinjustierung arbeiten. Sie bewerten beispielsweise die Antworten, die das System liefert, hinsichtlich ihrer Sinnhaftigkeit für den Dialogverlauf und speisen damit Feedback in das System ein.[13] Einige Autor*innen bezeichnen KI-Modelle und -Anwendungen, die etwa Texte, Bilder, Videos oder Audios erzeugen, auch als generative KI.[14]

1 Vgl. z. B. Bommasani et al.
2 Vgl. z. B. Maham/ Küspert. Der Begriff kann als problematisch angesehen werden, da er den Eindruck vermittelt, dass es KI-Systeme gibt, die für jeden Zweck infrage kommen. Das ist nicht der Fall. Die Anwendungszwecke sind lediglich vielfältiger, aber weiterhin begrenzt.
3 Vgl. Bommasani et al., S. 3.
4 Vgl. Deutscher Ethikrat, S. 88.
5 Vgl. ebd.
6 Vgl. Bommasani et al., S. 3.
7 Vgl. Harrod, o. S.
8 Vgl. Bommasani et al., S. 4.
9 Vgl. ebd., S. 75 f.
10 Vgl. OECD, S. 24.
11 Vgl. Maham/ Küspert, S. 12.
12 Vgl. Bommasani et al., S. 8.
13 Vgl. Albrecht, S. 10.
14 Vgl. ebd., S. 19.

Autorin

Jaana Müller-Brehm

given that training ML models requires significant computational resources, energy and materials. In the present article, we aim to quantify the carbon footprint of BLOOM, a 176-billion parameter language model, across its life cycle. We estimate that BLOOM's final training emitted approximately 24.7 tonnes of CO2eq if we consider only the dynamic power consumption, and 50.5 tonnes if we account for all processes ranging from equipment manufacturing to energy-based operational consumption. We also study the energy requirements and carbon emissions of its deployment for inference via an API endpoint receiving user queries in real-time. We conclude with a discussion regarding the difficulty of precisely estimating the carbon footprint of ML models and future research directions that can contribute towards improving carbon emissions reporting.","language":"en","note":"arXiv:2211.02001 [cs]","number":"arXiv:2211.02001","publisher":"arXiv","source":"arXiv.org","title":"Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model","URL":"http://arxiv.org/abs/2211.02001","author":[{"family":"Luccioni","given":"Alexandra Sasha"},{"family":"Viguier","given":"Sylvain"},{"family":"Ligozat","given":"Anne-Laure"}],"accessed":{"date-parts":[["2023",6,29]]},"issued":{"date-parts":[["2022",11,3]]}}],"schema":"https://github.com/citation-style-language/schema/raw/master/csl-citation.json"}

VERMESSEN

700.000 LITER WASSER FÜR EIN SPRACHMODELL

Basismodelle sollen das Potenzial besitzen, unsere Welt in vielen Bereichen nachhaltiger zu machen. Studien zeigen allerdings: Die Entwicklung und der Einsatz solcher Modelle sind bisweilen alles andere als nachhaltig.

Wie es um die Nachhaltigkeit von Basismodellen steht, ist gar nicht so einfach zu erfassen. Kommerzielle Entwickler, etwa *Google* oder *OpenAI*, geben die dafür benötigten Daten selten preis. Hinzu kommt, dass es bisher noch keine standardisierte Vorgehensweise gibt, um zum Beispiel den CO2-Ausstoß oder den Energie- und Ressourcenverbrauch dieser Modelle zu messen und zu dokumentieren. Das erschwert es, sie miteinander zu vergleichen.[1] Zudem konzentrieren sich die meisten Studien bisher vor allem auf die Trainingsphase eines Modells. Der gesamte Lebenszyklus – von der Produktion der nötigen Hardware über das Training und den Einsatz der Modelle bis hin zur Entsorgung einzelner Komponenten – wird bislang nur unzureichend erfasst.[2] Allerdings entstehen an all diesen Stellen CO2-Emissionen und Ressourcen und Energie werden verbraucht.

Aber immerhin: Das Thema erhält in jüngster Zeit mehr Aufmerksamkeit. 76 Prozent aller Studien zur Nachhaltigkeit von KI-Systemen sind seit 2020 erschienen.[3] Der Großteil beschäftigt sich vor allem mit ökologischen und technischen Faktoren, zum Beispiel dem CO2-Ausstoß oder der Energieeffizienz verschiedener Trainingsmethoden.[4]

Immer mehr, immer größer

Um Basismodelle leistungsfähiger zu machen, werden sie von ihren Entwickler*innen mit immer größeren Datensätzen trainiert.

- *Googles Pathways Language Model* (*PaLM*) wurde im Jahr 2022 mit 540 Milliarden Parametern auf einem Datensatz von 780 Milliarden Tokens – also Datenpunkten – trainiert.

- Zum Vergleich: *GPT-3*, eines der KI-Modelle hinter der kostenfreien Version des Chatbots *ChatGPT*, wurde im Jahr 2020 mit 175 Milliarden Parametern und etwa 300 Milliarden Tokens trainiert.

Um Datenmengen dieser Größenordnung verarbeiten zu können, werden immer mehr und immer leistungsstärkere Prozessoren benötigt.[5] Ihre Produktion ist in den vergangenen Jahren entsprechend deutlich angestiegen.[6]

- Laut dem Marktforschungsunternehmen *Gartner* wird der Anteil hochspezialisierter Chips, die für das Training großer KI-Modelle in Rechenzentren verwendet werden, von derzeit drei Prozent auf mehr als 15 Prozent im Jahr 2026 ansteigen.7

Um diese Chips herzustellen, werden unter anderem Schwermetalle und Seltene Erden abgebaut. Die dabei freigesetzten Emissionen sowie die schlechten bis ausbeuterischen Arbeitsbedingungen von Arbeiter*innen, zum Beispiel in Kupferminen, werden bislang kaum systematisch erfasst.[8]

Aufwendige Trainingsphase

Bevor Basismodelle einsatzbereit sind, müssen sie teils monatelang trainiert werden. Diese Trainingsphase kann extrem energieintensiv sein.

- Um *GPT-3* zu trainieren, wurden schätzungsweise 1.287 Megawattstunden (MWh) Strom verbraucht.[9] Das entspricht etwa dem jährlichen Strombedarf von 400 deutschen Durchschnittshaushalten.[10]

Wie hoch die Emissionen sind, die durch das Training verursacht werden, hängt unter anderem davon ab, wann und wo ein Modell trainiert wird. Steht ein Rechenzentrum etwa in Quebec und zieht viel Strom aus Wasserkraft, fallen die Emissionen geringer aus als bei einem Rechenzentrum, das in Tallinn steht, bei dem fossile Brennstoffe einen Großteil des Energiemix ausmachen. Während der Strom für das Training von *GPT-3* zum Beispiel aus kohlenstoffintensiven Energiequellen wie Gas oder Kohle kam, bezog *Googles* Rechenzentrum in Oklahoma, in dem *PaLM* trainiert wurde, laut eigenen Angaben 89 Prozent des benötigten Stroms aus kohlestofffreien Energiequellen.[11]

1 Vgl. Luccioni et al., S. 7.
2 Vgl. ebd., S. 1 ff.
3 Verdecchia et al., S. 1.
4 Vgl. ebd., S. 2.
5 Vgl. Narang, o. S.
6 Vgl. Khan/ Mann, S. 14.
7 Vgl. Mehta, o. S.
8 Vgl. Pattison, o. S.
9 Vgl. Luccioni et al., S. 7.
10 Vgl. Albrecht, S. 30.
11 Vgl. AW AlgorithmWatch 2023 a, S. 13.

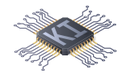
- Die Emissionen, die während der Entwicklungsphase moderner Large Language Models entstehen, sind vergleichbar mit den Emissionen, die fünf Autos über ihre gesamte Lebensdauer hinweg ausstoßen.[1]
- Als Beispiel: Das Training von *GPT-3* hat laut Äquivalenzrechnungen schätzungsweise einen Ausstoß von 502 Tonnen CO2 verursacht.[2]
- Um die Emissionen auszugleichen, die für das Training von *Googles* Sprachmodell *BERT* angefallen sind, müsste man 40 Bäume 10 Jahre lang wachsen lassen.[3]

Fünf Autos für das Training eines Large Language Models

Das Training eines Large Language Models verursacht ungefähr so viele Emissionen, wie fünf Autos über ihre gesamte Lebensdauer hinweg ausstoßen.[1]

1 Vgl. Strubell et al., S. 1.

x5

1x 

1 Vgl. ebd.
2 Vgl. Luccioni et al., S. 7.
3 Vgl. Bommasani et al., S. 140.

1 Vgl. Li et al., S. 1.
2 Vgl. ebd., S. 2.
3 Vgl. Li et al., S. 3.

Trinkwasser kühlt die Server

Um die maximale Leistung erbringen zu können, benötigen die Server, auf denen die Modelle laufen, nicht nur Strom. Damit die Prozessoren während der langwierigen und zunehmend komplexen Berechnungen nicht überhitzen, müssen sie gekühlt werden. Das passiert in der Regel mit großen Mengen Wasser. In ihrer Studie „Making AI Less Thirsty: Uncovering and Adressing the Secret Water Footprint of AI Models“ aus dem Jahr 2023 weisen die Autor*innen darauf hin, dass der Wasserverbrauch im Vergleich zu anderen Nachhaltigkeitsfaktoren bisher vernachlässigt wurde, obwohl dieser teilweise erheblich ist:

- Um *GPT-3* zu trainieren, mussten *Microsofts* Datenzentren mit etwa 700.000 Litern Wasser gekühlt werden.[1]
- *Googles* eigene Datenzentren in den USA haben im Jahr 2021 insgesamt 12,7 Milliarden Liter Wasser verbraucht, um Server zu kühlen. 90 Prozent davon waren Trinkwasser.[2]

Die Server müssen natürlich nicht nur während der Trainingsphase gekühlt werden, sondern auch später, wenn die Modelle zum Einsatz kommen. Für eine kurze Unterhaltung mit *ChatGPT* wird laut Äquivalenzrechnungen etwa ein halber Liter Wasser verbraucht.[3]

Durstiger als eine vierköpfige Familie

GPT-3 schluckt im Training etwa 700.000 Liter Wasser. Das ist dreimal mehr, als eine vierköpfige Familie durchschnittlich im Jahr verbraucht.[1]

1 Vgl. European Environment Agency, o. S.; Li et al., S. 1.

Eingeschränkter Zugang

Teure und hochspezialisierte Hardware, aufwendige und ressourcenintensive Trainingsprozesse, Zugang zu großen Datensets: Nicht jede*r kann sich die Entwicklung von Basismodellen beziehungsweise von komplexen generativen KI-Modellen leisten. So hat *OpenAI* Schätzungen zufolge etwa 4,6 Millionen Dollar in die Entwicklung von *GPT-3* investiert.[1] Hinzu kommt, dass die nötige technische Infrastruktur in den Händen einiger weniger Unternehmen liegt:

- *Google*, *Amazon* und *Microsoft* besitzen zusammen ca. 65 Prozent der weltweiten Cloud-Computing-Infrastruktur, auf der Modelle trainiert werden und später laufen.[2]

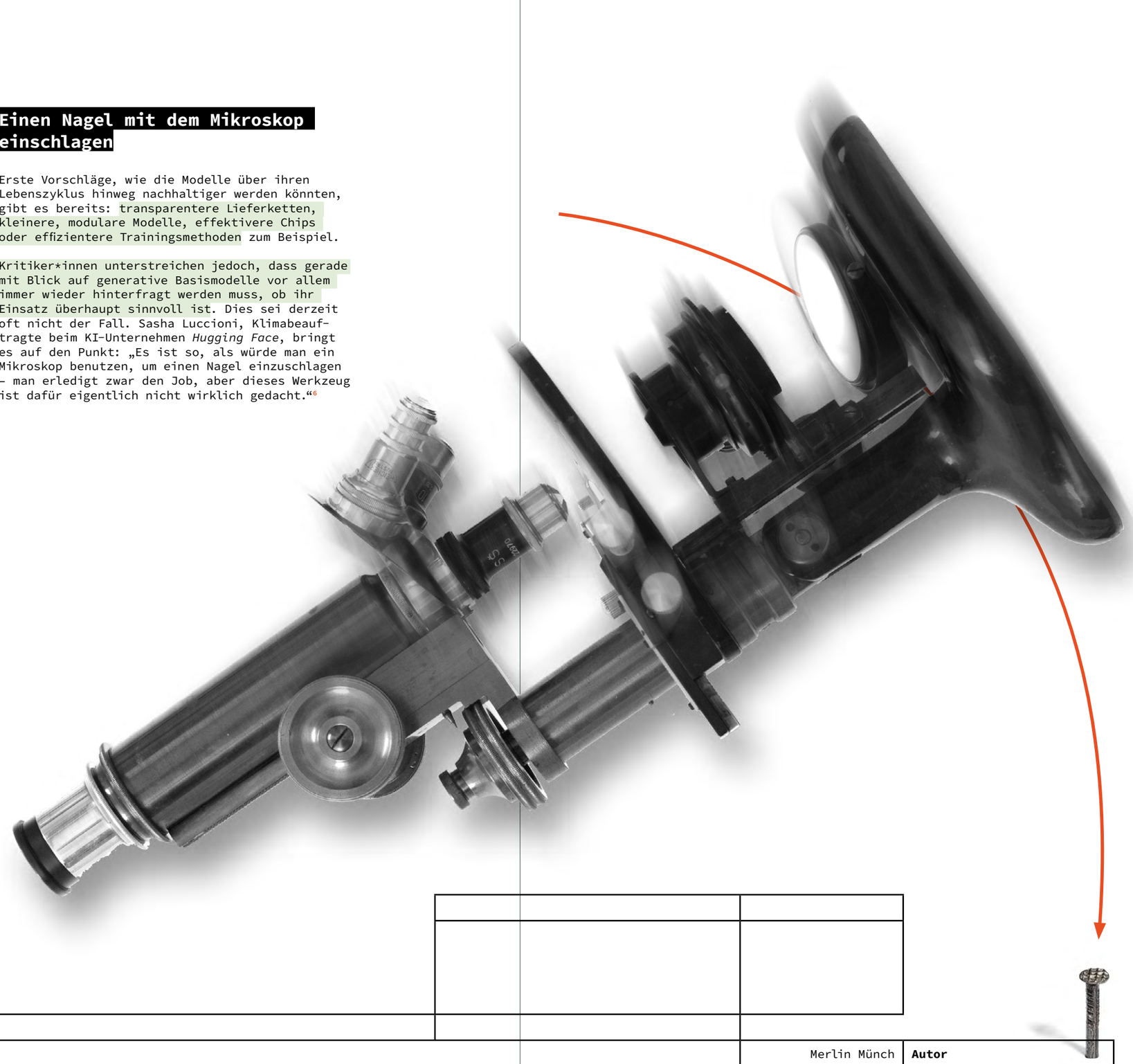
Das begünstigt eine Marktkonzentration, bei der einige wenige Akteure die Entwicklung großer Basismodelle dominieren.[3] So haben derzeit beinahe alle aktuellen Sprachmodelle ihren Ursprung in einem der wenigen Basismodelle wie *BERT*, *T5* (beide *Google* bzw. *Alphabet*) oder *RoBERTa* (*MetaAI*).[4]

- Die 24 bekanntesten generativen Basismodelle stammen von nur fünf Unternehmen. Nur zwei der 24 Modelle wurden ohne Beteiligung der Industrie von Universitäten entwickelt.[5]

Einen Nagel mit dem Mikroskop einschlagen

Erste Vorschläge, wie die Modelle über ihren Lebenszyklus hinweg nachhaltiger werden könnten, gibt es bereits: transparentere Lieferketten, kleinere, modulare Modelle, effektivere Chips oder effizientere Trainingsmethoden zum Beispiel.

Kritiker*innen unterstreichen jedoch, dass gerade mit Blick auf generative Basismodelle vor allem immer wieder hinterfragt werden muss, ob ihr Einsatz überhaupt sinnvoll ist. Dies sei derzeit oft nicht der Fall. Sasha Luccioni, Klimabeauftragte beim KI-Unternehmen *Hugging Face*, bringt es auf den Punkt: „Es ist so, als würde man ein Mikroskop benutzen, um einen Nagel einzuschlagen – man erledigt zwar den Job, aber dieses Werkzeug ist dafür eigentlich nicht wirklich gedacht.“[6]



Merlin Münch **Autor**

1 Vgl. Luccioni, o. S.
2 Vgl. Richter, o. S.
3 Vgl. Luccioni, o. S.
4 Vgl. Bommasani et al., S. 5.
5 Vgl. Albrecht, S. 31.
6 Originalzitat: „‘It’s frustrating because actually there are so many low-impact, efficient AI approaches and methods that people have developed over the years, but people want to use generative AI for everything’, said Luccioni. ‘It’s like using a microscope to hammer in a nail – it might do the job but that’s not really what this tool is meant for.’”, Singh, o. S.

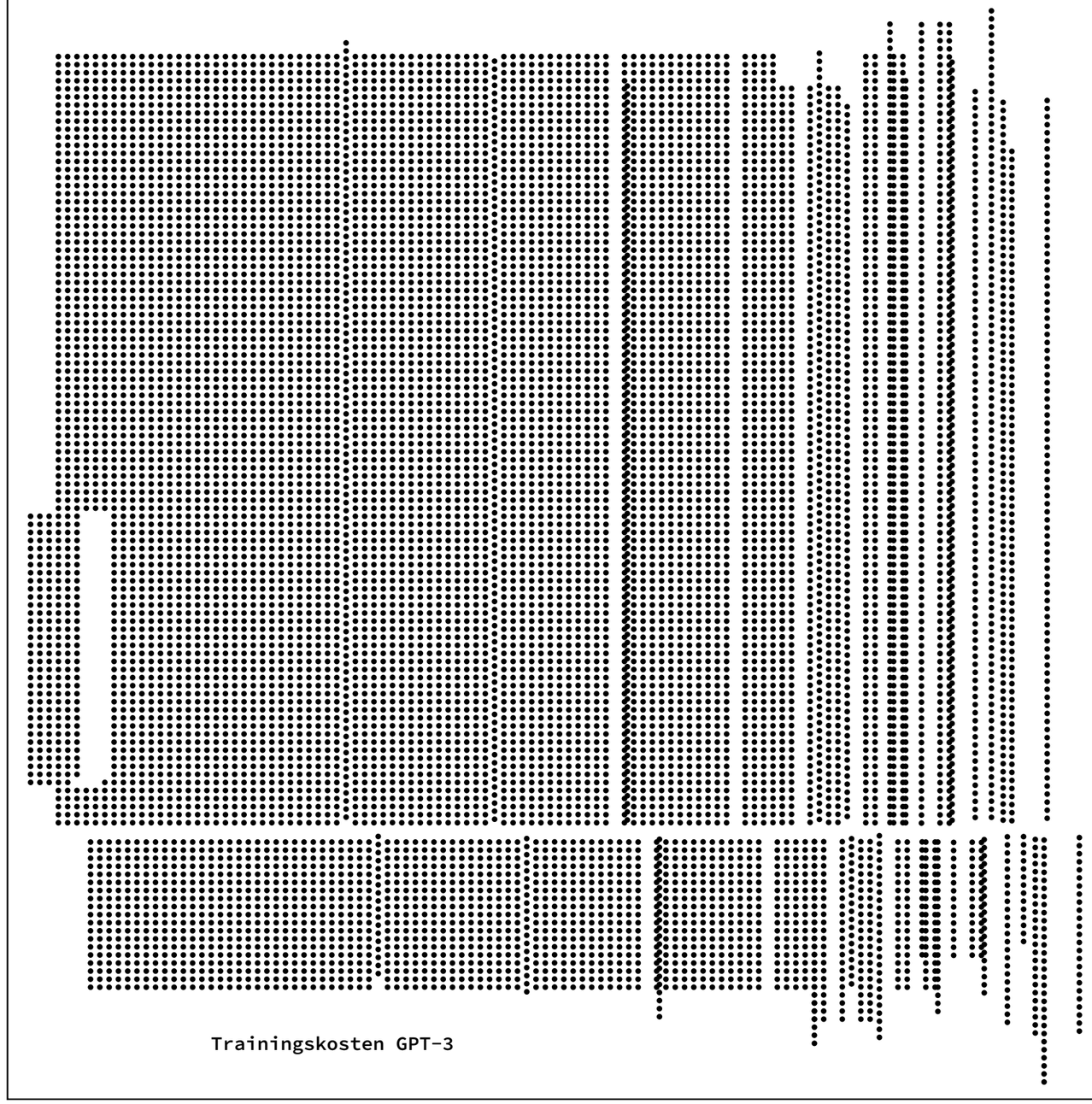
Wie viele Monatsgehälter kostet ein Large Language Model?

Hinter *ChatGPT* steckt viel menschliche Arbeit, die oft unter ausbeuterischen Bedingungen verrichtet wird. Clickworker*innen leben mit einem Stundenlohn von etwa zwei Dollar.[1]

1 Vgl. Perrigo, 2023, o. S.; Luccioni, 2023, o. S.

1 : 13.295

Gehalt einer Clickworkerin

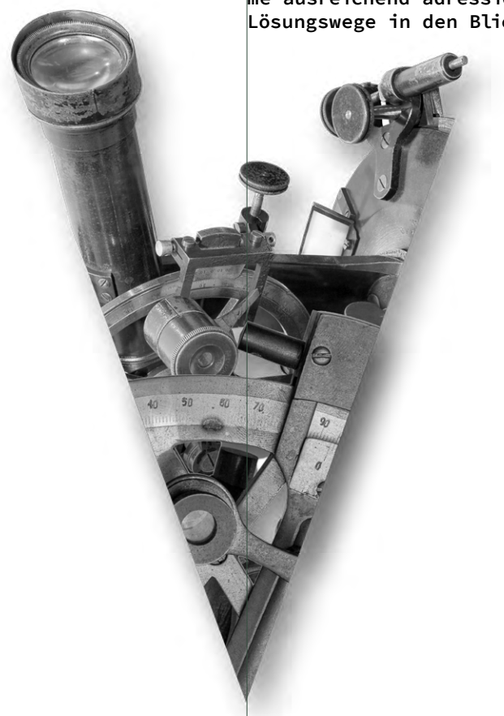


Trainingskosten GPT-3

VERSTEHEN

HINTER DEN GLITZER BLICKEN

Die neue Generation komplexer KI-Modelle versucht, durch Größe, Leistungsfähigkeit und vielfältige Anwendungsszenarien zu überzeugen. Teilweise gelingt das auch, denn auf diesen Basismodellen aufbauende KI-Systeme können beeindruckende Ergebnisse liefern – etwa sinnvoll wirkende Texte oder tanzbare Songs. Doch es gibt eine Kehrseite. Sie ist facettenreich, mit der Funktionsweise der Modelle verwoben und mit den Strukturen, in denen sie entstehen und wirken. Diese Kehrseite wird sichtbar, wenn wir darauf aufbauende KI-Systeme an der Frage messen, wie nachhaltig sie sind. Ist das notwendig? Wir kommen nicht darum herum, wenn wir KI-Systeme so gestalten und nutzen möchten, dass sie sich an unserer Umwelt sowie unserem Wohlergehen als Einzelne und als Gesellschaft orientieren. Deshalb treten wir jetzt einen Schritt zurück und erfassen die mit Basismodellen einhergehenden Probleme sowie ihr verändertes Ausmaß. Danach können wir prüfen, ob bestehende Überlegungen – beispielsweise der Regulierung – diese Probleme ausreichend adressieren, um neue Lösungswege in den Blick zu nehmen.



Fortschritt auf wackligem Fundament

Basismodelle verdeutlichen Probleme, die wir bereits von anderen KI-Modellen kennen. Dabei geht es zum Beispiel um mangelnde Nachvollziehbarkeit, Fairness, Marktmacht und Auswirkungen auf die Umwelt. Diese Aspekte bekommen durch die Funktionsweise der neuen KI-Modelle sowie die Art, wie sie entstehen und genutzt werden, neue Ausprägungsformen und eine besondere Brisanz. Zugleich verliert manch ein Lösungsansatz an Wirkung.

Bei komplexen Basismodellen handelt es sich um neue Entwicklungen, die eher Prototypen ähneln als umfassend erprobten Systemen. Die Folgen ihres Einsatzes sind schwer abzuschätzen, weil die Anwendungsbereiche nicht eindeutig definiert und damit auch die Nutzer*innen- und Betroffenengruppen unbekannt sind.[1] Betrachten wir Basismodelle und die Systeme, in denen sie entstehen und wirken, gemeinsam mit Nachhaltigkeitskriterien und -zielen, können wir zentrale übergeordnete Probleme skizzieren.

Macht den Mächtigen

Vornehmlich bereits etablierte und erfolgreiche Tech-Konzerne verfügen über die notwendigen Infrastrukturen, Kompetenzen und finanziellen Ressourcen, um die Entwicklung von Basismodellen voranzutreiben.[2] In einigen Fällen haben bestimmte Unternehmen in mehreren Bereichen eine ausgeprägte Marktposition: Für *Microsoft* trifft das zum Beispiel auf die Text- und Informationsverarbeitung zu. In beiden Feldern bietet das Unternehmen bereits Softwarelösungen an und kann gleichzeitig auf umfangreiche Daten zugreifen. Das ist ein zusätzlicher Vorteil bei der Entwicklung und Implementierung von Large Language Models. Er trägt dazu bei, die bestehende Marktmacht weiter auszubauen.[3]

Die Vormachtstellung von Unternehmen in diesem Feld schlägt sich auch in der Finanzierung und Ausrichtung von Forschungsarbeiten nieder. Es kann sein, dass kommerzielle Interessen mit einem gesellschaftlichen Nutzen einhergehen, beispielsweise in Bezug auf die Leistungsfähigkeit oder Sicherheit eines Systems. Jedoch trifft das nicht in allen Fällen zu. Es gibt etwa wenig erkennbare Anreize für privatwirtschaftliche Unternehmen, die eigenen Forschungen und Entwicklungen so auszurichten, dass marginalisierte oder ärmere Bevölkerungsteile davon profitieren. Ebenso wenig sind Tendenzen sichtbar, dass erfolgreiche Technikunternehmen daran interessiert sind, dezentrale, offene Ökosysteme und damit partizipative Ansätze für die Entwicklung von Basismodellen voranzutreiben.[4]

Da diese Modelle für eine Vielzahl von Anwendungen genutzt werden können und es kaum Alternativen zu den kommerziellen Angeboten gibt, kommt einigen wenigen großen Tech-Unternehmen eine Schlüsselfunktion zu. Sie legen etwa fest, wie viel ein Unternehmen bezahlen muss, um ihr Modell zu integrieren, und an welche Bedingungen sie sich dabei halten müssen.[5] Insgesamt verstärken sich dadurch bestehende wirtschaftliche Macht- und Kompetenzungleichgewichte.

Anschein von Realität

Die Schlüsselposition einiger weniger Unternehmen wird ebenfalls deutlich, wenn wir auf den Zugang zu Inhalten und die Darstellung von Sachverhalten blicken. Large Language Models kommen zum Beispiel vielfach zum Einsatz, um Informationen zu erfragen oder Inhalte zusammenzufassen. Das ausgegebene Ergebnis wirkt sich darauf aus, was wir zu einem bestimmten Thema wissen oder auch als relevant ansehen. Gewisse Informationen oder Darstellungen können dabei ausgeschlossen werden. Die an der Entwicklung Beteiligten entscheiden, welche Inhalte zugänglich sind und welche nicht. Das kann einerseits hilfreich sein, da wir dann beispielsweise bestimmte Informationen nicht erhalten, die als gefährlich gelten, wie eine Anleitung zum Bau einer Bombe. Andererseits bedeutet das auch, dass bestimmte Themen ausgeklammert werden können, beispielsweise aufklärende Inhalte zu Schwangerschaftsabbrüchen.[6]

Das gängige Nutzungsszenario, komplexe Dialogsysteme zu Informationszwecken zu verwenden, wirft eine weitere Frage auf: Sind solche Systeme überhaupt dafür geeignet, verlässliche Informationen zu liefern? Zahlreiche Berichte von falschen Darstellungen und inexistenten Links als Quellenverweisen bei der Nutzung von *ChatGPT* legen nahe, dass sie es nicht sind. Trotz der menschlichen Feinjustierung sind diese Systeme nicht dafür ausgelegt, faktisch korrekte Antworten zu geben oder diese nachzuweisen.[7]

Nicht zuletzt ermöglichen es vor allem generative KI-Anwendungen, gezielt irreführende Inhalte zu produzieren, zu personalisieren und zu streuen, um auf diese Weise Meinungsbildungsprozesse und öffentliche Diskurse zu beeinflussen. Ein Beispiel hierfür ist ein KI-generiertes Bild auf dem *Instagram*-Kanal des *AfD*-Abgeordneten Norbert Kleinwächter, mit dem er gezielt versucht, gegen Geflüchtete zu hetzen.[8]

1 Vgl. Bommasani et al., S. 7 ff.
2 Vgl. Ananthaswamy, S. 203 f.
3 Vgl. van Dis et al., S. 31.
4 Vgl. Bommasani et al., S. 9 f.
5 Vgl. Meineck, o. S.
6 Vgl. ebd.
7 Vgl. Albrecht, S. 20 f.
8 Vgl. Schneider, o. S.

Unsichtbare Perspektiven, verzerrte Bilder und überspitzte Meinungen

In den Ergebnissen generativer KI-Anwendungen werden häufig Stereotype sowie bestimmte Sichtweisen und Deutungen reproduziert, während weniger verbreitete Darstellungen und Lebensrealitäten unsichtbar bleiben. Auch in den Trainingsdaten inbegriffene dominante Weltdeutungen können in den Ergebnissen enthalten sein, etwa bestimmte Bilder von Weiblichkeit und Männlichkeit.[1] Im globalen Vergleich zeigt sich, dass bestimmte Perspektiven dominieren, während andere unterrepräsentiert oder schlichtweg nicht vorhanden sind. Der Blick auf die Veröffentlichungen von Large Language Models seit 2019 veranschaulicht das: Ungefähr 54 Prozent der Modelle stammen aus den USA, knapp 22 Prozent aus Großbritannien. Die restlichen Entwicklungen kommen aus sechs weiteren Ländern, unter denen kein südamerikanisches oder afrikanisches Land ist.[2]

Stereotype und Verzerrungen können auch die Repräsentanz von Meinungen betreffen, wie erste Untersuchungen eines Forscher*innenteams der *Vanderbild University* zeigen. Die Autor*innen der Studie messen in den mit *ChatGPT* generierten Inhalten eine erhöhte parteipolitische und rassistische Polarisierung im Vergleich zu Aussagen von Menschen mit den gleichen demografischen Merkmalen. Diese Merkmale wurden der Software als Eingabe zur Verfügung gestellt.[3] Stereotype Darstellungen könnten demnach beim Einsatz von *ChatGPT* in einer besonders zugespitzten Form fortgeschrieben werden und dadurch die Sichtbarkeit und Wahrnehmung polarisierender Meinungsbilder erhöhen.

Ein zunehmender Einsatz dieser Systeme führt dazu, dass wir bestimmte dominante Perspektiven automatisiert fortschreiben und verstärken. Darstellungen, die diesen Perspektiven nicht entsprechen, werden so noch seltener sichtbar. Das ist von besonderer Bedeutung, da die meisten komplexen KI-Systeme wiederum Weiterentwicklungen oder Anpassungen von einigen wenigen Basismodellen sind. Dadurch potenzieren sich die Mängel dieser Systeme. Neben bestimmten Darstellungen und Wirklichkeitsbildern können sich auch massenhaft enthaltene Verzerrungen verbreiten, die zu Diskriminierungen führen.[4]

Bislang fehlt ein Rahmen, damit Nutzer*innen und Verbraucher*innen die Ergebnisse komplexer KI-Modelle hinterfragen oder kritisch reflektieren können. Die Inhalte werden ohne weiteren Kontext wie Quellenangaben und Kennzeichnungen oder präzise Angaben über ihre Entstehung ausgegeben.

Unverstandene Systeme

Es gibt zahlreiche Gründe dafür, warum es zu folgenreichen Verzerrungen kommt, die zu Diskriminierungen führen. Sie stehen beispielsweise in Verbindung mit einer unausgewogenen Datenbasis, mit zugrunde liegenden Annahmen, die vorurteilsbehaftet sind, mit mangelhaften Prüfverfahren oder mit Fehlinterpretationen von Ergebnissen.[5] Im Zusammenhang mit Basismodellen kommen Besonderheiten in den Entwicklungsprozessen und Funktionsweisen der Modelle hinzu. Sie entstehen meist auf Grundlage massenhafter unstrukturierter und ungeprüfter Daten. Bisherige Ansätze, um die Datengrundlagen von KI-Systemen hinsichtlich folgenreicher Verzerrungen zu untersuchen, setzen voraus, dass genaue Kenntnisse über die Datenbasis vorliegen. Diese Ansätze greifen daher nicht. Das gilt zum Beispiel für, die die Repräsentanz in den Datensätzen im Zusammenhang mit dem gewünschten Ergebnis überprüfen.[6] Ähnliches trifft auf Erklärmethoden zu, die für kleinere KI-Modelle mit engen Anwendungskontexten entwickelt wurden.[7] Gründe dafür sind die ungelabelten Trainingsdaten, aber auch die Komplexität der Modelle an sich.[8] Hinzu kommt, dass erkannte Formen schädlicher Verzerrungen nicht systematisch dokumentiert werden.[9] Dadurch fehlt ein grundsätzliches Verständnis dafür, welche Schäden in diesen Zusammenhängen auf welche Weise entstehen können.[10]

Neben dem Training mit ungelabelten Daten, der Komplexität der Modelle und fehlenden Dokumentationen ist ein weiteres Problem, dass die Funktionsweise der Systeme bislang nicht einmal theoretisch komplett durchdrungen ist. Wir können das Vorgehen der Modelle also im Detail nicht nachvollziehen und erklären.[11] Diese Gemengelage erschwert es, neue aussagekräftige Erklärmethoden oder Ansätze der Nachvollziehbarkeit zu schaffen. Der Mangel an Informationen und Wissen über die Trainingsprozesse und die Datengrundlage führt dazu, dass kaum sichergestellt werden kann, dass keine personenbezogenen Daten verarbeitet und die Modelle datenschutzkonform entwickelt werden.[12] Nachvollziehbarkeit ist jedoch bei Modellen, die die Grundlage für eine Vielzahl von KI-Systemen und -Anwendungen bilden, von besonderer Bedeutung.[1] Die Notwendigkeit menschlicher Aufsicht, wie sie etwa die KI-Verordnungsentwürfe der *Europäischen Union* betonen, ist in diesem Zusammenhang schwer umsetzbar.[2]

„Warum brauchen wir Transparenz?“

Die zweite Ausgabe von Missing Link beschäftigt sich ausführlich mit dem Thema Transparenz, geeigneten Maßnahmen, um KI-Verfahren nachvollziehbarer zu gestalten, und mit den dafür notwendigen Kompetenzen.

Diese Ausgabe finden Sie hier: www.zvki.de > ZVKI Exklusiv > Fachinformationen.

Kein Modell ohne Menschen

Der Zugang zu Anwendungen, die auf Basismodellen aufbauen, ist nicht allen gleichermaßen möglich. Das hat verschiedene Ursachen. Eine ist darin begründet, dass die Systeme nicht für alle gleich gut funktionieren. Gängige Large Language Models sind beispielsweise nur auf einige wenige Sprachen ausgelegt. Darauf aufbauende Anwendungen und Geräte bleiben auch aus finanziellen Gründen für manche Menschen unzugänglich.[3] Daneben verfügen nicht alle über die notwendigen Kompetenzen, um entsprechende Anwendungen sinnvoll einzusetzen und ihre Ergebnisse einzuordnen. Erste Beobachtungen zur Nutzung von *ChatGPT* deuten darauf hin, dass bestehende Ungleichheiten bezüglich der Kompetenzen und Teilhabe eher verstärkt als abgebaut werden.[4]

Ähnliches zeigt sich auch im Zusammenhang mit Arbeitsprozessen. Mit Basismodellen wird häufig die Möglichkeit, Arbeitsprozesse effektiver und produktiver zu gestalten, in Verbindung gebracht. Auch wenn es gelingt, Arbeitsprozesse sinnvoll umzustrukturieren, werden davon nicht alle Betroffenen gleichermaßen profitieren. Das liegt unter anderem daran, dass verschiedene Aufgaben unterschiedlich stark vom Einsatz der Basismodelle betroffen sein können.[5] Darüber hinaus gilt auch hier, dass einige Systeme für manche Bevölkerungsgruppen und in einigen Arbeitsstrukturen nicht anwendbar sind. Forscher*innen des Projekts *Masakhane* stellten fest, dass Dialogsysteme wie *Cohere* oder *ChatGPT* für verbreitete afrikanische Sprachen wie Swahili oder Lingala nur unzureichend funktionieren, auch wenn sie zuvor mit entsprechenden Datensätzen angepasst wurden.[6] Das verringert die sozialen und ökonomischen Möglichkeiten im globalen Vergleich.[7]

Ungleichgewichte und Ungerechtigkeiten zeichnen sich noch deutlicher ab, wenn wir uns die Entwicklungsprozesse komplexer KI-Modelle mit breiten Anwendungsmöglichkeiten vor Augen führen.

Neben den großen Datenmengen und zahlreichen Parametern tragen zur Funktionsweise entsprechender KI-Systeme vor allem sehr große Mengen menschlicher Arbeitszeit bei, um die Modelle anzupassen und zu optimieren. Diese Arbeit wird meist an Menschen ausgelagert, die wenig verdienen sowie kaum Aus- und Weiterbildungsmöglichkeiten haben. Oft sind auch sie es, die von dem Einsatz der Systme nicht profitieren können.[8]

Kein Modell ohne Umwelt

Häufig sind die Bevölkerungsgruppen, die keinen Nutzen von den Basismodellen haben, auch diejenigen, die von den Umweltauswirkungen betroffen sind, die mit der Entwicklung und dem Einsatz dieser komplexen Modelle zusammenhängen. Die Autor*innen von „On the Dangers of Stochastic Parrots“ bringen dieses Problem bereits 2021 auf den Punkt: „Is it fair or just to ask, for example, that the residents of the Maldives (likely to be underwater by 2100) or the 800,000 people in Sudan affected by drastic floods pay the environmental price of training and deploying ever larger English LMs, when similar large-scale models aren’t being produced for Dhivehi or Sudanese Arabic?“[9]

Aufgrund der großen Trainingsdatenmengen und der Komplexität der Modelle sind sowohl die Trainingsprozesse als auch die Prozesse, um sie für bestimmte Anwendungen anzupassen, energie- und ressourcenintensiv.[10] Insgesamt ist ein deutlicher Trend zu erkennen, dass die Modelle zunehmend größer und rechenintensiver statt nachhaltiger werden. Das wirkt sich nicht nur auf Energie- und Ressourcenbedarfe wie Wasser oder Seltene Erden aus, sondern auch auf den Ausstoß von Treibhausgasen. Zu weiteren Anstiegen tragen außerdem unreflektierte Angebote und Nutzungsformen bei: Der Hype um breite KI-Modelle und -Anwendungen ist groß – die Sorge, abgehängt zu werden, ist es ebenfalls. Daher ziehen es zahlreiche Software-Anbieter vor, entsprechende Modelle auf irgendeine Weise einzubinden, ohne dabei die Sinnhaftigkeit dessen sicherzustellen.[11]

Autorin

Jaana Müller-Brehm

1 Vgl. Bender et al., S. 617.
2 Vgl. Human-Centred Artificial Intelligence (HAI), S. 58.
3 Vgl. Bisbee et al., S. 4.
4 Vgl. Bommasani et al., S. 131.
5 Vgl. Müller-Brehm, S. 18–26.
6 Vgl. Zhang, S. 2.
7 Vgl. Tang et al., o. S.
8 Vgl. ebd.
9 Vgl. ebd.
10 Vgl. Dev et al., S. 254.
11 Vgl. Albrecht, S. 43.
12 Vgl. ebd., S. 83.

1 Vgl. Bommasani et al., S. 156.
2 Vgl. OECD, S. 14 f.
3 Vgl. Bender et al., S. 613.
4 Vgl. Albrecht, S. 76.
5 Vgl. Bommasani et al., S. 151f.
6 Vgl. Ojo/ Ogueji, S. 1 ff.
7 Vgl. ebd.
8 Vgl. Albrecht, S. 85.
9 Bender et al., S. 612 f.
10 Vgl. Bommasani et al., S. 140.
11 Vgl. Ananthaswamy, S. 203 ff.

Warum die Regulierung bislang nicht genügt

Als Anfang 2023 KI-generierte Bilder des Papsts in weiß-goldenem Hiphop-Outfit kursierten[1] und an Schulen und Universitäten über den Umgang mit KI-erzeugten Textarbeiten diskutiert wurde,[2] standen auch die Pläne einer europäischen Regulierung von KI-Systemen auf dem Prüfstand. Mit seinem Vorschlag zur KI-Verordnung bessert das *Europäische Parlament* an entscheidenden Stellen nach und lässt doch viele Fragen unberücksichtigt.

Die *Europäische Kommission* legte im April 2021 einen KI-Verordnungsentwurf vor, der generative KI-Systeme oder Foundation Models nicht explizit benannte. Selbst die vorgesehene Kennzeichnungspflicht von KI-Systemen, die mit Menschen direkt interagieren, war nicht auf die neuen Modelle zugeschnitten.[3]

Vor dem Hintergrund der rasanten Weiterentwicklungen von KI-Modellen fordern immer mehr Expert*innen, die mit der Technik einhergehenden Risiken als solche wahrzunehmen und ihnen mit einem präzisierten Pflichtenkatalog zu begegnen.[4] So sprechen sich etwa die Autor*innen des Papers „Nachhaltigkeitskriterien für Künstliche Intelligenz“ dafür aus, nachhaltigkeitsbezogene Auswirkungen der immer größer werdender KI-Modelle bei ihrer Bewertung zu berücksichtigen.[5] Wissenschaftler*innen wie Rishi Bommasani und Andreas Jungherr mahnen zudem an, dass KI-vermittelte Diskriminierungen bei der Entwicklung und im Einsatz bedacht werden müssen.[6] Aus rechtswissenschaftlicher Sicht ist zu beachten, dass sich die Risiken von Basismodellen erst dann final bestimmen lassen, wenn sich ihr konkretes Einsatzfeld abzeichnet.[7]

Wirtschaftsnahe Interessenvertreter*innen wie der Rechtsanwalt und IT-Berater Andreas Splittgerber entgegnen, dass Basismodelle schon jetzt gesetzlichen Anforderungen unterliegen.[8] Online-Plattformen sind etwa durch den *Digital Services Act*[9] zur Löschung gemeldeter und rechtswidriger Inhalte verpflichtet, und zwar auch dann, wenn diese KI-generiert sind.[10] Die von der *Datenschutzgrundverordnung*[11] und künftig dem *Data Act*[12] gestellten Anforderungen an den Umgang mit Daten und ihr Management würden ebenfalls unverändert auch für komplexe KI-Systeme gelten. Keine der Regelungen ist jedoch auf die spezifischen Herausforderungen von Foundation Models ausgerichtet.

Parlamentsentwurf als Schritt in die richtige Richtung?

Der am 14. Juni 2023 vom *Europäischen Parlament* beschlossene Entwurf der KI-Verordnung adressiert gezielt einige Risiken, die mit Basismodellen in Verbindung stehen.[13] Zum ersten Mal in der europäischen Digitalregulierung werden Mindestanforderungen formuliert, die speziell auf Foundation Models und die auf ihnen basierenden KI-Systeme zugeschnitten sind.[14]

So sollen Hersteller*innen für die Entwicklung und den Betrieb ihrer Basismodelle beispielsweise nur qualitativ geeignete Datensätze verarbeiten und einbeziehen dürfen: Die verwendeten Daten müssen auf ihre grundsätzliche Eignung sowie auf bestehende Verzerrungen überprüft werden. Bestehende Verzerrungen oder Einschränkungen in der Nutzbarkeit müssen durch angemessene technische Maßnahmen ausgeglichen werden.[15] Gleichzeitig sollen die Hersteller*innen Foundation Models so konzipieren, dass es möglich ist, ihren Energie- und Ressourcenverbrauch sowie die hervorgerufenen Umweltauswirkungen - soweit technisch möglich - während ihres gesamten Lebenszyklus zu protokollieren. Auch sollen Basismodelle so (weiter-)entwickelt werden, dass ihr Ressourcen- und Energieverbrauch verringert und ihre Energieeffizienz gesteigert wird.[16] Für die mangelnde Präzision dieser Anforderungen erntet der Entwurf jedoch Kritik, beispielsweise von Jonas Andrulis, Geschäftsführer des Deutschen KI-Unternehmens *Aleph Alpha*: Die Anforderungen an Data Governance und Risikoverringerung, beispielsweise die Bewertung von Verzerrungen in Datenquellen, seien zu unpräzise gefasst und überforderten die Hersteller*innen und Anwender*innen von Foundation Models.[17] Es bleibt abzuwarten, wie genau diese Anforderungen von den europäischen Standardisierungsorganisationen *CEN* und *CENELEC* formuliert werden. Die Verpflichtungen zur Reduktion des Energie- und Ressourcenverbrauchs nach Vorstellung des *Europäischen Parlaments* stehen unter dem Vorbehalt, dass entsprechende Standards für die Praxis erarbeitet werden. Die Entwicklung dieser Standards ist nicht Teil des Auftrags der *Europäischen Kommission*.

„Welcome to Trilogue - Worüber jetzt noch diskutiert wird“

Mit dem Beschluss des Europäischen Parlaments vom 14. Juni 2023 ist die letzte Verhandlungsetappe auf dem Weg zu einem „europäischen KI-Gesetz“ erreicht. Welche Streitthemen bis Ende des Jahres im Mittelpunkt der politischen Debatte stehen werden, zeigen wir Ihnen in der vierten Ausgabe unseres Briefings zur KI-Verordnung.

Die Ausgabe zum Trilog finden Sie hier: Startseite > ZVKI Exklusiv > Fachinformationen.

Dort stehen auch alle weiteren Ausgaben des Briefings zur KI-Verordnung für Sie bereit.

1 Vgl. Klaus, o. S.
2 Vgl. Wilke, o. S.; vgl. Weßels, S. 3 ff.
3 Vgl. Artikel 52, Abs. 1 Kommissions-KIVO.
4 Vgl. Gebru et al., S. 4; vgl. Maham/ Küspert, S. 43 f.
5 Vgl. Rohde et al., S. 43 ff und S. 68 f.
6 Vgl. Jungherr, o. S.; vgl. Bommasani et al., S. 130-135; vgl. Müller-Brehm, S. 18-22.
7 Vgl. Helberger/ Diakopoulos, S. 3 f.; vgl. Hacker et al., S. 1114-1117.
8 Vgl. Kroker, o. S.
9 Vgl. Koch et al., o. S.
10 Vgl. epd/ dpa, o. S.
11 Vgl. Schmierer, o. S.; aus US-amerikanischer Perspektive vgl. Bommasani et al., S. 146 ff.
12 Vgl. acatech, o. S.; D16, S. 3.
13 Vgl. Erwägungsgrund 60g, Parlaments-KIVO.
14 Vgl. Artikel 28b, Parlaments-KIVO.
15 Vgl. Artikel 28b, Absatz 2 lit. b), Parlaments-KIVO.
16 Vgl. Artikel 28b, Absatz 2 lit. d), Parlaments-KIVO.
17 Vgl. Bienert et al., S. 9.

Nach dem Entwurf des Parlaments soll ein Foundation Model, das aufgrund seiner besonderen Einsatzweise oder Funktionalität ein signifikantes Risiko für die Gesundheit, die Sicherheit, die Grundrechte von Menschen oder die Umwelt darstellt, als Hochrisiko-KI-System behandelt werden. Als solches muss es zusätzliche Pflichten erfüllen.[1]

Weil diese Risikoeinordnung bei Systemen mit einer Vielzahl an Einsatzmöglichkeiten für die ursprünglichen Hersteller*innen schwerfällt, schlägt das Parlament vor, die Verantwortlichkeiten auszuweiten:[2] Unternehmen, die Foundation Models erwerben und zu eigenen, spezifischen Zwecken anpassen, sollen wie Hersteller*innen behandelt werden, wenn das Foundation Model substanziell weiterentwickelt wird und durch diese Veränderung ein signifikantes Risiko für Mensch oder Umwelt entsteht.[3] In diesen Fällen müssen Hersteller*innen und Weiterentwickler*innen Informationen über die Funktionsweise des KI-Modells und die eingesetzten (Trainings-)Daten austauschen.[4] Eine ähnliche Pflicht hatte auch schon der *Rat der Europäischen Union* vorgeschlagen.[5] Außerdem sind Weiterentwickler*innen nach Vorstellung des *EU-Parlaments* stets dazu angehalten, den von ihnen geplanten Einsatz von Basismodellen einer Risikoanalyse zu unterziehen.[6] Sie sollen die vom KI-Einsatz betroffenen Gruppen identifizieren und die vorhersehbaren Auswirkungen auf Grundrechte und Umwelt dokumentieren. Werden Nachteile und Risiken identifiziert, sollen die verantwortlichen Unternehmen durch einen detaillierten Plan Abhilfe schaffen. Andernfalls darf das Foundation Model in der geplanten Weise nicht eingesetzt werden.[7] Im Einzelnen bleibt bislang vieles unklar, etwa, welche Abhilfemaßnahmen in Betracht kommen oder ab welchem Umfang die vorgenommenen Anpassungen der Basismodelle als „signifikant“ gelten.[8] Auch in diesem Punkt könnten entsprechend formulierte Standards für mehr Klarheit sorgen. Ohne solche Regeln kann für Unternehmen ein wirtschaftliches Risiko entstehen: Kommen die Aufsichts- und Marktüberwachungsbehörden zu einer anderen Einschätzung als das Unternehmen (beispielsweise darüber, ob die vorgenommenen Abhilfemaßnahmen ausreichend sind), drohen diesem hohe Bußgelder.[9] Ohne präzise

1 Vgl. Artikel 6, Absatz 2, Parlaments-KIVO.
2 Vgl. Hacker et al., S. 1115-1117.
3 Vgl. Artikel 28, Absatz 1 lit. ba), Parlaments-KIVO.
4 Vgl. Artikel 28, Absatz 2, 2a und 2b, Parlaments-KIVO; vgl. Hacker et al., S. 1114-1117; vgl. Helberger/ Diakopoulos, S. 3 ff .
5 Vgl. Artikel 4b, Absatz 5, Rats-KIVO.
6 Vgl. Artikel 29a, Absatz 1, Parlaments-KIVO. Die Pflicht besteht nur beim Einsatz eines Hochrisiko-Systems. Ob im konkreten Fall ein Hochrisiko-System vorliegt, erfahren Weiterentwickler*innen aber erst durch die Risikoanalyse.
7 Vgl. Artikel 29a, Absatz 2, Parlaments-KIVO.
8 Vgl. Bienert, S. 18.
9 Vgl. Artikel 71, Kommissions-KIVO.

Standards – oder eventuelle Spezifikationen durch die *Europäische Kommission* – ist Rechtsklarheit erst nach Jahren der Aufsichts- und Gerichtspraxis zu erwarten.

Ist damit alles gut?

Ob sich das Parlament mit seinen Forderungen im Trilog gegenüber der *Europäischen Kommission* und dem *Europäischen Rat* durchsetzen wird, ist noch offen.[1] Einiger mit Foundation Models verbundenen Herausforderungen wird sich die *Europäische Union* aber selbst dann noch immer nicht angenommen haben.

Wiederholt kritisieren Expert*innen wie die Rechtswissenschaftlerin Natali Helberger und der Kommunikationswissenschaftler Nicholas Diakopoulos, dass die Kontroll- und Rechenschaftspflichten des *Digital Services Act* in Bezug auf Deep Fakes und Desinformation nicht auf die Hersteller*innen und die Weiterentwickler*innen von Basismodellen ausgeweitet werden.[2] Die im Parlamentsentwurf vorgesehene Pflicht, dass Erzeugnisse von generativer KI rechtskonform sein müssen,[3] genüge nicht, die von der Technik ausgehenden Gefahren für die öffentliche Meinungsbildung zu verringern.[4] Zugleich sind in der KI-Verordnung anders als in den übrigen Gesetzen der Daten- und Digitalstrategie keine Maßnahmen vorgesehen, um neu entstehenden oder sich verstärkenden Marktmächten zu begegnen.[5] Stattdessen scheint die vorgesehene Förderung von kleinen und mittleren Unternehmen vor allem die Schaffung europäischer Champions im Blick zu haben.[6] Das wären solche Unternehmen, die sich aufgrund ihrer eigenen starken Marktposition auch im Wettbewerb mit US-amerikanischen oder chinesischen Konkurrenten behaupten können.[7]

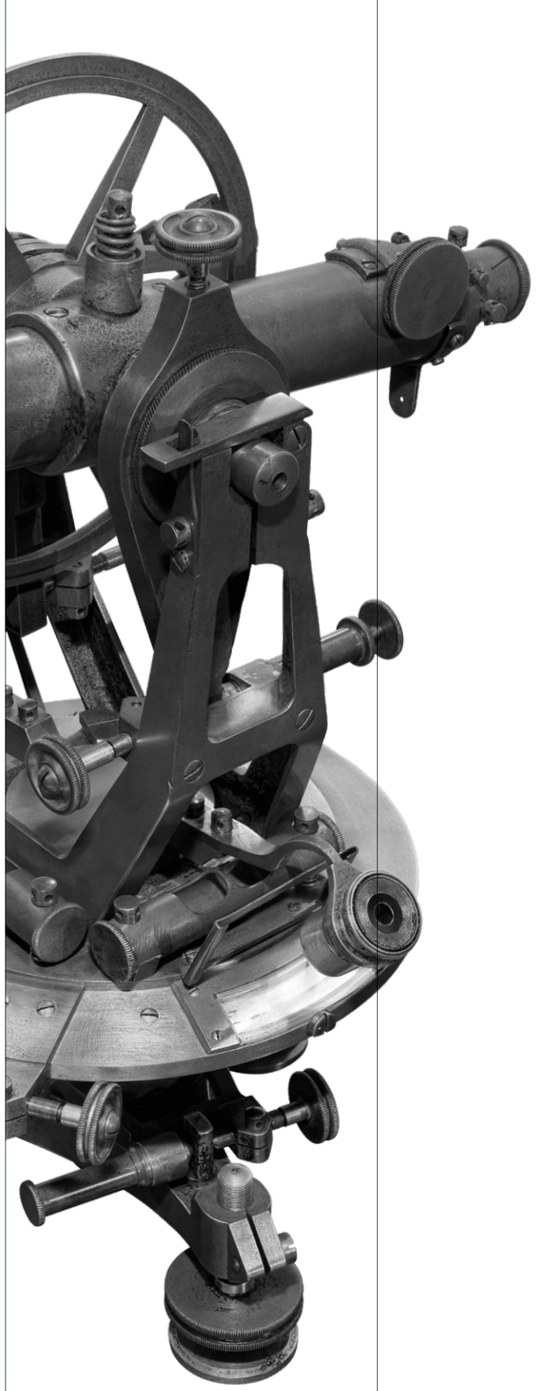
Die bestehenden und geplanten Gesetze begrenzen die Risiken von Foundation Models daher nicht ausreichend und lassen bislang viele Fragen unbeantwortet.

„Diskriminierung begegnen, Fairness stärken“

Die dritte Ausgabe von Missing Link beschäftigt sich ausführlich mit dem Thema KI- und algorithmenvermittelter Diskriminierung und erklärt, wie sie zustande kommen. Außerdem gibt das Magazin einen Überblick über mögliche Begegnungsansätze.

Die Ausgabe zum Trilog finden Sie hier: Startseite > ZVKI Exklusiv > Fachinformationen.

Diese Ausgabe finden Sie hier: www.zvki.de > ZVKI Exklusiv > Fachinformationen.



Matthieu Binder und Gergana Baeva

Autor*innen

1 Vgl. Binder, S. 9.
2 Vgl. Hacker, o. S.; Vgl. Helberger/ Diakopoulos, S. 4.
3 Zur Abwägung von Persönlichkeitsrecht, Urheberrecht und Meinungsfreiheit im bestehenden Recht vgl. Lantwin 2019 und Lantwin 2020.
4 Zu diesen Risiken und Regelungsbedarf vgl. Lossau, S. 5 ff.
5 Vgl. Bommasani et al., S. 151.
6 Vgl. AFP, France 24, o. S.; vgl. Maham/ Küspert, S. 44.
7 Vgl. Valero, o. S.; vgl. Stam, o. S.

NACHHALTIGE BASIS FÜR KI

Es braucht maßgeschneiderte Ansätze, um regulatorische Ideen für nachhaltigere Basismodelle zu konkretisieren und um darüber hinausgehende Maßnahmen zu verwirklichen:[1] Dazu zählen umfangreiche Dokumentationen, Kosten-Nutzen-Analysen, schlankere KI-Modelle und die Förderung bestimmter Kompetenzen.

Umfangreiche Dokumentationen als Grundlage

Bestimmte Kriterien vertrauenswürdiger KI[2] sind als übergreifende Begegnungsstrategien für nachhaltigere Basismodelle besonders bedeutsam. Dazu zählen etwa Transparenz, Rechenschaftspflicht und Fairness. Diese Ansätze gründen auf umfangreichen Dokumentationen der Trainingsdaten von komplexen KI-Modellen. Dokumentationen vereinfachen es Entwickler*innen beispielsweise, folgenreiche Verzerrungen zu erkennen, um ihnen im Anschluss begegnen zu können.[3]
Eine Herausforderung hinsichtlich der Dokumentation von Trainingsdaten bei Foundation Models besteht darin, dass die zugrunde liegenden Datenmengen sehr umfangreich sind. [4] Baut die Entwicklung zusätzlich auf unstrukturierten Datensätzen auf, ist eine umfassende Dokumentation kaum möglich.[5] Demnach ist es notwendig, den Trainingsprozess so zu gestalten, dass strukturierte Datensätze genutzt werden.[6] Die Autor*innen des Papers „On the Dangers of Stochastic Parrots” schlagen zudem vor, nur so viele Daten zu sammeln, wie auch kuratiert und dokumentiert werden können.[7] Dies bedeutet, dass Entwickler*innen die Datengrundlagen der entsprechenden Basismodelle im Zweifel verkleinern müssen. Dokumentationen bilden für viele Nachhaltigkeitsmaßnahmen über den gesamten KI-Lebenszyklus hinweg eine zentrale Grundlage: Mit ihrer Hilfe können bestimmte Nachhaltigkeitsindikatoren – etwa zum Energieverbrauch oder CO2-Ausstoß, aber auch zu partizipativen Designansätzen oder zum Datenschutz – erfasst werden – vom Auftrag über die Entwicklung bis hin zum Einsatz und der Evaluierung eines KI-Modells. Solche Dokumentationen ermöglichen Analysen und im nächsten Schritt gezielte Maßnahmen, um die Nachhaltigkeit eines KI-Modells zu verbessern.[8]

Für eine erfolgreiche Umsetzung von Nachhaltigkeitsansätzen gilt es, die entsprechenden Indikatoren bereits vor der Entwicklung eines Basismodells zu berücksichtigen. Beispielsweise sollten Auftraggeber*innen, Entwickler*innen und Wissenschaftler*innen die Umweltkosten komplexer KI-Modelle und die Frage, wie diese verringert werden können, von Anfang an mitdenken.[9] Im Zusammenhang mit Foundation Models kann das bedeuten, dass auf unstrukturierte und sehr umfangreiche Datensätze verzichtet werden muss.

Schlankere KI-Modelle und Kosten-Nutzen-Analysen

Bevor eine KI-Anwendung programmiert wird, sollten Auftraggeber*innen die Vor- und Nachteile abwägen, die die Entwicklung eines Basismodells für Einzelne, die Gesellschaft und die Umwelt hätte.[10] Eine solche Kosten-Nutzen-Rechnung stellen die Autor*innen von „On the Opportunities and Risks of Foundation Models“ vor.[11] Zu den sozialen sowie ökologischen Vorteilen gehört beispielsweise, dass ein Foundation Model die Gesundheitsversorgung verbessern oder dem Klimawandel entgegenwirken kann. Diese Vorteile werden mit den sozialen sowie ökologischen Kosten in Bezug gesetzt, zum Beispiel den Energiekosten des Basismodells oder den sozialen Kosten der Schadstoffemissionen. Nur wenn die Entwicklung und der Einsatz eines komplexen KI-Modells mehr Nutzen als Kosten haben, sollten Unternehmen Foundation Models in Auftrag geben.[12] Die hier aufgestellte Rechnung umfasst allerdings nur bestimmte soziale und ökologische Vor- und Nachteile und vernachlässigt daher einige weitere Nachhaltigkeitskriterien.

Eine konsequente Kosten-Nutzen-Rechnung lässt sich auch auf andere Risiken und Herausforderungen bezüglich der Nachhaltigkeit von Basismodellen übertragen: Die allem zugrunde liegende Begegnungsstrategie umfasst an erster Stelle die Entscheidung darüber, ob ein Foundation Model überhaupt die richtige Wahl für einen bestimmten Anwendungskontext ist oder ob die sozialen, kulturellen, ökonomischen und ökologischen Kosten dafür zu hoch wären. Hierbei geht es nicht nur um die Frage, ob ein Modell entwickelt werden sollte oder nicht, sondern auch darum, unter welchen Bedingungen es gebaut werden sollte und welche Kriterien und Grundsätze hierfür gelten.[13] Diese Fragen müssen politische Entscheidungsträger*innen, zivilgesellschaftliche Organisationen, Bürger*innen und Unternehmen gemeinsam beantworten.

1 Vgl. Bommasani et al., S. 146.
2 Vgl. High-Level Expert Group on AI, S. 17 f.
3 Vgl. Bender et al., S. 615.
4 Vgl. ebd.
5 Vgl. Tang et al., o. S.
6 Vgl. ebd.
7 Vgl. Bender et al., S. 615.
8 Vgl. Rohde et al., S. 22.
9 Vgl. Schulzki-Haddouti, o. S.
10 Vgl. Bommasani et al., S. 140.
11 Vgl. ebd., S. 140 ff.
12 Vgl. ebd., S. 140.
13 Vgl. ebd., S. 159.

Neue Perspektiven fördern und Umweltschäden reduzieren

Vor allem für die Herausforderungen von Machtoligopolen gibt es keine offensichtlichen oder einfachen Lösungen. So können Foundation Models laut der Autor*innen von „On the Opportunities and Risks of Foundation Models“ einerseits weiterhin zu einer Machtkonzentration weniger großer Unternehmen beitragen. Andererseits könnten Basismodelle aber auch die Grundlage für eine deutlich vereinfachte Anpassung und Implementierung anwendungsbezogener KI-Systeme bilden. Dadurch bestünde die Möglichkeit, ein breiteres Spektrum von Anwendungen und gleichmäßiger verteilten Wohlstand zu ermöglichen.[1] Entsprechend beschreiben einige wissenschaftliche Veröffentlichungen umfassende Open-Source-Ansätze als eine mögliche Maßnahme, um globalen Machtungleichheiten, Oligopolstellungen weniger Unternehmen und der Dominanz hegemonialer Perspektiven etwas entgegenzusetzen.[2]

Einige Beispiele zeigen, dass die Veröffentlichung von Quellcodes die Marktmacht bereits erfolgreicher Unternehmen sogar stärken kann. Ein geleaktes Dokument eines *Google*-Mitarbeiters beschreibt beispielsweise, wie *Meta* von der Offenlegung des KI-Modells *LLaMA* zur Sprachverarbeitung profitierte.[3] So können offene Modelle beispielsweise kostengünstig durch die Open-Source-Community auf dem neuesten Stand gehalten werden, etwa wenn bessere Datensätze verfügbar sind.[4] Das geleakte Dokument zeigt einerseits, dass Open-Source-Ansätze bei großen Technikunternehmen ankommen, und andererseits, dass Unternehmen daraus resultierende Weiterentwicklungen nutzen, um Profit zu erwirtschaften: Wenn Entwickler*innen die veröffentlichten Modelle kostenlos optimierten, könne *Meta* diese Verbesserungen in weitere – nicht öffentlich zugängliche – KI-Anwendungen einfließen lassen.[5] Ähnlich äußert sich Emad Mostaque, der CEO von *Stability AI*, dem Unternehmen hinter *Stable Diffusion*:[6] Es plant, die Innovationen, die unabhängige Entwickler*innen durch die

Nutzung und Anpassung ihrer Modelle einbringen, in maßgeschneiderte Produkte für verschiedene Kund*innen einfließen zu lassen.[7] Große Technikunternehmen können sich demnach offene Modelle finanziell zunutze machen, ohne dass der Zugang zu den aus ihnen resultierenden Anwendungen offen gestaltet ist. In diesem Zusammenhang stellt sich die Frage, wie Open-Source-Ansätze gelingen können, von denen alle beteiligten Akteur*innen profitieren.

Damit Open-Source-Ansätze perspektivisch ein Gegengewicht zu Machtkonzentrationen bilden können, genügt eine alleinige Veröffentlichung des Quellcodes nicht. Weitere Maßnahmen sind erforderlich. Dazu zählen unter anderem eine partizipative Entwicklung der Basismodelle, die Veröffentlichung der Datensätze sowie die Förderung von Kompetenz- und Wissensaustausch. Beispiele hierfür sind Organisationen wie *Masakhane*[8] oder *EleutherAI*[9] und das Unternehmen *HuggingFace*[10]. Sie treiben die unabhängige Forschung und Entwicklung von offenen KI-Modellen voran.[11] Allerdings sind weitere Forschungsarbeiten notwendig, um besser beurteilen zu können, ob und wie umfassende Open-Source-Ansätze tatsächlich zum Aufbrechen von hegemonialen Perspektiven und Machtoligopolen beitragen können.

Open-Source-Ansätze können auch dazu beitragen, schädliche Umweltauswirkungen von Basismodellen zu reduzieren. So können Entwickler*innen oder Forscher*innen beispielsweise bereits vortrainierte Modelle durch Feinjustierungen an spezifische Anwendungskontexte anpassen und müssen diese nicht von Grund auf neu trainieren[12] – ein Vorteil, den Sasha Luccioni, Forscherin bei *Hugging Face,* betont.[13]

Zudem arbeiten Wissenschaftler*innen derzeit daran, wie KI-Systeme bei – womöglich besserer Leistungsfähigkeit – schlanker und energieeffizienter programmiert werden können.[14] Dabei helfen zum Beispiel Komprimierungsverfahren für bereits vortrainierte KI-Modelle, die öffentlich zugänglich sind, um diese in ihrer Größe zu reduzieren.[15] Eine weitere Möglichkeit ist, effiziente Modelle zu nutzen, „die mit weniger Parametern die gleiche Leistung erbringen.“[16]

Kompetenzen stärken

Damit Bürger*innen und Verbraucher*innen KI-Systeme, die auf Basismodellen aufbauen, reflektiert nutzen und mit ihren Ergebnissen umgehen können, sind neben Transparenz- und Prüfmaßnahmen auch entsprechende Kompetenzen notwendig. Denn generative Basismodelle können für Desinformationskampagnen genutzt werden, um in großer Anzahl und mit wenig Aufwand beispielsweise synthetische Bilder, Videos und Texte zu erstellen. Medienkompetenzen und sogenannte AI Literacy der Rezipient*innen sind daher unabdingbar. AI Literacy beschreibt eine Sammlung von Fähigkeiten, die es ermöglichen, dass Nutzer*innen KI-Systeme selbstbestimmt anwenden, technische Grundlagen verstehen und ethische Herausforderungen reflektieren.[1] In Bezug auf komplexe Basismodelle sind vor allem die Einordnung und die Bewertung KI-generierter Inhalte relevant.

Beim Aufbau von AI Literacy können sowohl grundlegende Medienkompetenztrainings als auch neuere Methoden, die Deepfakes in den Fokus stellen, hilfreich sein.[2] Da das Forschungsfeld zu Medienkompetenzen, insbesondere im Kontext von Basismodellen und KI-generierten Inhalten noch recht neu ist, müssen Wissenschaftler*innen genauer untersuchen, welche Programme zum Aufbau der notwendigen Kompetenzen effektiv beitragen können.[3]

Allerdings reicht es nicht aus, nur die individuellen Kompetenzen zu fördern.[4] Die Verantwortung, reflektiert mit komplexen KI-Systemen und ihren Ergebnissen wie KI-erzeugten Inhalten umzugehen, sollte nicht allein an Bürger*innen und Verbraucher*innen ausgelagert werden. Stattdessen müssen zivilgesellschaftliche und politische Akteur*innen eine Auseinandersetzung über die Bewertung und Nutzung komplexer KI-Modelle vorantreiben.[5]

Autorin

Franziska Busse

1 Vgl. ebd., S. 149.
2 Vgl. ebd., S. 151.
3 Vgl. Patel/ Ahmad, o. S.
4 Vgl. ebd., o. S.
5 Vgl. ebd., o. S.
6 Vgl. Heaven, o. S.
7 Vgl. ebd., o. S.
8 Vgl. Masakhane, o. S.
9 Vgl. EleutherAI, o. S.
10 Vgl. Hugging Face, o. S.
11 Vgl. Bommasani, S. 151.
12 Vgl. Luccioni, o. S.
13 Vgl. AlgorithmWatch 2023 b, S. 16.
14 Vgl. Ananthaswamy, S. 205.
15 Vgl. Albrecht, S. 47.
16 Vgl. ebd.

1 Vgl. Ng et al., S. 2.
2 Vgl. Hwang et al., S. 191.
3 Vgl. Helmus, S. 16.
4 Vgl. McCosker, S. 14.
5 Vgl. ebd.

VERARBEITEN

MIT OPEN SOURCE DIE ÖKONOMISCHE TEILHABE STÄRKEN

Wenige Akteure verfügen über die Ressourcen, die Infrastruktur und das Wissen, um Basismodelle zu entwickeln. Viele davon teilen ihre Modelle und wichtige zusätzliche Informationen nicht. Die Forschungsgruppe *EleutherAI* stellt eigene *GPT*-Varianten und Datensätze mit umfangreichen Dokumentationen zur Verfügung. Das ermöglicht das Erforschen von Mängeln und Gefahren beim Einsatz von Basismodellen.

Wer und wann?

EleutherAI ist eine Forschungsorganisation, die sich im Juli 2020 unter dem Namen *LibreAI* auf einem *Discord*-Server, einer selbstmoderierten Plattform eines Onlinedienstleisters, zusammengeschlossen hat. Das Ziel war, Entwicklungen im Bereich der Künstlichen Intelligenz und des maschinellen Lernens in einer Gemeinschaft von Forscher*innen zu diskutieren. Auslöser waren die Veröffentlichungen der *GPT*-Modelle des US-Unternehmens *OpenAI*. Im Januar 2023 wurde *EleutherAI* auch formell als nicht profitorientiertes Forschungsinstitut gegründet.[1]

Was?

EleutherAI verfolgt einen dezentralen Open-Source-Ansatz in der Forschung. Die Organisation fördert gemeinschaftsorientierte Beiträge, bei denen Forscher*innen und Entwickler*innen mit unterschiedlichen fachlichen Hintergründen gemeinsam an Projekten arbeiten. Das Institut stellt seine Quellcodes, Datensätze und Forschungsergebnisse offen zur Verfügung, sodass die Gemeinschaft und alle Interessierten auf ihre Arbeit zugreifen, sie prüfen und darauf aufbauen können. Dieser offene Ansatz soll die Transparenz sowie die Zusammenarbeit verschiedener Forscher*innen fördern und den Fortschritt in diesem Bereich beschleunigen.

Wie?

Im Dezember 2020 veröffentlichte *EleutherAI* „The Pile“, eine Sammlung von Datensätzen zum Training und Testen großer Sprachmodelle.[2] Die Sammlung umfasst verschiedene Wissensgebiete und -formate wie Bücher, Quellcodes, Webseiten, Chat-Protokolle sowie Aufsätze zu Medizin, Physik, Mathematik und Philosophie. Mithilfe der Datensätze können Entwickler*innen große Sprachmodelle mit Daten aus verschiedenen Bereichen trainieren oder prüfen, wie geeignet diese sind, um eine allgemeine, nicht auf einen Wissensbereich beschränkte Textmodellierung auszuführen.

Wenig später, im März 2021, veröffentlichte die Gruppe mit den *GPT-Neo*-Modellen erste kleinere Basismodelle zur Textgenerierung.[3] Die umfangreichste Version umfasst 2,7 Milliarden Parameter. Im Juni 2021 folgte mit *GPT-J-6B* das größte *GPT-3* ähnliche Modell mit 6 Milliarden Parametern. Im Februar 2022 wurde *GPT-NeoX* veröffentlicht, das mit 20 Milliarden Parametern das größte Open-Source-Modell zu diesem Zeitpunkt war.[4] Die Gruppe arbeitet nicht an noch größeren Modellen, da die bisherigen Modelle für die vorgesehenen Zwecke ausreichen.[5]

Bei anderen Systemen, die mit solch umfangreichen Parametern funktionieren, handelt es sich um Entwicklungen von Big-Tech-Unternehmen, von for-profit-Forscher*innen, die auf private Rechenzentren zugreifen können, und von einigen chinesischen Universitäten mit Zugriff auf staatliche Supercomputer und enger staatlicher Bindung, so die *EleutherAI*-CEO Stella Biderman.[6]

Anfänglich nutzte die Gruppe *Googles TPU Research Cloud Program*, eine Plattform, die kostenfrei zur Forschung an Machine-Learning-Systemen verwendet werden kann. Im Zuge der Nutzung verpflichtet sich die Organisation zur Veröffentlichung der Ergebnisse und Weitergabe bestimmter Informationen an *Google*.[7] Nach ersten Erfolgen gewann *EleutherAI* verschiedene Sponsor*innen und Spender*innen. *CoreWeave*, ein in den Vereinigten Staaten ansässiges Cryptomining-Unternehmen, unterstützt das Institut mit Cloudservices, um ihre Anwendungen zu hosten, und stellt Rechenleistung zum Training zur Verfügung.[8] Im Gegenzug sind die *EleutherAI*-Anwendungen für Kund*innen von *CoreWeave* direkt nutzbar.[9] Darüber hinaus wird die Gruppe durch Rechenleistung sowie Sponsoring von den KI-Unternehmen *Hugging Face* und *Stability AI*, dem ehemaligen *GitHub*-CEO Nat Friedman sowie den Unternehmen *Lamda Labs* und *Canva* unterstützt.[10]

Im März dieses Jahres gehören 20 Vollzeitkräfte in der Forschung zum Team, das aber offen für Kollaborator*innen bleibt und zum niedrigschwelligen Austausch weiterhin den *Discord*-Server betreibt.[11]

1 Vgl. Biderman et al. 2023, o. S.
2 Vgl. Gao et al.
3 Vgl. Biderman et al. 2021, o. S.
4 Vgl. Black et al., o. S.
5 Vgl. Leahy, o. S.
6 Vgl. Biewald, o. S.
7 Vgl. Google, o. S.
8 Vgl. Black et al., o. S.
9 Vgl. Wiggers, o. S.
10 Vgl. ebd.
11 Vgl. Biderman et al. 2023, o. S.

Ist die Forschung wirklich frei?

EleutherAI ist auf Finanzierungen aus der Privatwirtschaft angewiesen und kann etwa auf die Rechenleistung von *Stability AI* zählen, nachdem die Forschungsorganisation das Unternehmen bei der ersten Version von *Stable Diffusion* unterstützt hatte. Kritiker*innen befürchten, dass die verschiedenen Sponsoren Einfluss auf die Ausrichtung des Forschungsinstituts und die von ihm formulierten politischen Empfehlungen nehmen könnten.[1]

Stella Biderman argumentiert, dass *EleutherAI* durch das breite Portfolio an Investor*innen bestmöglich abgesichert ist: Sich nur auf ein Big-Tech-Unternehmen zu stützen, würde eine größere Abhängigkeit mit sich bringen.[2] Die Frage bleibt, ob das Forschungsinstitut diesen Weg beibehalten kann und in welchem Ausmaß seine Arbeit zu einem Gegengewicht zur Marktmacht großer Technikunternehmen beitragen kann.

		Autor
	Paul Ritzka	

1 Vgl. ebd., o. S.
2 Vgl. ebd., o. S.

NACHFRAGEN

WOFÜR WOLLEN WIR UNSERE RESSOURCEN NUTZEN?

Fortschritt in der KI-Entwicklung heißt gerade vor allem eines: immer größere Modelle. Nachhaltigkeit spielt bislang eine untergeordnete Rolle. Wenn wir einen positiven Nutzen aus dem Einsatz von Basismodellen ziehen wollen, muss sich das ändern, sagt Friederike Rohde. Dazu gehören auch eine gleichberechtigte politische Teilhabe von zivilgesellschaftlichen Organisationen, ein diversifizierter Markt und ein verantwortungsvoller Umgang mit unseren Ressourcen.

Welche Herausforderungen kommen Ihnen als erste in den Sinn, wenn Sie an die Nachhaltigkeit von Foundation Models und darauf aufbauenden Anwendungen denken?

Wenn komplexe und immer größere KI-Systeme in zunehmend mehr Bereichen eingesetzt werden, ist es wichtig, Verbräuche und Emissionen bei der Entwicklung und im Einsatz zu beachten. Dabei ist nicht nur die Modellarchitektur entscheidend, sondern auch der Einsatzkontext: Bei einem System zum intelligenten Quartiersmanagement kann es zum Beispiel sein, dass es nur einmal täglich angewendet wird, während bei anderen Systemen, etwa im Onlinehandel oder zu Werbezwecken, millionenfach Abfragen an einem Tag stattfinden. Darüber hinaus müssen wir uns damit befassen, wo die Ressourcen herkommen und in welchen Regionen die Rechenzentren stehen sollen, ohne die die Modelle und Anwendungen nicht funktionieren. Es geht also auch darum, wie wir digitale Infrastrukturen gestalten und wie wir sie in Anspruch nehmen. Dabei ist es wichtig, dass wir uns fragen: Wofür wollen wir die Energie oder das Wasser, das wir haben, einsetzen?

Kommen diese Themen in den politischen Debatten an?

Ja, so langsam werden solche Fragen Teil politischer Diskurse. Im Juni fand beispielsweise ein Fachgespräch im Bundestag zu *ChatGPT* statt. Dort konnten wir das Thema Wasser- und Energieverbrauch einbringen. Wir sehen auch im Fall der KI-Verordnung, dass die Auseinandersetzung mit Foundation Models und entsprechenden Anwendungen ein Stück weit angekommen ist. Der Parlamentsentwurf sieht bestimmte Anforderungen wie das Tracken von Energieverbräuchen vor. Dabei ist es wichtig, dass wir das Thema Nachhaltigkeit nicht zu sehr verkürzen und nur noch über Energieverbrauch und CO2-Ausstoß sprechen. Wir müssen uns auch mit den gesellschaftlichen Risiken beschäftigen. Deshalb ist es sinnvoll, ein umfassendes Nachhaltigkeitsverständnis zu etablieren.

Sie haben zuvor von der Größe der Modelle gesprochen und dem Trend, dass sie immer größer werden. Warum ist das so?

Fortschritt wird derzeit an der Entwicklung von immer größeren KI- und Transformer-Modellen festgemacht. Größere Modelle sollen die Genauigkeit verbessern, erhöhen aber auch die Komplexität. Die Frage, ob das im Verhältnis zum daraus hervorgehenden Nutzen überhaupt notwendig ist, spielt meist nicht wirklich eine Rolle.

Fortschritt könnte prinzipiell auch etwas anderes bedeuten – zum Beispiel spezialisierte Modelle für Einsatzzwecke, in denen sie einen wichtigen Mehrwert bieten. Ihre Komplexität wäre dann tendenziell begrenzter beziehungsweise würde ihre Größe ins Verhältnis zu anderen Zielen gesetzt werden. Die zentrale Frage ist hier: Welchen Weg wollen wir gehen?

Wir erleben derzeit die nächste Stufe der Automatisierung. Sie zu gestalten, kann nicht allein in den Händen profitorientierter Unternehmen liegen. Natürlich gibt es viele sinnvolle Einsatzzwecke für komplexe KI-Modelle, etwa verbesserte Klimamodelle zum Monitoring des Klimawandels. Ihr Nutzen hängt vor allem davon ab, wie wir die Modelle gestalten und einsetzen. Die notwendigen strukturellen Veränderungen zur Bekämpfung des Klimawandels kann uns aber kein Algorithmus abnehmen.

Es gibt auch Stimmen, die in Foundation Models Chancen für mehr Nachhaltigkeit sehen, zum Beispiel, dass mehr Entwickler*innen auf bestehende Modelle zugreifen könnten, anstatt sie komplett neu aufzusetzen. Die Voraussetzung wäre, dass die Modelle dementsprechend gestaltet sind und alle notwendigen Informationen zur Verfügung stehen. Was halten Sie davon?

Solche Open-Source-Ansätze sind definitiv eine Chance und es gibt bereits Beispiele dafür wie das Start-up *Hugging Face*, das eine Plattform für das Teilen von Modellen und der notwendigen zusätzlichen Informationen bietet. Das ist eine sinnvolle Entwicklung und wird in Zukunft eine größere Rolle spielen. Modelle wie *Koala*, die auf offenen Datensätzen basieren, sind mittlerweile fast genauso gut wie *ChatGPT* und das zu einem Bruchteil der Entwicklungskosten. Große Technikunternehmen verfügen aber weiterhin über Datenmonopole und haben dadurch ganz andere Möglichkeiten, Anwendungen zu entwickeln. Aus Nachhaltigkeitssicht ist es aber dringend notwendig, den Markt zu diversifizieren. Wenn es mit Open-Source-Modellen gelingt, Eintrittsbarrieren zu verringern, stärkt das auch kleinere und mittlere Unternehmen. Wenn Unternehmen bewerten möchten, ob ihre Modelle bereits bestimmten Nachhaltigkeitskriterien entsprechen, können Sie zum Beispiel das Bewertungstool ausprobieren, das wir im Projekt *SustAIn* entwickelt haben. Ein Fragenkatalog hilft dabei, die soziale, ökologische und ökonomische Nachhaltigkeit zu bewerten.[1]

„Dabei ist es wichtig, dass wir das Thema Nachhaltigkeit nicht zu sehr verkürzen und nur noch über Energieverbrauch und CO2-Ausstoß sprechen. Wir müssen uns auch mit den gesellschaftlichen Risiken beschäftigen.“

1 Vgl. SustAIn, o. S.

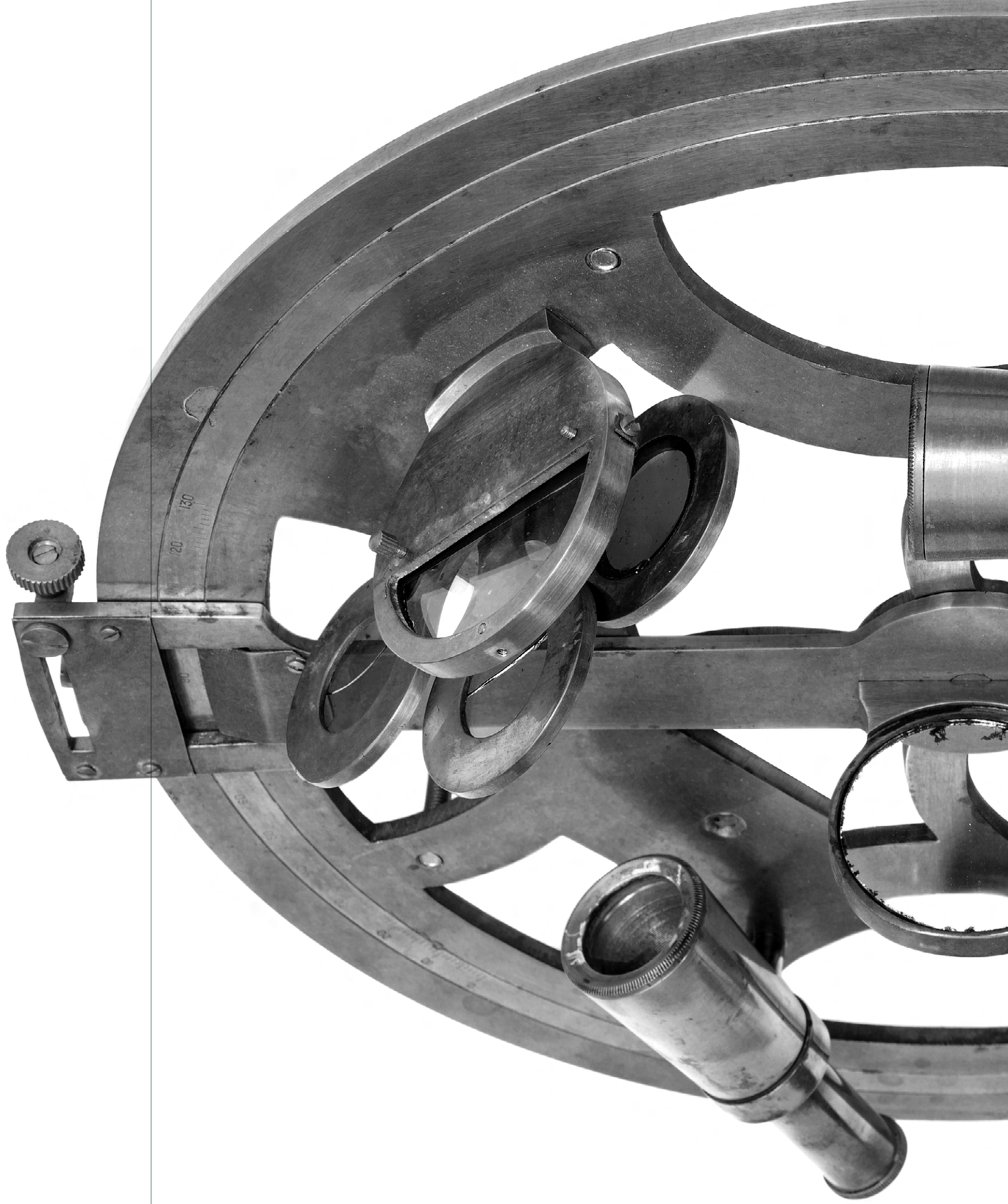
„Große Technikunternehmen verfügen aber weiterhin über Datenmonopole und haben dadurch ganz andere Möglichkeiten, Anwendungen zu entwickeln. Aus Nachhaltigkeitssicht ist es aber dringend notwendig, den Markt zu diversifizieren.“

Sie haben dieses Jahr gemeinsam mit zwei weiteren Autor*innen einen Artikel veröffentlicht, der ausführt, dass es wichtig ist, die Zivilgesellschaft einzubinden, um eine nachhaltige Digitalisierung zu gestalten. Warum ist das so?

Es ist notwendig, die Voraussetzungen für politische Teilhabe so zu verändern, dass sich verschiedene Akteur*innen gleichberechtigt beteiligen können. Das ist derzeit oft nicht gegeben. Big-Tech-Unternehmen können im Kontext von Regulierungsvorhaben umfangreiche finanzielle und zeitliche Ressourcen in die Lobbyarbeit stecken. Demgegenüber stehen häufig prekär finanzierte zivilgesellschaftliche Organisationen. Kann man denen und ihren teilweise ehrenamtlich tätigen Mitarbeiter*innen tatsächlich zumuten, sich zum Beispiel 500 Seiten Gesetzestext in kurzer Zeit durchzulesen, um danach dazu Stellung nehmen zu können? Die zivilgesellschaftlichen Organisationen haben nicht immer die Kapazitäten, jeden Tag an die Tür von Politiker*innen zu klopfen. Hier herrscht ein Ungleichgewicht, das ein Risiko für Regulierungsvorhaben wie die KI-Verordnung ist, wenn nicht Profitorientierung, sondern Menschen im Mittelpunkt stehen sollen. Um ein Gegengewicht zur intensiven Lobbyarbeit großer Tech-Konzerne zu etablieren, sind klare Regeln wichtig, wann welche Stakeholder einbezogen werden müssen. Eine Form aufsuchender Beteiligung ist notwendig, damit sich verschiedene Stakeholder gleichberechtigt einbringen können.

© privat

Friederike Rohde ist Nachhaltigkeitsforscherin und Techniksoziologin. Sie beschäftigt sich mit den Wechselwirkungen zwischen technischen und sozialen Veränderungsdynamiken. Am Institut für ökologische Wirtschaftsforschung (IÖW) arbeitet sie als wissenschaftliche Mitarbeiterin. Ihre Schwerpunkte sind die nachhaltige Gestaltung des digitalen Wandels sowie Narrative und Zukunftsvisionen zur Digitalisierung. Rohde forscht derzeit zu sozialen Innovationen im Energiebereich und Nachhaltigkeitskriterien für Künstliche Intelligenz.



Interview von Jaana Müller-Brehm

WELCHE FRAGEN BLEIBEN, WENN DER HYPE ABEBBT?

Die Aufmerksamkeitswelle um neue KI-Modelle flacht nur langsam ab. Dabei ist die KI-Berichterstattung von Erzählungen einer „Superintelligenz“ und von Vermenschlichung geprägt. Beides hat im Kontext von KI-Modellen Tradition, aber in den Debatten eigentlich nichts zu suchen, sagt Andreas Jungherr. Er beschreibt, wie der Hype entstehen konnte und welche für unser Zusammenleben wichtigen Aspekte bislang zu wenig Aufmerksamkeit erhalten.

KI-Modelle sind spätestens seit dem Launch von *ChatGPT* kein Nischenthema mehr. Wir haben in den letzten Monaten sehr viele Berichte über „Superintelligenzen“ lesen, sehen und hören können. Wie haben Sie die mediale Auseinandersetzung erlebt?

Die neuen Sprachmodelle haben eine Aufmerksamkeitswelle ausgelöst. Als es Ende 2022 losging, dachte ich noch: Okay, wir reden jetzt einmal wieder ein bis zwei Wochen über einen neuen Prototypen und dann ebbt das wieder ab. Dieses Mal war es anders. Der Launch von *ChatGPT* und die Kommunikation dazu erfolgte gut dosiert. *OpenAI* und *Microsoft* veröffentlichten alle paar Wochen etwas Neues dazu. Dadurch konnten sie Aufmerksamkeit binden. Außerdem war es für viele zum ersten Mal möglich, ein solches komplexes KI-System selbst auszuprobieren - ohne Vorwissen oder besondere Kenntnisse. Zusätzlich können wir mit den Ergebnissen, den Texten von *ChatGPT*, alle etwas anfangen. Sie spielen in vielen Arbeitskontexten eine Rolle. Die Automatisierung betrifft damit plötzlich nicht mehr nur bestimmte industrielle Produktionsprozesse, sondern für uns bislang menschlich besetzte Tätigkeiten wie das Verfassen von einfachen Texten.

Die Bilder, die Medienberichte häufig von neuen KI-Modellen zeichnen, sind stark von Vermenschlichung geprägt. Warum ist das so?

Für viele Journalist*innen ist es einfacher, mit Vermenschlichung zu arbeiten, weil es sowohl für sie als auch für die Rezipient*innen verständlicher sein kann, als über Gewichtungen und Modellarchitekturen zu sprechen. Diese Bilder erzeugen Vorstellungskraft und Anschlussfähigkeit. Zur Frage, ob meine Arbeit von einer Maschine ersetzt werden könnte, kann ich ohne genauere Kenntnisse eine Meinung haben. Auch Technikkonzerne fördern diese Narrative gezielt. Sie haben ein Interesse daran, dass wir ihren Produkten eine unglaubliche Fähigkeit und Reichweite zuschreiben - ungeachtet dessen, ob das der Realität entspricht. Wenn in einem Produkt die Fähigkeit vermutet wird, die Welt aus der Balance bringen zu können, sind die dahinterstehenden Unternehmen wichtig. In sie zu investieren, wirkt dann bedeutsam.

Ist das problematisch?

Wenn wir davon sprechen, dass Sprachmodelle „verstehen“ oder „halluzinieren“, sind das letztlich Begriffe, die in diesem Kontext nichts zu suchen haben. Sie entsprechen nicht den dahinterstehenden Mechaniken. Das bedeutet nicht, dass in diesen KI-Systemen und ihrer Entwicklung kein Wert begründet ist, darin keine Leistung steckt oder ihre Entwicklung kein Fortschritt ist. Während wir allerdings über die scheinbaren Fähigkeiten der neuen KI-Modelle sprechen, beachten wir andere wichtige Fragen nicht. Dazu zählen etwa Moderationsentscheidungen, auf denen entsprechende KI-Systeme basieren, oder Fragen nach der Urheberschaft und Entlohnung, die gerade bei generativen KI-Systemen wichtig sind. Wenn wir diese Welle der Aufmerksamkeit hinter uns haben, kommt wahrscheinlich eine Welle der Ernüchterung. Wir müssen erst einmal sehen, wie entsprechende Anwendungen auf diesen prototypischen Modellen aufbauen und wie sie in der Praxis tatsächlich funktionieren.

> „Wenn wir davon sprechen, dass Sprachmodelle ‚verstehen‘ oder ‚halluzinieren‘, sind das letztlich Begriffe, die in diesem Kontext nichts zu suchen haben. Sie entsprechen nicht den dahinterstehenden Mechaniken.“

Abgesehen davon, auf welche Weise wir uns mit KI auseinandersetzen, hat der Hype um *ChatGPT* das Thema auch stärker in die breite Öffentlichkeit getragen. Das bietet auch die Chance, darauf aufzubauen, um uns als Gesellschaft mit wichtigen Fragen auseinanderzusetzen. Geht diese Anschlussfähigkeit womöglich auch mit einem Abebben der Aufmerksamkeit verloren?

Wie nachhaltig die Wahrnehmung von KI ist, ist damit verbunden, wie viele Produkte aus den bislang existierenden Prototypen entstehen. Hinzu kommt die Frage, ob wir diese Produkte und ihre Funktionen tatsächlich noch als KI wahrnehmen. Wenn Sprachmodelle beispielsweise in Textverarbeitungsprogrammen ausgerollt werden, nehmen wir das als neue Funktionen der Software wahr. Die KI-Modelle sind für uns dann weniger präsent. Der Neuigkeitswert verliert sich und wir sprechen nicht mehr so viel darüber. Das ist derzeit im Zusammenhang mit vielen KI-Anwendungen der Fall, zum Beispiel bei Suchmaschinen.

Auch wenn der Hype abflacht, ist es wichtig, sich mit der Entwicklung, Funktionsweise und dem Einsatz von KI-Modellen zu beschäftigen. Sie können sich auf unser Zusammenleben auswirken. Inwiefern berühren komplexe KI-Systeme unsere demokratischen Strukturen?

In einer Demokratie sollen wir in der Lage sein, selbstbestimmt politische Entscheidungen zu treffen, zum Beispiel bei Wahlen. Das erfordert, dass wir rechtlich dazu in die Lage versetzt werden, aber eben auch, dass wir eine Meinung äußern und bilden können. Deshalb ist die Struktur von Informations- und Kommunikationsumgebungen extrem wichtig. In diesem Zusammenhang spielen durch KI-Systeme strukturierte Kommunikationsräume eine Rolle. Das betrifft gängige KI-Anwendungskontexte wie im Fall von sozialen Netzwerken oder Suchmaschinen. Von außen ist hier das tatsächliche Ausmaß des Einflusses der KI-Systeme schwer zu durchschauen. Die Annahme, dass Menschen auf Grundlage manipulierter Informationsumgebungen Entscheidungen treffen, öffnet in jedem Fall die Tür dafür, demokratische Prozesse wie Wahlen infrage zu stellen. Es gibt dann viele Gründe, die Wahlverlierer*innen anführen können, warum eine Wahl nicht legitim war. Das ist eine extrem gefährliche Situation, wie wir es derzeit in den USA sehen. Das betrifft aber auch Deutschland, wo ebenfalls Wahlergebnisse durch bestimmte Akteur*innen delegitimiert werden.

> „Solche strukturellen Veränderungen können zu einer kontrollierteren Kommunikationsumgebung beitragen, aber eben auch zu einer, der es an Diversität fehlt.“

Perspektivisch ist das dialogische Format von Suchanfragen, das Large Language Models ermöglichen, ein zusätzliches Problem. Es stellen sich Fragen danach, welche Moderationsregeln hinterlegt sind und ob die Ergebnisse korrekte und ausgewogene Darstellungen wiedergeben. Zugleich ist denkbar, dass künftig dialogische Antworten auf Suchanfragen Nutzer*innen die Notwendigkeit nehmen, die jeweiligen Quellen zu besuchen. Damit verlieren mittlere und kleine Informationsangebote Monetisierungsoptionen. Über die Zeit wird damit die Zahl politischer Informationsangebote und Quellen zurückgehen. Solche strukturellen Veränderungen können zu einer kontrollierteren Kommunikationsumgebung beitragen, aber eben auch zu einer, der es an Diversität fehlt.

Andreas Jungherr ist Inhaber des Lehrstuhls für Politikwissenschaft, insbesondere Digitale Transformation, an der Universität Bamberg. Er untersucht die Auswirkungen der Digitalisierung auf Politik und Gesellschaft. Ihn interessieren unter anderem Einstellungsstrukturen zu Künstlicher Intelligenz. Jungherr befasst sich außerdem mit den Herausforderungen und Chancen bei der Nutzung neuer Datenquellen, von KI und computerbasierten Methoden in der Sozialwissenschaft.

Interview von
Jaana Müller-Brehm

KOMBINIEREN

FORTSCHRITT WOHIN?

Der Mainstream der KI-Entwicklung macht Fortschritt vor allem an zwei Dingen fest: an der Komplexität der Modelle und daran, wie gut sie Aufgaben lösen, die bislang Menschen erledigt haben. Wenn wir dieses Verständnis von Fortschritt ins Verhältnis zu anderen gesellschaftlichen Anliegen wie Nachhaltigkeitszielen setzen, ergibt das keinen Sinn. Um die Probleme der Gegenwart zu erkennen und Lösungen dafür zu finden, brauchen wir ein neues Verständnis.

Nachhaltigkeitsdiskurse und -ansprüche gewinnen an Bedeutung. Zugleich sind sie von tiefen Zielkonflikten und vielen offenen Fragen geprägt – zum Beispiel, wie Leitlinien des Umweltschutzes, der sozialen Gerechtigkeit und des individuellen Wohlstands konkret umgesetzt werden können.[1] Einige hoffen, dass KI-Verfahren die fehlenden Antworten liefern. Schließlich können KI-Systeme dabei helfen, das Waldsterben und die Biodiversität von Ökosystemen zu überwachen, alternative Materialien für knappe Rohstoffe zu finden oder die Kreislaufwirtschaft zu verbessern.[2] Das ist zwar korrekt, zugleich aber lediglich eine Momentaufnahme einer längeren Geschichte.

KI, wie wir sie derzeit in der Breite entwickeln und einsetzen, ist nicht die Antwort auf die Frage, wie wir Nachhaltigkeitsziele umsetzen. Vielmehr deutet der Trend in die entgegengesetzte Richtung: Basismodelle und darauf aufbauende Systeme werden unter sehr viel größeren Ressourcen- und Energieaufwänden entwickelt als kleinere KI-Modelle mit klaren Einsatzzwecken. Das führt wiederum zu steigenden Ausstößen von Treibhausgasen und endet nicht mit dem Trainingsprozess. Wenn Systeme wie *ChatGPT* für eine durchschnittliche Anfrage etwa einen halben Liter Trinkwasser verbrauchen, kommt es spätestens mit der massenhaften Nutzung zu erheblichen Verbräuchen. Genaue Zahlen dazu gibt es derzeit nicht. Denn bislang ist es weder aus rechtlichen Gründen noch aus solchen der Produktionslogik notwendig, Verbräuche bei der KI-Entwicklung und entlang der gesamten Wertschöpfungskette genau zu erfassen und zu dokumentieren.

Auch die Infrastrukturen, die Basismodelle benötigen, erhalten zu wenig Aufmerksamkeit. Stehen die meisten Rechenzentren, die mit Wasser gekühlt werden, in von Dürre geplagten Regionen? Aus welchen Energieträgern speist sich der Energiebedarf für Training und Anwendung? Woher stammen die für die Hardware notwendigen Seltenen Erden und unter welchen Bedingungen findet eigentlich der Abbau statt? Diese Fragen werden zwar langsam Teil politischer Debatten um KI im Allgemeinen. Doch in der Praxis sind sie vor allem in Bezug auf Foundation Models noch lange nicht angekommen.

Die Entwicklung von Basismodellen wird derzeit maßgeblich von großen Technikunternehmen vorangetrieben. Als Oligopole haben sie weitreichenden Zugriff auf wichtige Datengrundlagen, Fachkräfte und finanzielle Ressourcen, um die Entwicklung der komplexen Modelle weiter voranzutreiben. Damit können einige wenige Unternehmen Standards setzen, nach denen sich die anderen Marktteilnehmer unter den vorgeschriebenen Bedingungen richten müssen. Die Bedeutung dieser wenigen Unternehmen nimmt stetig zu. Im Gegensatz dazu arbeiten umfassende Open-Source-Ansätze mit dem Teilen von Datensätzen, Modellen und Kenntnissen. So können beispielsweise sinnvolle Strukturen des organisationsübergreifenden Wissensaufbaus entwickelt und erprobt werden. Solche Ansätze ermöglichen unabhängige Forschung. Das allein genügt jedoch nicht, um Oligopole aufzubrechen und einen vielfältigeren Markt zu verwirklichen. Weitere Maßnahmen werden unter anderem darin gesehen, eine unabhängige KI-Forschung sowie kleine und mittlere Unternehmen zu fördern.

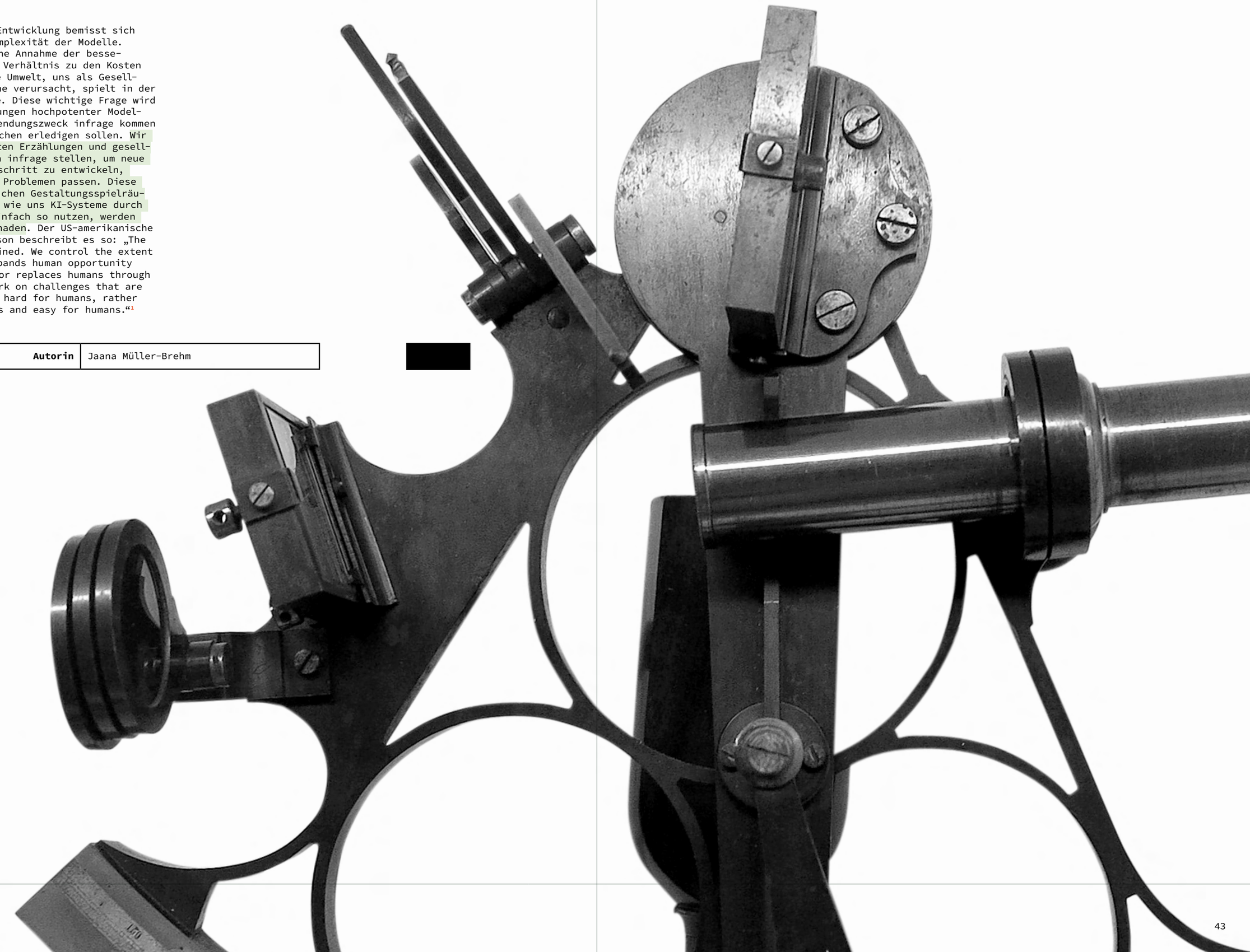
Bestehende Gesetze wie der *Digital Markets Act* oder die *DSGVO* begegnen den besonderen Herausforderungen von Foundation Models nicht. Der Parlamentsentwurf der KI-Verordnung enthält zwar einige Regeln, die sich explizit an die Anbieter*innen von Basismodellen und ihre Anwender*innen richten. Der Entwurf bleibt dabei aber unspezifisch. Zugleich können Anforderungen an Dokumentationen oder die Datenqualität mit bisherigen Ansätzen nur schwer verwirklicht werden. Es ist beispielsweise nicht möglich, die Repräsentativität von Datensätzen für die vorgesehene Aufgabe zu prüfen, wenn Informationen zur Datengrundlage gleichermaßen fehlen wie solche zum Einsatzzweck. Ähnliches gilt für Transparenzmaßnahmen, die auch aufgrund der Komplexität der Modelle schwer umzusetzen sind. Das steht im Gegensatz zu zentralen Anliegen der KI- Verordnung wie dem Gewährleisten von menschlicher Aufsicht, dem Schaffen von Transparenz und dem Prüfen der Datenqualität. Dabei sind gerade diese Punkte für Foundation Models von besonderer Bedeutung: Die Modelle dienen als Grundlage für eine Vielzahl von Anwendungen. Dadurch verbreiten sich enthaltene Fehler oder Verzerrungen, die zu Diskriminierungen führen können, massenhaft.

Wissenschaftler*innen arbeiten an neuen Ansätzen, um diese Besonderheiten zu berücksichtigen. Aus ihren Arbeiten lassen sich etwa Strategien ableiten, um eine systematische Dokumentation des Energie- und Ressourcenverbrauchs, des CO2-Ausstoßes oder wiederkehrender Schädigungen durch den Einsatz umzusetzen. Forscher*innen entwickeln zudem Methoden, um komplexe Modelle zu verschlanken. Aus solchen Arbeiten lassen sich dann wiederum übergeordnete Maßnahmen ableiten, um die Nachhaltigkeit von Basismodellen zu verbessern. Das allein ist allerdings nicht die Lösung. Wir kommen nicht um die Fragen herum, wann der Einsatz komplexer KI-Modelle überhaupt sinnvoll ist und wie ein solcher sinnvoller Einsatz gestaltet sein muss. Wir benötigen entsprechende Kompetenzen und das Wissen, um die mit Basismodellen verbundenen Probleme zu erkennen, zu analysieren und sie in einen größeren Kontext zu setzen, etwa den der Nachhaltigkeit.

1 Vgl. Henkel et al., S. 12.
2 Vgl. Müller-Brehm/ Otto.

Im Mainstream der KI-Entwicklung bemisst sich Fortschritt an der Komplexität der Modelle. Ob die damit verbundene Annahme der besseren Funktionsweise im Verhältnis zu den Kosten steht, die sie für die Umwelt, uns als Gesellschaft und als Einzelne verursacht, spielt in der Breite kaum eine Rolle. Diese wichtige Frage wird überlagert von Erzählungen hochpotenter Modelle, die für jeden Anwendungszweck infrage kommen und Aufgaben von Menschen erledigen sollen. Wir müssen diese etablierten Erzählungen und gesellschaftlichen Praktiken infrage stellen, um neue Perspektiven auf Fortschritt zu entwickeln, die besser zu unseren Problemen passen. Diese Perspektiven verdeutlichen Gestaltungsspielräume. Denn ebenso wenig wie uns KI-Systeme durch ihre reine Existenz einfach so nutzen, werden sie uns einfach so schaden. Der US-amerikanische Ökonom Erik Brynjolfsson beschreibt es so: „The future is not preordained. We control the extent to which AI either expands human opportunity through augmentation or replaces humans through automation. We can work on challenges that are easy for machines and hard for humans, rather than hard for machines and easy for humans.“[1]

Autorin | Jaana Müller-Brehm



1 Brynjolfsson, S. 14.

VERBINDEN

WOZU ALL DAS?

Wozu gibt es dieses Magazin? Warum braucht es ein *Zentrum für vertrauenswürdige Künstliche Intelligenz (ZVKI)*? Wir möchten Wissen und verschiedene Perspektiven rund um den großen Themenkomplex Künstliche Intelligenz miteinander verbinden, um neue Erkenntnisse zu gewinnen. Im Fokus stehen dabei wir Menschen, unsere Grundrechte, unser Wohlergehen und unser Zusammenleben. Wir wollen herausfinden, wann die Zuschreibung ‚vertrauenswürdig' gerechtfertigt ist.

Algorithmische Systeme und Verfahren der Künstlichen Intelligenz sind Begriffe aus Fachdiskussionen und zugleich Teil unseres Alltags. Mit ihrer Hilfe werden Entscheidungen getroffen, die Auswirkungen auf unser Leben haben. Um diese Auswirkungen zu erfassen, verständlich darzustellen und mit ihnen umzugehen, müssen wir aus Echokammern ausbrechen, Silodenken hinter uns lassen und Brücken zwischen Insellösungen bauen.

Das *ZVKI* ist ein zentraler Ort der Debatte in Deutschland. Es macht die Entwicklungen rund um gesellschaftliche Fragen zu Künstlicher Intelligenz und algorithmischen Systemen greifbar. Zugleich ist es eine unparteiische Schnittstelle zwischen Wissenschaft, Wirtschaft, Politik und Zivilgesellschaft, die gemeinsam mit ihren Partner*innen Instrumente entwickelt, um vertrauenswürdige KI zu bewerten.

Die Ziele des *ZVKI* sind unter anderem:

- Informieren, Wissen vermitteln und aufklären: Verständnis ist die Voraussetzung, um Vertrauen aufzubauen. Deshalb bündeln wir Informationen und bereiten sie für verschiedene Zielgruppen auf.

- Forschen und wissenschaftliche Erkenntnisse verständlich darstellen: Wir untersuchen unter anderem, welche Schritte unternommen werden müssen, um negative Auswirkungen von KI-Systemen zu erkennen und ihnen zu begegnen.

- Evaluieren, Regulieren, Zertifizieren: Wir übersetzen Fachthemen rund um KI und Regulierung für Fachpublikum und (politische) Entscheidungsträger*innen und ermächtigen sie so, Rahmenbedingungen für Verbraucher*innenschutz zu gestalten.

- Netzwerken und unterstützen: Um möglichst viele Stakeholder*innen sowie deren Ansätze und Ideen zusammenzubringen, bieten wir verschiedene Formate des Austauschs an.

Um diese Ziele zu erreichen, arbeiten wir als interdisziplinäres Team und mit verschiedenen Partner*innen zusammen:

Mit Unterstützung des *Bundesministeriums für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz (BMUV)* baut der unabhängige Think Tank *iRights.Lab*, in Zusammenarbeit mit den *Fraunhofer-Instituten AISEC* und *IAIS* sowie der *Freien Universität Berlin*, das *ZVKI* auf.

Mitmachen

Das *Zentrum für vertrauenswürdige KI - ZVKI* versteht sich als unparteiische Schnittstelle zwischen Disziplinen und Akteur*innen, zwischen Nutzer*innen und Expert*innen. Treten Sie mit uns in Kontakt und in den Austausch. Sie erreichen uns über zvki@irights-lab.de.

Sie finden uns auch auf Instagram: zvki.de, X (Twitter): ZVKI_de und LinkedIn: ZVKI

BELEGEN

WOHER STAMMEN DIE INFORMATIONEN?

Quellen

acatech – Deutsche Akademie der Technikwissenschaften e.V. (2023): Europäisches Datengesetz: Wie die EU-Regeln KI-Innovationen fördern können. Lernende Systeme. Die Plattform für Künstliche Intelligenz. 23. Februar 2023. https://www.plattform-lernende-systeme.de/aktuelles-newsreader/europaeisches-datengesetz-wie-die-eu-regeln-ki-innovationen-foerdern-koennen.html.

AFP (2023): France to invest €500 million to fund AI ‚champions', Macron says. France 24. 15. Juni 2023. https://www.france24.com/en/europe/20230615-macron-wants-to-boost-ai-calls-for-smart-rules-that-don-t-impede-tech-growth.

Albrecht, Steffen (2023): ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen. TAB-Hintergrundpapier Nr. 26. Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag. https://www.bundestag.de/resource/blob/944148/30b0896f6e49908155fcd01d77f57922/20-18-109-Hintergrundpapier-data.pdf.

Ananthaswamy, Anil (2023): In AI, Is Bigger Always Better? In: Nature 615 (7951), S. 202–205. https://doi.org/10.1038/d41586-023-00641-w.

AW AlgorithmWatch (2023a): KI Terribilis. In: SUSTAIN. KI und ihre Folgen für die Nachhaltigkeit, Magazin 2, Frühjahr 2023, S. 10–13.

AW AlgorithmWatch (2023b): Gemeinschaft statt Größenwahn: Mit Open Source zu nachhaltigen Lösungen. In: SUSTAIN. KI und ihre Folgen für die Nachhaltigkeit, Magazin 2, Frühjahr 2023, S. 14–18.

Bender, Emily M./ Gebru,Timnit / McMillan-Major, Angelina und Shmitchell, Shmargaret (2021): On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, S. 610–623. https://doi.org/10.1145/3442188.3445922.

Biderman, Stella/ Huebner, Curtis / Leahy, Connor und Hallahan, Eric (2021): What A Long, Strange Trip It's Been: EleutherAI One Year Retrospective. EleutherAI. 7. Juli 2021. https://blog.eleuther.ai/year-one/.

Biderman, Stella/ Huebner, Curtis / Leahy, Connor und Hallahan, Eric (2023): The View from 30,000 Feet: Preface to the Second EleutherAI Retrospective. EleutherAI. 2. März 2023. https://blog.eleuther.ai/year-two-preface/.

Bienert, Jörg/ Just, Vanessa / Rothe, Rasmus / Abbou, Daniel / Handy, Philipp und Blank, Alessandro (2023): The EU AI Act. Towards the Finish Line: Key Issues and Proposals of the Trilogy Negotiations. German AI Association. https://ki-verband.de/wp-content/uploads/2023/07/Position-Paper_AI-Act-Trilogue_GermanAIAssociation.pdf.

Biewald, Lukas (2023): How EleutherAI Trains and Releases LLMs: Interview with Stella Biderman. Gradient Dissent: Exploring Machine Learning, AI, Deep Learning, Computer Vision. 4. Mai 2023. https://player.captivate.fm/episode/1f33b906-03a8-47cf-81cb-068497d4419a.

Binder, Matthieu (2023): Welcome to Trilogue – Worüber jetzt noch diskutiert wird. ZVKI. 10. Juli 2023. https://www.zvki.de/zvki-exklusiv/fachinformationen/ki-vo-briefing-4.

Bisbee, James/ Clinton, Joshua/ Dorff, Cassy / Kenkel, Brenton und Larson, Jennifer (2023): Artificially Precise Extremism: How Internet-Trained LLMs Exaggerate Our Differences. Preprint. 4. Mai 2023. SocArXiv. https://doi.org/10.31235/osf.io/5ecfa.

Black, Sid/ Biderman, Stella/ Hallahan, Eric/ Anthony, Quentin/ Gao, Leo / Golding, Laurence/ He, Horace u. a. (2022): GPT-NeoX-20B: An Open-Source Autoregressive Language Model". arXiv. 14. April 2022. http://arxiv.org/abs/2204.06745.

Bommasani, Rishi/ Hudson, Drew A./ Adeli, Ehsan/ Altman, Russ/ Arora, Simran/ von Arx, Sydney/ Bernstein, Michael S. u. a. (2022): On the Opportunities and Risks of Foundation Models. 12. Juli 2022. http://arxiv.org/abs/2108.07258.

Brooks, Rodney (2018): „Die Ursprünge der Künstlichen Intelligenz". reframe[tech]. 22. November 2018. https://www.reframetech.de/2018/11/22/die-ursprunge-der-kuenstlichen-intelligenz/.

Brynjolfsson, Erik (2022): „The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence". In: Daedalus 151 (2), S. 272–287. https://doi.org/10.1162/daed_a_01915.

Bubeck, Sébastien/ Chandrasekaran, Varun/ Eldan, Ronen/ Gehrke, Johannes/ Horvitz, Eric/ Kamar, Ece/ Lee, Peter u. a. (2023): Sparks of Artificial General Intelligence: Early Experiments with GPT-4. 13. April 2023. http://arxiv.org/abs/2303.12712.

Crawford, Kate/ Calo, Ryan (2016): There Is a Blind Spot in AI Research. In: Nature 538 (7625), S. 311–313. https://doi.org/10.1038/538311a.

D16 Digitalministertreffen (2022): Positionierung der Länder gegenüber der geplanten KI-Verordnung der Europäischen Union. 12. Dezember 2022. https://www.baden-wuerttemberg.de/fileadmin/redaktion/m-im/intern/dateien/pdf/221214_PM_Digitalministertreffen_D16_Positionierung_KI-Verordnung_der_Europäischen_Union.pdf.

Deutscher Ethikrat (2023): Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz. Stellungnahme. 20.03.2023. https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf.

Dev, Sunipa/ Sheng, Emily / Zhao, Jieyu/ Amstutz, Aubrie/ Sun, Jiao/ Hou, Yu/ Sanseverino, Mattie u. a. (2022): On Measures of Biases and Harms in NLP. 13. Oktober 2022. https://doi.org/10.48550/arXiv.2108.03362.

Dis, Eva A. M. van/ Bollen, Johan/ Zuidema, Willem/ van Rooji, Robert und Bockting, Claudi L. (2023): ChatGPT: five priorities for research. nature. 3. Februar 2023. https://www.nature.com/articles/d41586-023-00288-7.

EleutherAI (2023): Empowering Open-Source Artificial Intelligence Research. https://www.eleuther.ai.

epd und dpa (2022): Neues EU-Gesetz gegen Hass und Gewalt im Netz. ZDFheute. 23. April 2022. https://www.zdf.de/nachrichten/digitales/eu-digital-services-act-100.html.

Europäische Kommission (2021): Artificial Intelligence Act – Amendments Adopted by the European Parliament on 14 June 2023 on the Proposal for a Regulation of the European Parliament and of the Council on Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)). https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf.

European Environment Agency (2018): Water Use in Europe – Quantity and Quality Face Big Challenges. 30. August 2018. https://www.eea.europa.eu/signals/signals-2018-content-list/articles/water-use-in-europe-2014.

Gao, Leo/ Biderman, Stella/ Black, Sid/ Golding, Laurence/ Hoppe, Travis/ Foster, Charles/ Phang, Jason u. a. (2023): The Pile: An 800GB Dataset of Diverse Text for Language Modeling. EleutherAI. https://pile.eleuther.ai/paper.pdf.

Gebru, Timnit/ Hanna, Alex/ Kak, Amba/ Myers West, Sarah/ Solaiman, Irene/ Khan, Metab/ Talat, Zeerak u. a. (2023): Five considerations to guide the regulation of "General Purpose AI" in the EU's AI Act:

Policy guidance from a group of international AI experts. https://algorithmwatch.org/de/wp-content/uploads/2023/04/GPAI-Policy-Brief-.pdf.

Generalversammlung Vereinte Nationen (Hrsg.) (2015): „Resolution der Generalversammlung, verabschiedet am 25. September 2015. https://www.un.org/Depts/german/gv-70/band1/ar70001.pdf.

Google (Hrsg.): TPU Research Cloud. Google. O. J. https://sites.research.google/trc/about/.

Hacker, Philipp (2023). So kriegen wir nie ein europäisches ChatGPT. ZEIT Online, 14. Juni 2023. https://www.zeit.de/digital/2023-06/ki-verordnung-eu-parlament-gesetz/.

Hacker, Philipp/ Engel, Andreas/ Mauer, Marco (2023): Regulating ChatGPT and Other Large Generative AI Models. In: 2023 ACM Conference on Fairness, Accountability, and Transparency, S. 1112–1123. https://doi.org/10.1145/3593013.3594067.

Harrod, Jordan (2021): What Are Foundation Models? On the Opportunities and Risks of Foundation Models. 19. Oktober 2021. https://www.youtube.com/watch?v=TmY-Wz4XfRM0.

Heaven, Will Douglas (2023): The open-source AI boom is built on Big Tech's handouts. How long will it last? MIT Technology Review. 12. Mai 2023. https://www.technologyreview.com/2023/05/12/1072950/open-source-ai-google-openai-eleuther-meta/.

Helberger, Natali/ Diakopoulos, Nicholas (2023): ChatGPT and the AI Act. In: Internet Policy Review 12 (1), S. 1–12. https://doi.org/10.14763/2023.1.1682.

Helmus, Todd C. (2022): Artificial Intelligence, Deepfakes, and Disinformation A Primer. RAND Corporation. Juli 2022. https://www.jstor.org/stable/resrep42027.

Henkel, Anna/ Wendt, Björn/ Barth, Thomas/ Besio, Cristina/ Block, Katharina/ Böschen, Stefan/ Dickel, Sascha u. a. (2021): Zur Einleitung: Kernaspekte einer Soziologie der Nachhaltigkeit. In: Soziologie der Nachhaltigkeit, herausgegeben von SONA – Netzwerk Soziologie der Nachhaltigkeit, S. 9–31. https://doi.org/10.14361/9783839451991.

High-Level Expert Group on AI (2019): Ethics Guidelines for Trustworthy Artificial Intelligence. Herausgegeben von Europäische Kommission. Publications Office of the EU. https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1.

Hugging Face: The AI community building the future. O. J. https://huggingface.co.

Hwang, Yoori/ Ryu, i Youn/ Jeong, Se-Hoon (2021): Effects of Disinformation Using Deepfake: The Protective Effect of Media Literacy Education. In: Cyberpsychology, Behavior, and Social Networking 24 (3), S. 188–193. https://www.liebertpub.com/doi/full/10.1089/cyber.2020.0174.

Jungherr, Andreas (2023): Artificial Intelligence and Democracy: A Conceptual Framework. In: Social Media + Society 9 (3). https://doi.org/10.1177/20563051231186353.

Khan, Saif/ Mann, Alexander (2020): AI Chips: What They Are and Why They Matter. Center for Security and Emerging Technology. April 2020. https://doi.org/10.51593/20190014.

Klaus, Julia (2023): Warum das Papst-Foto nicht nur witzig ist. ZDFheute. 27. März 2023. https://www.zdf.de/nachrichten/digitales/papst-daunenjacke-fake-ki-kuenstliche-intelligenz-100.html.

Koch, Moritz/ Volkery, Carsten/ Neuerer, Dietmar (2023): So will die EU KI-Betrug vorbeugen. Handelsblatt. 5. Juni 2023. https://www.handelsblatt.com/politik/deutschland/kuenstliche-intelligenz-so-will-die-eu-ki-betrug-vorbeugen/29188272.html.

Kroker, Michael (2023): „Dieses Gesetz wird die KI-Landschaft verändern". WirtschaftsWoche. 12. Juli 2023. https://www.wiwo.de/technologie/digitale-welt/ai-act-der-europaeischen-union-dieses-gesetz-wird-die-ki-landschaft-veraendern/29251780.html.

Lantwin, Tobias (2019): Deep Fakes – Düstere Zeiten für den Persönlichkeitsschutz? In: MMR – Zeitschrift für IT-Recht und Recht der Digitalisierung 574, S. 574–578. https://beck-online.beck.de/Dokument?vpath=bibdata%2Fzeits%2FMMR%2F2019%2Fcont%2FMMR.2019.574.1.htm.

Lantwin, Tobias (2020): Strafrechtliche Bekämpfung missbräuchlicher Deep Fakes. In: MMR – Zeitschrift für IT-Recht und Recht der Digitalisierung 78, S. 78–82.

Leahy, Connor (2023): Why release a large language model? EleutherAI. 2. Juni 2023. https://blog.eleuther.ai/why-release-a-large-language-model/.

Li, Pengfei/ Yang, Jianyi/ Islam, Mohammad A. und Ren, Shaolei (2023): Making AI Less „Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models. 6. April 2023. http://arxiv.org/abs/2304.03271.

Lossau, Norbert (2020): Deep Fake Gefahren, Herausforderungen und Lösungswege. 2020. Analysen & Argumente Digitale Gesellschaft, Nr. 382. Konrad-Adenauer-Stiftung. Februar 2020. https://www.kas.de/documents/252038/7995358/AA382+Deep+Fake.pdf/de479a86-ee42-2a9a-e038-e18c208b93ac.

Luccioni, Alexandra Sasha/ Viguier, Sylvain/ Ligozat, Anne-Laure (2022): Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. 3. November 2022. http://arxiv.org/abs/2211.02001.

Luccioni, Sasha (2023): The mounting human and environmental costs of generative AI. Ars Technica. 12. Juli 2023. https://arstechnica.com/gadgets/2023/04/generative-ai-is-cool-but-lets-not-forget-its-human-and-environmental-costs/?comments=1&comments-page=1.

Maham, Pegah/ Küspert, Sabrina (2023): Governing General Purpose AI. A Comprehensive Map of Unreliability, Misuse and Systemic Risks. Stiftung Neue Verantwortung. https://www.stiftung-nv.de/sites/default/files/snv_governing_general_purpose_ai_pdf.pdf.

Masakhane (2023): Masakhane. A grassroots NLP community for Africa, by Africans. 27. Juli 2023. https://www.masakhane.io.

Maslej, Nestor/ Loredana, Fattorini/ Brynjolfsson, Erik/ Etchemendy, John/ Ligett, Katrina/ Lyons, Terah/ Manyika, James u. a. (2023): Artificial Intelligence Index Report 2023. Institute for Human-Centered AI (HAI), Stanford University. Februar 2023. https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf.

McCosker, Anthony (2022): Making sense of deepfakes: Socializing AI and building data literacy on GitHub and YouTube. In: New Media & Society 0(0). Mai 2022, S. 1–18. https://journals.sagepub.com/doi/10.1177/14614448221093943.

Mehta, Chavi (2023): Nvidia results show its growing lead in AI chip race. Reuters. 23. Februar 2023. https://www.reuters.com/technology/nvidia-results-show-its-growing-lead-ai-chip-race-2023-02-23/.

Meineck, Sebastian (2023): Das Hype-Theater um moderne Chatbots. Netzpolitik.org. 4. Juli 2023. https://netzpolitik.org/2023/olimpias-augen-das-hype-theater-um-moderne-chatbots/?via=nl#6.

Müller-Brehm, Jaana (2023): Technik und Macht. Missing Link. Magazin für vertrauenswürdige Künstliche Intelligenz, Mai 2023. https://www.zvki.de/storage/publications/zvki_missinglink_3.pdf.

Müller-Brehm, Jaana/ Otto, Philipp (2022): Smarte Technologie gegen den Klimawandel. 15 Fakten über Künstliche Intelligenz. Heinrich Böll Stiftung. https://www.boell.de/sites/default/files/2022-04/BoellFakten_Smarte_Technologie_gegen_den_Klimawandel_15_Fakten_ueber_Kuenstliche_Intelligenz.pdf.

Narang, Sharan (2022): Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance. Google Research. April 2022. https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html.

Ng, Davy Tsz Kit/ Leung, Jac Ka Lok/ Chu, Samuel Kai Wah und Qiao, Maggie Shen (2021): Conceptualizing AI literacy: An exploratory review. In: Computers and Education: Artificial Intelligence 2. https://www.sciencedirect.com/science/article/pii/S2666920X21000357.

OECD (Hrsg.) (2023): AI Language Models: Technological, Socio-Economic and Policy Considerations. OECD Digital Economy Papers, OECD Digital Economy Papers, 352. https://doi.org/10.1787/13d38f92-en.

Ojo, Jessica/ Ogueji, Kelechi (2023): How Good Are Commercial Large Language Models on African Languages? 10. Mai 2023. http://arxiv.org/abs/2305.06530.

Patel, Dylan/ Ahmad, Afzal (2023): Google „We Have No Moat, And Neither Does OpenAI". Leaked Internal Google Document Claims Open Source AI Will Outcompete Google and OpenAI. SemiAnalysis. 4. Mai 2023. https://www.semianalysis.com/p/google-we-have-no-moat-and-neither.

Pattison, Pete (2021): 'Like Slave and Master': DRC Miners Toil for 30p an Hour to Fuel Electric Cars. The Guardian. 8. November 2021. https://www.theguardian.com/global-development/2021/nov/08/cobalt-drc-miners-toil-for-30p-an-hour-to-fuel-electric-cars.

Perrigo, Billy (2023): Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. Time. 18. Januar 2023. https://time.com/6247678/openai-chatgpt-kenya-workers/.

Potsdam-Institut für Klimafolgenforschung (2022): „UN-Nachhaltigkeitsziele nur begrenzt wirksam, warnen Forschende". Potsdam-Institut für Klimafolgenforschung. 20. Juni 2022. https://www.pik-potsdam.de/de/aktuelles/nachrichten/un-nachhatligkeitsziele-wirkungslos-warnen-forschende.

Rat der Europäischen Union (2022): Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union. https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/de/pdf.

Reiner, Laura (2019): Die SDGs – Kein gutes Leben für alle? Deutsche Gesellschaft für die Vereinten Nationen e.V. (blog). 18. Dezember 2019. https://dgvn.de/meldung/die-sdgs-kein-gutes-leben-fuer-alle.

Richter, Felix (2023): Amazon Maintains Lead in the Cloud Market. statista. 8. August 2023. https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers/.

Rohde, Friederike/ Wagner, Josephin/ Reinhard, Philipp/ Petschow, Ulrich/ Meyer, Andreas/ Voß, Marcus und Mollen, Anne (2021): Nachhaltigkeitskriterien für künstliche Intelligenz. Entwicklung eines Kriterien- und Indikatorensets für die Nachhaltigkeitsbewertung von KI-Systemen entlang des Lebenszyklus. Schriftenreihe des IÖW, Nr. 220/21. https://www.ioew.de/fileadmin/user_upload/BILDER_und_Downloaddateien/Publikationen/2021/IOEW_SR_220_Nachhaltigkeitskriterien_fuer_Kuenstliche_Intelligenz.pdf.

Schmierer, Anna-Lena (2023): Datenschutzkonferenz nimmt ChatGPT unter die Lupe. Netzpolitik.org. 7. April 2023. https://netzpolitik.org/2023/openai-datenschutzkonferenz-nimmt-chatgpt-unter-die-lupe/.

Schneider, Jan (2023): Wie KI für Fake-Bilder missbraucht wird. ZDFheute. 24. März 2023. https://www.zdf.de/nachrichten/panorama/deep-fake-fotos-viral-trump-putin-xi-afd-100.html.

Schulzki-Haddouti, Christiane (2023): Nachhaltigkeitsdebatte: Kann Künstliche Intelligenz das Klima noch retten? RiffReporter. 19. Juni 2023. https://www.riffreporter.de/de/umwelt/nachhaltigkeit-ki-kuenstliche-intelligenz-organisation-macht-kontrolle-wbgu-europa?utm_source=newsletter&utm_medium=email&utm_campaign=riff&utm_term=article_button_link.

Sheng, Emily/ Chang, Kai-Wei/ Natarajan, Premkumar und Peng, Nanyun (2021): Societal Biases in Language Generation: Progress and Challenges. 22. Juni 2021. http://arxiv.org/abs/2105.04054.

Singh, Maanvi (2023): As the AI industry booms, what toll will it take on the environment?The Guardian. 8. Juni 2023. https://www.theguardian.com/technology/2023/jun/08/artificial-intelligence-industry-boom-environment-toll.

Stam, Claire (2019): „European Champions: Berlin und Paris legen nach. Euractiv. 20. Februar 2019. https://www.euractiv.de/section/handel-und-industrie/news/european-champions-berlin-und-paris-legen-nach/.

Stoltenberg, Ute (2020): Kultur als Dimension eines Bildungskonzepts für eine nachhaltige Entwicklung. KULTURELLE BILDUNG ONLINE. https://www.kubi-online.de/artikel/kultur-dimension-eines-bildungskonzepts-nachhaltige-entwicklung.

Strubell, Emma/ Ganesh, Ananya/ McCallum, Andrew (2019): Energy and Policy Considerations for Deep Learning in NLP. http://arxiv.org/abs/1906.02243.

SustAIn (2023): Der Fragebogen zur Selbstbewertung gibt Ihnen Orientierung, wie nachhaltig Ihre KI-Systeme sind und wo Sie nachbessern können. SustAIn. Der Nachhaltigkeits-Index für Künstliche Intelligenz. https://sustain.algorithmwatch.org/bewertungstool/.

Tang, Nan/ Yang, Chenyu/ Fan, Ju und Cao, Lei (2023): VerifAI: Verified Generative AI. 6. Juli 2023. http://arxiv.org/abs/2307.02796.

Valero, Jorge (2019): Wettbewerb und Handel: Wie schafft man eigentlich „European Champions"? Euractiv. 26. Juni 2019. https://www.euractiv.de/section/handel-und-industrie/news/industrie-und-handel-wie-schafft-man-eigentlich-european-champions/.

Van Wynsberghe, Aimee (2021): Sustainable AI: AI for Sustainability and the Sustainability of AI. In: AI and Ethics 1 (3), S. 213–218. https://doi.org/10.1007/s43681-021-00043-6.

Verdecchia, Roberto/ Sallou, June/ Cruz, Luís (2023): A Systematic Review of Green AI. 5. Mai 2023. http://arxiv.org/abs/2301.11047.

Vereinte Nationen (2023): Ziele für nachhaltige Entwicklung. https://unric.org/de/17ziele/.

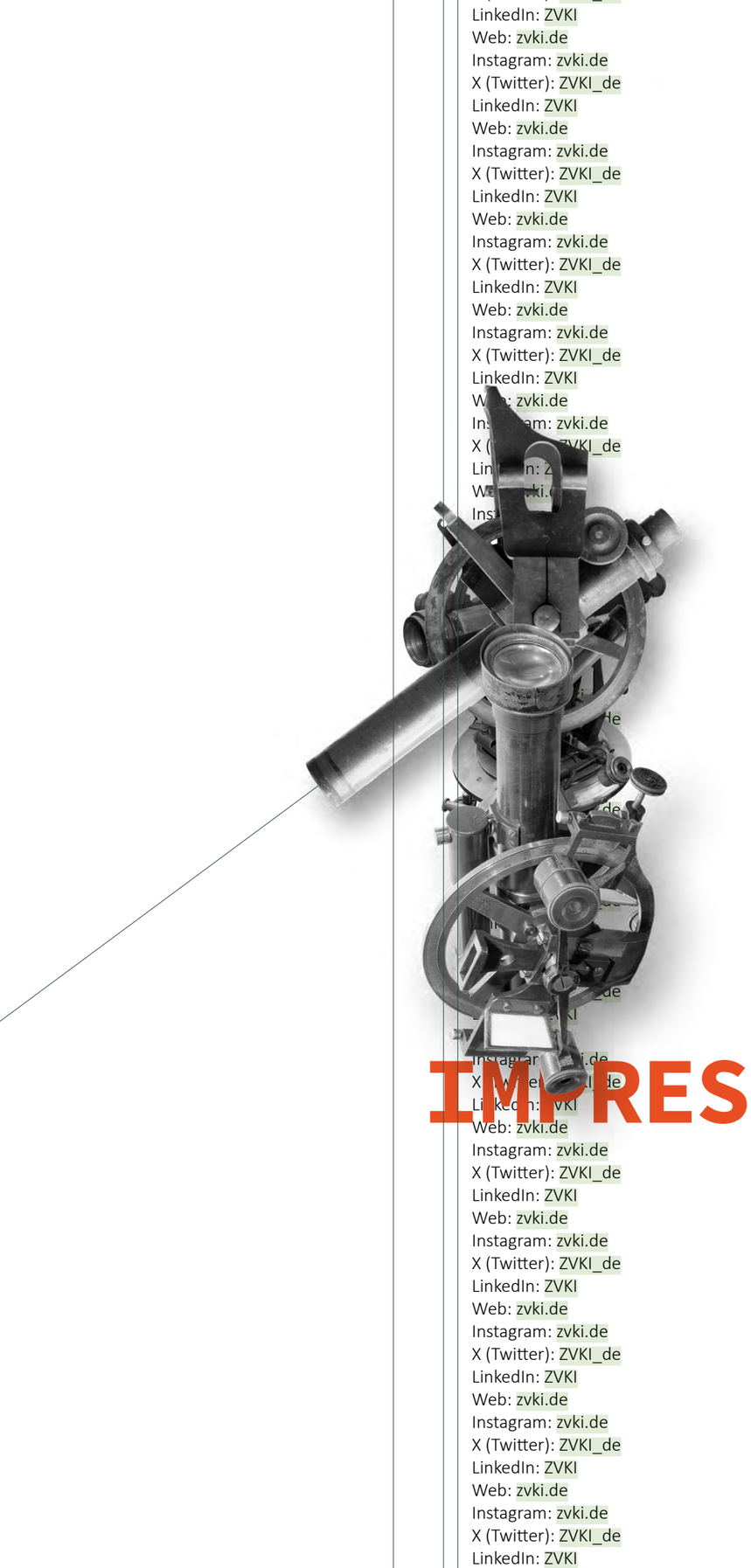
Vinuesa, Ricardo/ Azizpour, Hossein/ Leite, Iolanda/ Balaam, Madeline/ Dignum, Virginia / Domisch, Sami/ Felländer, Anna u. a. (2020): The Role of Artificial Intelligence in Achieving the Sustainable Development Goals. In: Nature Communications 11 (1), S. 1–10. https://doi.org/10.1038/s41467-019-14108-y.

WBGU – Wissenschaftlicher Beirat der Bundesregierung Globale Umweltveränderungen (Hrsg.) (2019): Hauptgutachten. Unsere gemeinsame digitale Zukunft. 12. April 2019. https://www.wbgu.de/fileadmin/user_upload/wbgu/publikationen/hauptgutachten/hg2019/pdf/wbgu_hg2019.pdf.

Wiggers, Kyle (2023): Stability AI, Hugging Face and Canva back new AI research nonprofit. TechCrunch. 2. März 2023. https://tcrn.ch/3Zutx74.

Wilke, Maik (2023): Wie Hochschulprüfer auf KI-geschriebene Hausarbeiten reagieren wollen. Südwest Presse. 16. März 2023. https://www.swp.de/lokales/albstadt/hochschule-albstadt-sigmaringen-chatgpt_-soll-ich-meine-hausarbeit-selber-schreiben_-69714715.html.

Zhang, Jizhi/ Bao, Keqin/ Zhang, Yang/ Wang, Wenjie/ Feng, Fuli und He, Xiangnan (2023): Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation. 3. Juli 2023. http://arxiv.org/abs/2305.07609.



IMPRESSUM

Herausgeber und inhaltlich Verantwortlicher i. S. d. § 55 Abs. 2 RStV:
Philipp Otto

iRights.Lab GmbH
Oranienstr. 185
D-10999 Berlin
Telefon: +49 (0)30 40 36 77 230
Fax: +49 (0)30 40 36 77 260
E-Mail: **zvki@irights-lab.de**

Geschäftsführer*in: Philipp Otto, Dr. Wiebke Glässer
Registergericht: Amtsgericht Berlin-Charlottenburg
Registernummer: HRB 185640 B
Finanzamt für Körperschaften II
USt-IdNr.: DE311181302

Projektleitung: Philipp Otto

Autor*innen:

Dr. Gergana Baeva, Matthieu Binder, Franziska Busse, Jaana Müller-Brehm, Merlin Münch, Paul Ritzka

Chefredaktion: Jaana Müller-Brehm

Redaktion: Dr. Gergana Baeva, Matthieu Binder, Merlin Münch, Philipp Otto, Verena Till

Inhaltliche Mitarbeit: Michael Puntschuh, Paul Ritzka

Gestalterische Konzeption und Layout: Christoph Löffler

Lektorat: text|struktur

Die **Online-Version** von *Missing Link* und weitere Informationen zum Projekt *ZVKI* finden Sie unter: **www.zvki.de**.

Weitere Informationen zu diesem und anderen Themen gibt es auf unseren Social-Media-Kanälen:
Instagram: **zvki.de**
X (Twitter): **ZVKI_de**
LinkedIn: **ZVKI**

Projektpartner*innen ZVKI: *iRights.Lab*, *Fraunhofer AISEC*, *Fraunhofer IAIS*, *Freie Universität Berlin*

Gefördert durch: *Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz (BMUV)*

Das Projekt *ZVKI* wird vom unabhängigen Think Tank *iRights.Lab* verantwortet und durchgeführt. Das *iRights.Lab* entwickelt Strategien und praktische Lösungen, um die Veränderungen in der digitalen Welt vorteilhaft zu gestalten. Wir unterstützen öffentliche Einrichtungen, Stiftungen, Unternehmen, Wissenschaft und Politik dabei, die Herausforderungen der Digitalisierung zu meistern und die vielschichtigen Potenziale effektiv und positiv zu nutzen.

Weitere Informationen über das *iRights.Lab* finden Sie unter **www.irights-lab.de**.

Gefördert durch:

Bundesministerium
für Umwelt, Naturschutz, nukleare Sicherheit
und Verbraucherschutz

aufgrund eines Beschlusses
des Deutschen Bundestages

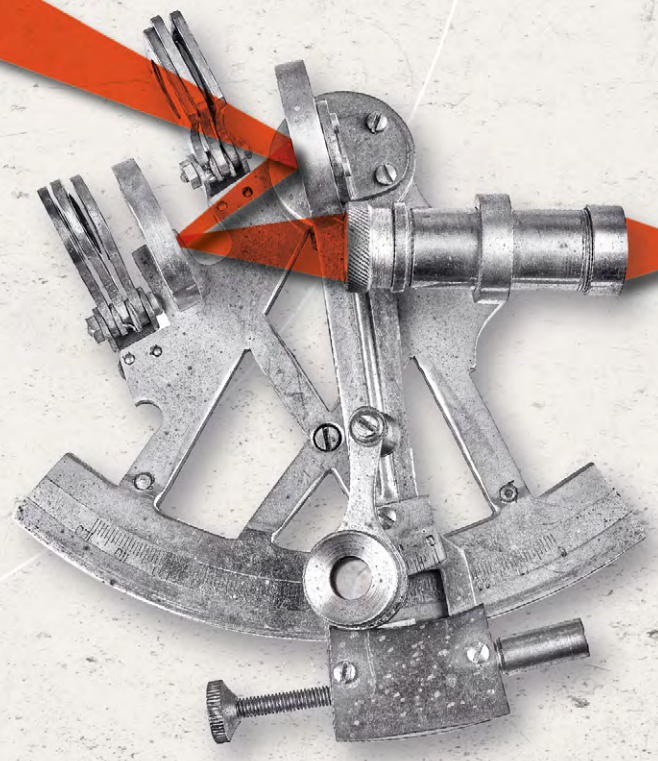
iRights.Lab
Think Tank für die digitale Welt

ISSN (Online) 2751-2207

Missing Link

Magazin für vertrauenswürdige Künstliche Intelligenz

Foundation Models – KI zwischen Wachstum und Nachhaltigkeit



Heft #4 / Oktober 2023