iit perspektive

Working Paper des Instituts für Innovation und Technik (iit) Nr. 69

Was bedeutet Open Source für Künstliche Intelligenz (KI)?

Ansätze für Unternehmen im Sinne einer intelligenten, offenen und vernetzten Gesellschaft

Sabine Weber, Juliane Hanel

iit perspektive

Working Paper des Instituts für Innovation und Technik (iit) Nr. 69

Sabine Weber, Juliane Hanel

Was bedeutet Open Source für Künstliche Intelligenz (KI)? – Ansätze für Unternehmen im Sinne einer intelligenten, offenen und vernetzten Gesellschaft

Einleitung: KI-Modelle auf dem Vormarsch

Amazon Alexa legt Musik auf und startet die Kaffeemaschine, Google Maps berechnet den schnellsten Weg zur Arbeit und ChatGPT formuliert schon mal die ersten E-Mails: Mit smarten Assistenzsystemen fühlt es sich so an, als sei man endlich in der Zukunft angekommen. Obwohl die Schwelle zur technologischen Singularität noch lange nicht überschritten ist, ist der Einfluss der künstlichen Intelligenz (KI) spätestens seit dem Aufkommen von Chat-GPT spürbar. Insbesondere die großen KI-Modelle, sogenannte Foundation Models, haben die KI-Landschaft in den vergangenen Jahren grundlegend verändert. Foundation Models unterscheiden sich technologisch auf den ersten Blick nicht von herkömmlichen Machine-Learning-Modellen, da sie wie diese auf Basis von tiefen neuronalen Netzen und selbstüberwachtem Lernen arbeiten. Dennoch wird bei dieser neuen Generation von KI-Modellen immer häufiger von einem Paradigmenwechsel gesprochen. Zentral ist hier der Umfang der Datensätze, auf denen sie trainiert werden, also der Umgang mit dem „Wissen", das dem Modell zur Verfügung steht. Durch das Training auf einer sehr großen Menge öffentlich verfügbarer Daten aus dem Internet erkennt das Modell statistische Zusammenhänge und Muster aus einer Vielzahl unterschiedlicher Quellen.

Wenn ein Modell auf einem derart großen Datensatz trainiert wurde, kann man es mit geringem Ressourcenaufwand an spezifische Wissensdomänen angepassen. Während zuvor das ressourcenintensive Training der Foundation Models Rechenleistung und Datensätze in einem Umfang erfordert, der – mit wenigen Ausnahmen – nur großen Technologieanbietern wie Meta oder Google vorbehalten gewesen ist, kann der relativ geringe Aufwand für das Anpassen an ein spezifisches Problem nun mit Standard-Machine-Learning-Hardware durchgeführt werden. Dies eröffnet neue Perspektiven für Anwender:innen, da die Einstiegshürde für die Anwendung innovativer KI-Methoden gesenkt wird.

Andererseits bedeutet dies auch, dass einige wenige meist wirtschaftlich agierenden Akteure maßgeblich für die Entwicklung der KI-Landschaft verantwortlich sind, weil sie ihre vortrainierten Modelle veröffentlichen und damit die Basis für den Großteil der modernen Machine-Learning-Anwendungen bereitstellen. Die KI-Community diskutiert dieses Potenzial für weitere Monopolisisierung kritisch: Wenn es um öffentliche Interessen geht, sollten dann nicht eher öffentliche Kontrollinstanzen als gewinnorientierte Unternehmen das Sagen haben? Expert:innen hinterfragen vor allem die Rolle von politischen Entscheidungsträger:innen und Zivilgesellschaft im Prozess der Marktkonzentration.

Teile der Community sehen eine große Chance in Open-Source-Ansätzen für KI. In einem Policy Paper appellieren beispielsweise einige der größten Stakeholder, darunter GitHub und Creative Commons, an die EU und erläutern die zentrale Rolle, die Open Source ihrer Meinung nach bei der KI-Regulierung durch den AI-Act spielen wird: „Erstens sind die Werte der soliden Forschung, der Reproduzierbarkeit und der Transparenz, die durch die offene Wissenschaft gefördert werden, für die Entwicklung sicherer und verantwortungsvoller KI-Systeme von entscheidender Bedeutung. Zweitens kann die Open-Source-Entwicklung Wettbewerb und Innovation durch neue Marktteilnehmer und kleinere Akteure, auch in der EU, ermöglichen." (CreativeCommons, et al., 2023). Die Stakeholder sehen daher in Open Source Software einen Katalysator für verantwortungsbewusste Innovation, der die Grundlage für einen chancengleichen Wettbewerb bildet.

Die Herausforderung: Open Source für KI

Open Source, also quelloffene Software, ist in vielen Bereichen mittlerweile fester Bestandteil des professionellen und privaten Alltags. Die Philosophie der freien Software hat in den 1980er-Jahren als Bewegung gegen proprietäre Software angefangen und beruht auf vier wesentlichen Freiheiten (FreeSoftwareFoundation, 2023):

1. Die Freiheit, das Programm auszuführen wie man möchte, für jeden Zweck.
2. Die Freiheit, die Funktionsweise des Programms zu untersuchen und eigenen Datenverarbeitungbedürfnissen anzupassen.
3. Die Freiheit, das Programm zu redistribuieren und damit Mitmenschen zu helfen.
4. Die Freiheit, das Programm zu verbessern und diese Verbesserungen der Öffentlichkeit freizugeben, damit die gesamte Gesellschaft davon profitiert.

Diese auf den ersten Blick utopischen Prinzipien werden heute nicht mehr vorwiegend mit dem Idealismus ihrer Anfänge verfolgt, sondern haben sich auch in der freien Marktwirtschaft bewährt. Webbrowser wie Mozilla Firefox und Betriebssysteme wie Android basieren beispielsweise auf frei zugänglicher Software und profitieren somit auch selbst von der dezentralen und kollaborativen Entwicklung durch interessierte Communities. Diese Communities melden und beheben Fehler im Quellcode, entwickeln die Software weiter und sorgen für größtmögliche Transparenz und Qualität.

Besonderheiten von KI im Kontext von Open Source
Wie lassen sich diese zwei Konzepte, Foundation Models und Open Source, verbinden? Können die Vorteile des Open-Source-Prinzips auf die neue Generation der Machine-Learning-Modelle angewendet werden? Bei näherer Betrachtung fallen einige grundlegenden Unterschiede zwischen Machine-Learing-Modellen und herkömmlicher Software ins Auge, die vor allem den Umfang der Faktoren sowie die Fehlerbehebung betreffen:

KI ist mehr als nur Code: Der Code, der benutzt wird, um ein Modell zu trainieren ist zwar wichtig für das spätere Verhalten des Modells, doch spielen andere Faktoren eine ebenso große Rolle. Während es bei Open-Source-Software nur ein zu veröffentlichendes Artefakt gibt (den Code der Software), sind es im Falle von KI bis zu fünf:

1. Code, um ein Modell zu trainieren,
2. Trainingsdaten,
3. Testdaten,
4. Metadaten über das Trainingsregime und
5. das fertig trainierte Modell.

KI-Modelle sind schwieriger von Fehlern zu bereinigen: Anders als bei Software kann unerwünschtes Verhalten bei einem Modell nur durch erneutes Trainieren des Modells mit geänderten Daten oder Modellparametern behoben werden. Des Weiteren lässt es sich nicht immer eindeutig ermitteln, was die Ursache von Fehlverhalten ist. Der Entwicklungszyklus eines Machine-Learing-Modells beinhaltet daher oft das Trainieren von Modellen mit verschiedenen Architekturen und Trainingsdaten.

Um diesen Unwägbarkeiten des Entwicklungsprozesses zu begegnen und um verschiedene Modelle besser untereinander vergleichen zu können, gibt es Bemühungen um eine standardisierte Dokumentation. Data Statements (Bender & Friedman, 2018) und Model Cards (Mitchell, 2019) sind zwei beispielhafte Wege, die dabei helfen, Klarheit über die einzelnen Bestandteile eines Modells zu schaffen. Dennoch hat sich noch kein einheitlicher Dokumentationsstandard durchgesetzt.

Open Source: Verschiedene Wege der Veröffentlichung
Um am wissenschaftlichen Diskurs teilzunehmen, veröffentlichen Universitäten in der Regel ihren Code und ihre Erkenntnisse in Publikationen für Konferenzen, in wissenschaftlichen Zeitschriften und auf pre-print-Portalen wie arXiv. Für Veröffentlichungen zu Konferenzen oder in wissenschaftlichen Zeitschriften liegt die Entscheidung ob ein Modell ausreichend dokumentiert ist, stets bei einer von der jeweiligen Publikation ausgewählten Gruppe aus mehreren Wissenschaftler:innen, die die Arbeiten auf Richtigkeit überprüfen, also im sogenannten Peer-Review.

Anders als wissenschaftliche Akteure unterliegen Unternehmen jedoch keinem Publikationsdruck. Hier sind andere Faktoren entscheidend darüber, was veröffentlicht wird. Wenn ein Unternehmen möchte, dass Nutzende für den Zugang zu dem Modell bezahlen, dann wäre es kontraproduktiv, das Modell kostenlos zur Verfügung zu stellen. Soll verhindert werden, dass das Modell von Nutzenden mit eigenen Daten weiter trainiert wird, so wäre es nicht ratsam, Informationen über Trainingsregime zu veröffentlichen. Und wenn die Daten, auf denen das Modell trainiert wurde, einen Wettbewerbsvorteil gegenüber Mitwettbewerbern darstellen können, so besteht auch dort ein Interesse an der Geheimhaltung – insbesondere, wenn zusätzlich urheberrechtliche Bedenken vorliegen.

Es besteht also ein Spannungsverhältnis zwischen den Interessen der Unternehmen, die die Modelle herstellen, und den Menschen, die sie nutzen wollen. Auf der einen Seite steht das Bedürfnis der Unternehmen, ihr geistiges Eigentum und ihre Produkte vor Nachahmung und Diebstahl zu schützen. Auf der anderen Seite steht das Bedürfnis der Öffentlichkeit nach Auditierbarkeit und dem Aufdecken möglicher Fehler. Doch auch

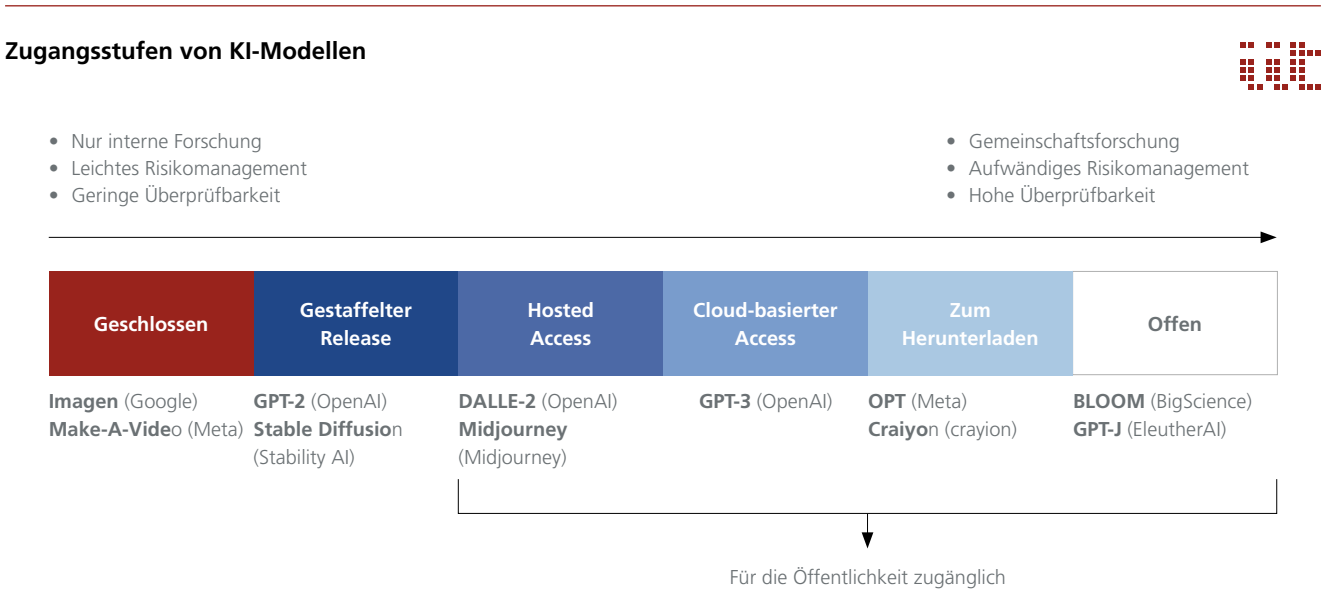

Abb. 1: Zugangsstufen von KI-Modellen, eigene Darstellung übersetzt und adaptiert basierend auf Solaiman 2023.

die Unternehmen müssen abwägen, wie sehr sie von der Einbindung der Öffentlichkeit profitieren könnten, zum Beispiel durch neue Perspektiven von Forschenden außerhalb des Unternehmens, Feedback durch eine Open-Source-Community und unter Umständen sogar kostenlose Arbeit freiwilliger Entwickler:innen.

Die Tatsache, dass es bei KI-Systemen nicht nur *ein* Artefakt, sondern mehrere gibt, ermöglicht einen fließenderen Übergang zwischen den zwei Extremen von Open und Closed Source. Eine mögliche Einteilung sieht drei Kategorien der Veröffentlichung vor: Das Modell, Artefakte zur Ermöglichung einer weiteren Risikoanalyse und Artefakte zur Modellreplikation (Solaiman, 2023). Diese Kategorien überlappen sich teilweise, da zum Beispiel Trainingsdaten sowohl für die Replikation eines Modells als auch für die Analyse von Biases, also systematischer Vorurteile und Verzerrungen eines Systems bestimmten Sachverhalten oder Personengruppen gegenüber, genutzt werden können (siehe Abbildung 1). Generell beziehen sich Artefakte für eine weitere Risikoanalyse aber eher auf inhaltliche und soziale Faktoren der Trainingsdaten, während bei Komponenten für die Modellreplikation eher technische Gesichtspunkte im Vordergrund stehen. Unternehmen können diese verschiedenen Strategien nutzen, um die Auditierbarkeit zu verbessern, ohne wirtschaftlich relevante Daten preiszugeben.

Die zeitliche Achse fügt eine weitere Dimension hinzu: Ein KI-System kann zum Beispiel zunächst nur über eine Schnittstelle zugänglich gemacht werden, und erst zu einem späteren Zeitpunkt werden das Modell und die Trainingsdaten veröffentlicht. Die Zeit zwischen den einzelnen Phasen der Veröffentlichung kann von Unternehmen genutzt werden, um Risiken zu minimieren, die Aktivitäten potenzieller Angreifender zu überwachen und mögliche Schäden zu untersuchen. Dennoch ist dieses Vorgehen kein Patentrezept: So kommt es zum Beispiel immer wieder zu ungewollten Veröffentlichungen von Modellen durch Dritte vor dem geplanten Zeitplan, sogenannten Leaks.

Das Center for Research on Foundation Models an der Stanford University hat mit seinem Foundation Model Transparency Index (Bommasani, et al., 2023) erstmals eine Reihe von großen Machine-Learning-Modellen auf verschiedene Kriterien der Transparenz hin untersucht. Die Forschenden definierten dabei 100 Indikatoren in 13 verschiedenen Kategorien wie Daten, Rechenleistung und Methoden. Aber auch gesellschaftliche Faktoren wie Arbeitsaufteilung, Nutzung und Zugang zu dem Modell werden einbezogen. Der Index bietet vor allem denjenigen eine Übersicht, die sich bei der Wahl eines großen Sprachmodells an Open-Source-Prinzipien orientieren wollen. Am besten schneidet dabei Metas Llama 2 und das als Open-Science-Projekt entwickelte BLOOMZ ab[1]. Allerdings erreichte keines der audi-

1 Für mehr Informationen: https://crfm.stanford.edu/fmti/

tierten Modell die volle Punktzahl. Dies zeigt, dass zurzeit auch Open-Source-Projekte nicht alle möglichen Aspekte der Transparenz abdecken.

Einbindung der Community über die Veröffentlichung hinaus

Die Veröffentlichung ist nicht der einzige Kanal, über den die breite Öffentlichkeit in die Entwicklung von KI-Systemen eingebunden werden kann. Einerseits gibt es das Argument, dass modernes Machine Learning ein inhärent kooperatives Unterfangen ist, da Trainingsdatensätze oft von vielen global weit verteilten Menschen – sogenannten Crowdworker:innen – über Plattformen wie Amazon Mechanical Turk erstellt werden. Andererseits aber haben die einzelnen Crowdworker:innen sehr wenig Einfluss auf die entstehenden Modelle, da die Entscheidungsgewalt stets bei den beauftragenden Unternehmen liegt. In ihrer idealen Form sollte kollaborative Entwicklung eine Bereicherung für beide Seiten darstellen und kein unilaterales Abhängigkeitsverhältnis schaffen. Im besten Fall führt sie so nicht nur zu besseren Modellen, sondern zu individueller und kollektiver Befähigung der Beteiligten und zu positiven sozialen und strukturellen Veränderungen.

In ihrem Paper untersuchen Birhane et al. (2022) beispielhaft Fälle von kollaborativer KI-Entwicklung. Ein prominentes Beispiel ist Masakhane, eine Graswurzelbewegung, die sich online zusammengefunden hat, um die maschinelle Übersetzung für afrikanische Sprachen zu ermöglichen, für die es noch keine kommerziellen Übersetzungssysteme gibt. Die Gemeinschaft, deren Name aus isiZulu übersetzt so viel wie „Wir kreieren gemeinsam" bedeutet, besteht aus mehr als tausend Freiwilligen aus 30 afrikanischen Ländern, die seit der Gründung der Bewegung Übersetzungssysteme für 38 afrikanische Sprachen erstellt und veröffentlicht haben. In einem ähnlichen Projekt hat die Gemeinschaft der Maori in Neuseeland Sprachdaten zur Entwicklung von Sprachsystemen gesammelt. Durch ein Data Sovereignty Protocol konnte die Gemeinschaft sicherstellen, dass kommerzielle Akteure die Daten nicht nutzen dürfen. Dadurch soll garantiert werden, dass die Daten nur zum Vorteil der Maori-Gemeinschaft eingesetzt werden können. In beiden Beispielen liegt der volle Entwicklungszyklus von Machine-Learning-Modellen in der Hand der entsprechenden Sprachcommunity. Es wurden explizit nicht nur Menschen mit Kenntnissen in Informatik und Machine Learning, sondern ein weites Spektrum aus verschiedenen gesellschaftlichen Bereichen involviert.

Diese Fälle können als Goldstandard für die kollaborative Entwicklung von KI gelten. Aber auch große Technologiefirmen wählen öffentlichkeitswirksame Events, um größere Nutzergemeinschaften in die Entwicklung von KI einzubeziehen. Die selbstorganisierte Interessengemeinschaft AI Village veranstaltete beispielsweise auf der Hackermesse DEF CON 31 zusammen mit anderen KI-Ethik-fokussierten Organisationen das erste öffentliche Red Team Event für generative KI. Dabei wurden große Sprachmodelle von Anthropic, Google, Hugging Face, NVIDIA, OpenAI und Stability unter Beteiligung von Microsoft auf einer von Scale AI entwickelten Bewertungsplattform von Freiwilligen getestet. Ziel der verschiedenen Testszenarien war es, zum Beispiel Texteingaben zu schreiben, die das Modell dazu bringen sollten, falsche Aussagen zu treffen, rassistische Vorurteile zu zeigen oder Rechenaufgaben falsch zu lösen.

Ein zentrales Spannungsfeld dieser jüngsten Welle der Partizipation ist die Frage, ob die dabei entstehenden Mechanismen lediglich dazu dienen, relevante maschinelle Lernsysteme zu verfeinern oder ob sie einen umfassenderen humanitären oder gesellschaftlichen Nutzen gewährleisten. Das Red Team Event ist ein Beispiel für sogenannte „Bias Bounties", in denen Unternehmen Preise ausschreiben, die Freiwillige dazu motivieren sollen, schädliche Biases in Machine-Learning-Systemen zu finden. Bias Bounties ermöglichen zwar ein Feedback von marginalisierten Gruppen, aber die Entscheidung darüber, ob und wie das Feeback umgesetzt wird, liegt stets bei den Unternehmen. Diese scheinbare Beteiligung marginalisierter Gruppen ohne klare Verantwortungsstrukturen bei den Unternehmen wird (analog zum „green-washing") als „participation-washing" bezeichnet (Dennler, et al., 2023). Kerngedanke der kollaborativen KI-Entwicklung ist: Wenn sich ein KI-System auf eine Gemeinschaft auswirkt, dann sollte diese Gemeinschaft ein Stakeholder in der Entwicklung und Evaluation des Systems sein. Bisher sind große Technologiefirmen nicht bereit, diesen Schritt zu gehen.

Die Beteiligung von vielfältigen Interessengruppen und der breiten Öffentlichkeit in Open-Source-Ansätzen verspricht eine Reihe von Vorteilen, birgt jedoch auch Herausforderungen und potenzielle Gefahren. Was sind Chancen und Risiken von Open Source in der KI und was bedeutet das für Unternehmen in Deutschland?

Chancen und Risiken von Open Source

Die Idee scheint zuerst kontraintuitiv: Software, deren Code für alle offen zugänglich ist, ist sicherer als Software, deren Code nur innerhalb des entwickelnden Unternehmens bekannt ist. Öffnet eine Publikation des Quellcodes nicht vielmehr Tor und Tür für Hacker:innen, die Schwachstellen so leichter finden und ausnutzen können? Die Entwicklungen der vergangenen zehn Jahre zeichnen ein anderes Bild. Open-Source-Software hat sich von einer akademischen Nische zu einer übergreifenden Strategie entwickelt und gilt heute in vielen Bereichen als Standard.

Es sind vor allem drei Punkte, die Open Source mit Hinblick auf Sicherheitsbedenken so vielversprechend machen:

- **Transparenz:** Nicht nur Hacker:innen können den Quellcode lesen, sondern auch diejenigen, die den Code aktiv nutzen und an seiner Sicherheit interessiert sind. Je erfolgreicher eine Software, desto mehr Nutzende gibt es und desto höher die Wahrscheinlichkeit, dass diese Schwachstellen finden, bevor es Hacker:innen tun.

- **Unabhängigkeit**: Ein Unternehmen mag Interesse daran haben, Sicherheitslücken zu vertuschen, um kostspielige Reparaturen am Code zu vermeiden. Eine weltweite Community aus Freiwilligen hat diese Motivation nicht und ist so besser in der Lage Fehler aufzuzeigen.

- **Schnelle Fehlerbehebung:** Die Open-Source-Community ist um ein vielfaches größer als die meisten Softwareunternehmen. Wenn Freiwillige Fehler ausfindig machen und beheben, werden Sicherheitslücken schneller geschlossen, als es ein Unternehmen mit fest angestellten Programmierer:innen bewerkstelligen könnte.

Doch es gibt auch klare Vorbehalte: Open-Source-Software gewährt geistigem Eigentum keinen Schutz. Unternehmen müssen neue Strategien finden, um Gewinne zu erzielen, wenn ihr Produkt eine Open-Source-Software ist. Und eine aktive Community für ein Open-Source-Projekt aufzubauen kostet Zeit. Schließlich ist das Ziel, dass Menschen unbezahlt und aus eigenem Antrieb an dem Code arbeiten und ihn gegen immer neue Angriffe absichern. Es ist nicht der offene Quellcode selbst, der eine Verbesserung der Sicherheit bewirkt, sondern die Menschen, die mit ihm interagieren.

Was bedeutet es also, wenn Open-Source-Prinzipien auf die Herausforderungen der Entwicklung von KI-Systemen treffen?

Chancen

Eines der größten Versprechen der Open-Source-Bewegung ist die Demokratisierung. Die Verantwortung für die Entwicklung und Wartung wichtiger Werkzeuge liegt im Fall von Open-Source-Software nicht mehr bei großen Unternehmen, sondern bei den Menschen, die die Werkzeuge nutzen. Damit wird einer Machtkonzentration und Monopolstellung von wenigen Akteuren entgegengewirkt. Mit der zunehmenden Bedeutung von KI ist diese Idee wichtiger denn je: Wenn KI Systeme jeden Bereich der Gesellschaft beeinflussen, sollte die Macht über sie nicht bei wenigen Großunternehmen liegen.

Aber die Natur des maschinellen Lernens und vor allem der großen Sprachmodelle (Large Language Models, LLMs) führt fast automatisch zu einer Machtkonzentration. Um leistungsstarke LLMs zu trainieren, braucht man große Mengen an Daten und einen Zugang zu performanten Rechenzentren: So etwas können in der Regel nur sehr große Unternehmen wie Google, Meta oder Microsoft leisten. Schon seit einigen Jahren dominieren die Forschungsabteilungen dieser Unternehmen den wissenschaftlichen Diskurs zum Thema Machine Learning, weil Universitäten und staatliche Forschungseinrichtungen finanziell nicht mithalten können. Prominente Forscher:innen stellen zudem immer wieder den Einfluss von kommerziellen Akteuren in der universitären Forschung in Frage (Abdalla & Abdalla, 2021). Sie fürchten, dass kommerziell dominante Technologien von der universitären Forschung nicht kritisch genug untersucht werden. Langfristig könnte dies dazu führen, dass nur jene Technologien erforscht werden, die für die Entscheidungsträger:innen in den Unternehmen monetarisierbar erscheinen, während eigentlich notwendige Grundlagenforschung in andere Richtungen vernachlässigt wird.

Während Universitäten dazu verpflichtet sind, Forschungsergebnisse der Öffentlichkeit zugänglich zu machen, sind diese Ergebnisse bei weitem nicht die einzigen Quellen von Open-Source-KI-Projekten. Die Entwicklung von KI ist schon seit geraumer Zeit von sehr erfolgreichen Open-Source-Projekten geprägt. Im Folgenden werden vier Projekte besonders in den Fokus gerückt, die die Chancen für Open-Source-KI verdeutlichen.

1) Bau dir dein eigenes neuronales Netzwerk: PyTorch und TensorFlow

Sowohl PyTorch als auch TensorFlow sind Open-Source-Bibliotheken in der Programmiersprache Python. Die Bibliotheken bieten baukastenartige Elemente, die es ermöglichen, verschiedene KI-Architekturen zu implementieren und zu testen, ohne dabei ihre mathematischen Grundstrukturen ausformulieren zu müssen. Beide Bibliotheken sind ursprünglich von großen Softwareunternehmen, nämlich Google und Meta, entwickelt und dann als Open-Source-Bibliotheken veröffentlicht worden, was die Weiterentwicklung von KI Systemen erheblich beschleunigt hat. Beide Frameworks haben große und aktive Benutzer:innen-Communities, was den Austausch von Wissen, bewährten Methoden und Ressourcen fördert.

2) Lade bereits trainierte Modelle herunter: Hugging Face

Es ist Teil des Open-Source-Gedankens, dass nicht nur der Code für den Bau von Modellen, sondern auch die fertig trainierten Modelle frei zur Verfügung stehen sollten. Die Open-Source-Community um das Unternehmen Hugging Face bietet sowohl vortrainierte Modelle als auch Python-Bibliotheken, die es ermöglichen, die vortrainierten Modelle einzubinden. Des Weiteren stellt Hugging Face auch eine Auswahl aus öffentlichen Datensätzen zum Trainieren und Testen von Modellen zur Verfügung. Hugging Face entwickelte außerdem die Transformers-Bibliothek, eine der am weitest verbreiteten Bibliotheken zur Arbeit mit Transformer-basierten Modellen, zu denen unter anderem die Modelle der GPT-Familie zählen.

Einschätzungen zu Risiken und Hinderungsgründen beim Einsatz von KI

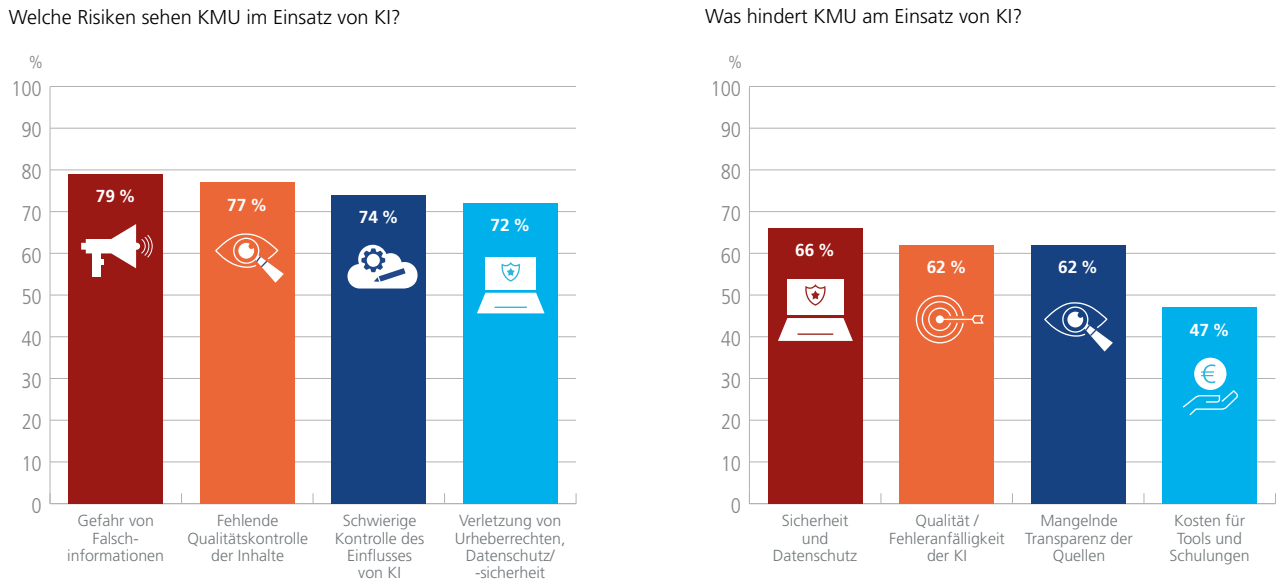

Abb. 2 a und b: Einschätzungen zu Risiken und Hinderungsgründen beim Einsatz von KI gemäß einer Online-Umfrage der YouGov Deutschland GmbH unter ca. 1.000 Entscheider:innen aus Unternehmen mit bis zu 250 Mitarbeitenden, eigene Darstellung adaptiert von YouGov Deutschland (2023).

3) Große Sprachmodelle für alle: EleutherAI

Die selbstorganisierte Non-Profit-Organisation EleutherAI entstand als Reaktion auf das Erscheinen von GPT-3. Zu diesem Zeitpunkt gab es noch kein Open-Source-basiertes großes Sprachmodell der GPT-Familie und die Gründer von EleutherAI beschlossen, das zu ändern. Der Grundsatz der Organisation baut auf der Überzeugung auf, dass eine so einflussreiche Technologie wie große Sprachmodelle breiten Gesellschaftsschichten zugänglich sein sollte. Der Zugang zu EleutherAI ist niederschwellig: Es gibt einen Discord-Server, bei dem jede:r auch ohne Vorwissen Mitglied werden und sich in die Konversation einbringen kann. Zu den Erfolgen der Organisation gehören die Veröffentlichungen einiger der zurzeit stärksten Modelle für Sprachverarbeitung, Text-zu-Bild-Anwendungen und Bioinformatik: GPT-J, GPT-NeoX, BLOOM, VQGAN-CLIP, Stable Diffusion und OpenFold. Außerdem bringt sich die Organisation durch Publikationen auf internationalen KI-Konferenzen in den wissenschaftlichen Diskurs ein. EleutherAI sieht sich als Vorreiter auf dem Gebiet der Open-Source-LLMs und konzentriert sich nun auf Forschung zur Interpretabilität und Sicherheit von KI-Modellen.

Risiken

Obwohl die im vorigen Abschnitt gelisteten Projekte Erfolgsgeschichten erzählen, gibt es auch viele Bedenken, was den Einsatz von KI betrifft (siehe Abbildung 2). Ein kritischer Punkt: Was soll passieren, wenn Open-Source-Entwickler: innen schädliche Biases oder Sicherheitsrisiken in einem Modell finden? Programmcode lässt sich unter Umständen schnell ändern, aber ein KI-Modell mit neuen Daten oder sogar einer neuen Architektur zu trainieren ist kostspielig und kann oft nicht von einzelnen Akteuren in Angriff genommen werden. Momentan ergreifen Unternehmen, die LLMs veröffentlichen, eine Reihe von Sicherheitsmaßnahmen, um eine unerwünschte Nutzung ihrer Systeme zu vermeiden. Diese Sicherheitsmaßnahmen spannen einen weiten Raum zwischen rein technischen Lösungen wie der Begrenzung der Anzahl der Zugriffe auf ein System bis hin zu vor- und nachgeschalteten Machine-Learning-Modellen, die die Eingaben der Nutzenden und die Ausgaben des Modells auf schädliche Aussagen prüfen.

Eine weitere häufige Vorsichtsmaßnahme sind regel- und stichwort-basierte Filter, die bei verdächtigen Eingaben zu einer leeren oder vorbestimmten Antwort des Systems führen. Eine häufige Antwort dieser Art ist zum Beispiel „Ich bin nur ein Sprachmodell und kann keine Aussagen zu diesem Thema machen". Da diese Filter händisch eingestellt werden müssen, spiegeln sie das Werteverständnis der entsprechenden Unternehmen oder Communities wider, aus denen sie stammen. So verhinderte beispielsweise anfänglich das Bildgenerierungssystem Stable Diffusion das Generieren sexuell expliziter Bilder, aber nicht das Generieren gewaltverherrlichender Bilder – ein

Feature, das erst nach öffentlicher Kritik angepasst wurde. Wenn Filter gesellschaftlich sensitive Inhalte blockieren, kann das auch dazu führen, dass marginalisierte Gruppen von der Nutzung der Technologie ausgeschlossen werden. So blockierten die großen Sprachmodellen anfänglich auch solche Eingaben, die Identitätsbegriffe wie „lesbisch" oder „schwul" enthielten.

Laut Erkenntnissen von Forschenden der Carnegie Mellon University kann Machine Learning auch eingesetzt werden, um gezielt Prompts, also direkte Handlungsanweisungen an die KI zur Erzielung bestimmter Ergebnisse, zu finden, die schädliche Modell-Outputs produzieren (Zou, Wang, Kolter, & Fredrikson, 2023). Die Ergebnisse zeigen auch auf, dass diese schädlichen Prompts sogar zwischen Modellen übertragbar sind. So können öffentlich zur Verfügung stehende LLMs genutzt werden, um ein Modell zu trainieren, das schädliche Prompts erzeugt, die nicht nur bei dem entsprechenden Open-Source-Modell wirken, sondern auch bei Closed-Source-LLMs. Damit ermöglichen die offen verfügbaren LLMs Transferattacken auf kommerzielle LLMs. Die Forschenden empfehlen, dieses Wissen über mögliche Attacken explizit in das Training von neuen Modellen mit einzubeziehen. Generell lässt sich dieses Problem nicht allein durch die manuelle Anpassung von regelbasierten Filtern korrigieren, sondern erfordert ein erneutes Training des Modells sowie eine Änderung des Trainingsprotokolls und der Trainingsdaten.

Ein weiteres Sicherheitsproblem liegt in der individuellen Anpassung von offen zur Verfügung stehenden Modellen, dem sogenannten Fine-Tuning. Wenn ein Modell und sein entsprechender Trainingscode frei zur Verfügung stehen, kann jede:r Nutzende das Modell mit eigenen Daten weitertrainieren und für eine spezifische Anwendung optimieren. Das kann einerseits hilfreich sein, wenn man die breiten Fähigkeiten eines Sprachmodells auf eine bestimmte Domäne anpassen will, oder dafür sorgen will, dass das Modell für den eigenen Anwendungsfall besonders gut funktioniert. Akteure mit unlauteren Motiven können diese Funktionalität aber auch dafür nutzen, um schädliche Anwendungen zu produzieren, zum Beispiel um automatisiert Hassreden oder politische Propaganda zu produzieren, die dank der Fähigkeiten der Sprachmodelle überzeugend menschlich klingt und auf verschiedene Zielgruppen angepasst werden kann. Forschende der Stanford University schlagen als Gegenmaßnahme sogenannte „selbstzerstörende" Modelle vor, die auf ein Fine-Tuning für eine unterwünschte Anwendung mit einem großen Verlust an Outputqualität reagieren (Henderson, Mitchell, Manning, Jurafsky, & Finn, 2023). Dieser Ansatz steckt jedoch noch in den Kinderschuhen und es bleibt abzuwarten, wie gut er sich für verschiedene große Sprachmodelle und für mehrere unerwünschte Nutzungsszenarien gleichzeitig umsetzen lässt.

Ausblick

Trotz möglicher Risiken ist die Bilanz der Open-Source-Bemühungen im KI-Bereich durchaus positiv. Große Sprachmodelle und Foundation Models liegen nicht mehr allein in den Händen von wenigen Großunternehmen, die sie verkaufen wollen. Für viele Unternehmen hierzulande kommt es nicht in Frage, die eigenen Daten über eine Schnittstelle an Server in den USA zu schicken. Für sie bieten die vielen frei zur Verfügung stehenden Modelle eine gute Alternative zu kommerziellen Anbietern. Außerdem verfügt Deutschland über gute Ausgangsbedingungen: Die deutsche Sprache ist neben anderen europäischen Sprachen wie Spanisch und Französisch durchaus in den Trainingsdaten großer Sprachmodelle vertreten. Man muss also für deutschsprachige Anwendungen erst einmal nur mit geringen Qualitätseinbußen rechnen. An den meisten deutschen Unternehmensstandorten stehen gute Internetverbindungen zur Verfügung, die es ermöglichen, flexibel die Speicher- und Rechenleistung von Cloud-Dienstleistern zu beziehen, die zum Trainieren oder Nutzen von KI-Modellen notwendig sind.

Je nach Anwendungsfall eines Unternehmens und den zur Verfügung stehenden Ressourcen kann der komplette Entwicklungszyklus eines KI-Modells auf Open-Source-Basis ablaufen. Dabei kann ein Unternehmen sowohl Open-Source-Komponenten nutzen als auch eigene Ergebnisse unter Open-Source-Lizenzen veröffentlichen. Im einfachsten Fall kann man ein Modell über Hugging Face herunterladen und es unmodifiziert für weitere Aufgaben anwenden. Falls öffentlicher Trainingscode und vielleicht sogar Trainingsmeilensteine vorliegen, lässt sich ein Modell auch mit eigenen Daten im Sinne des Fine-Tunings anpassen. Falls sehr viele Trainingsdaten und das Know-how für Modellarchitekturen zur Verfügung stehen, können mit PyTorch oder Tensorflow eigene künstliche neuronale Netze konzipiert werden. Damit hat man dann alle Schritte selbst in der Hand: Datenauswahl, Modellarchitektur, Trainingsregime und die weitere Verwendung.

Das Spektrum zwischen Open und Closed Source ist somit sehr breit zu sehen. Unternehmen können selbst abwägen, wo sie ihre eigenen Produkte verorten. Sind sie daran interessiert, eine Community aus Freiwilligen um ihre Codebasis aufzubauen und unter Umständen statt einer Software oder dem Zugang zu einer Schnittstelle eher Services wie Wartung und individuell angepasste Lösungen anzubieten? Oder soll das auf eigenen Daten trainierte Modell das eigentliche Produkt sein, und sollen erneutes Training oder Fine-Tuning durch andere verboten werden? In jedem Fall sind dank der Entwicklungen im Bereich der Open-Source-KI nun auch kleinere Unternehmen in einer guten Startposition, um von den Möglichkeiten, die die KI eröffnet, zu profitieren und Wege zu finden, die neuen Technologien zu integrieren und Innovation voranzutreiben.

Einschätzung zur Nutzung von KI-Tools im eigenen Unternehmen

Kommt die Nutzung von KI Tools in Ihrem Unternehmen in Frage?

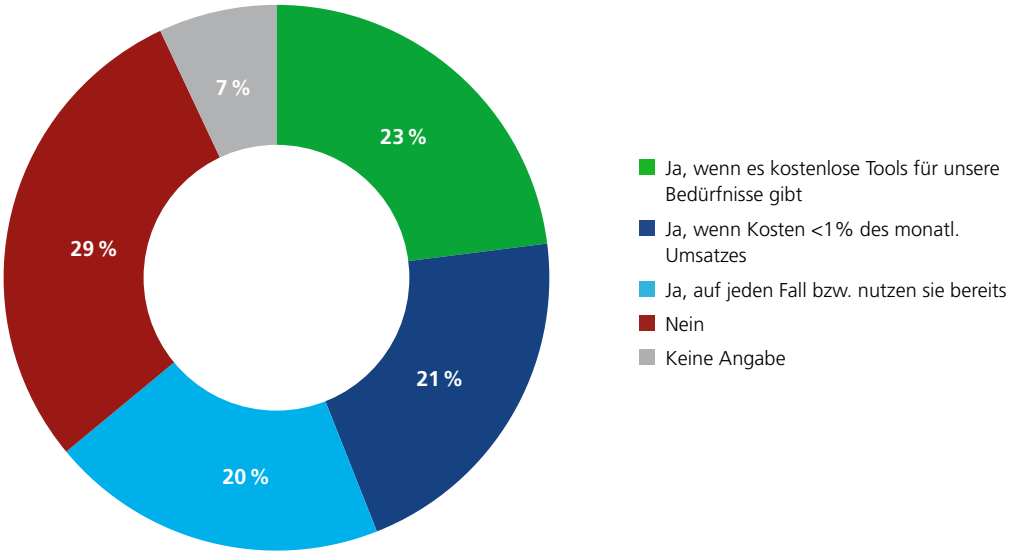

Abb. 3: Einschätzung zur Nutzung von KI-Tools im eigenen Unternehmen (eigene Darstellung, adaptiert von: https://www.presseportal.de/pm/160685/5619053, Online-Umfrage der YouGov Deutschland GmbH unter ca. 1.000 Entscheider:innen aus Unternehmen mit bis zu 250 Mitarbeitenden (August 2023)

Insbesondere bietet dies eine Chance für den deutschen Mittelstand, da durch die gesenkte Zugangshürde das Investitionskapital und somit auch das Risiko für die Erprobung innovativer KI-Technologien geringer ausfällt (siehe Grafik 3). Dank bestehender Open-Source-Lösungen können KI-Systeme auch mit knappen personellen Ressourcen implementiert werden. Darüber hinaus ermöglicht Open-Source-KI kleineren Unternehmen den Wettbewerb auf Augenhöhe mit größeren Konkurrenten, indem sie Zugang zu datengesteuerten Erkenntnissen und personalisierten Kundeninteraktionen verschafft. KI-gesteuerte Tools können auch dazu beitragen, den Kundenservice zu verbessern, Vertriebsprognosen zu optimieren und maßgeschneiderte Marketingstrategien zu entwickeln, die auf den individuellen Bedürfnissen der jeweiligen Zielgruppen basieren. Dieser dank Open-Source-Lösungen erschwingliche Zugang zu leistungsstarker Technologie kann kleineren Unternehmen helfen, sich in einem zunehmend wettbewerbsintensiven Markt zu behaupten und ihr Wachstumspotenzial zu maximieren.

Literatur

Abdalla, M., & Abdalla, M. (2021). The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, S. 287–297.

Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics 6*, S. 587–604.

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the people? opportunities and challenges for participatory AI. *Equity and Access in Algorithms, Mechanisms, and Optimization*, S. 1–8.

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., … Liang, P. (07. 11 2023). *The Foundation Model Transparency Index*. Von arXiv: https://arxiv.org/abs/2310.12941 abgerufen

CreativeCommons, EleutherAI, GitHub, Face, H., LAION, & OpenFuture. (07. 11 2023). *https://creativecommons.org*. Von https://creativecommons.org/wp-content/uploads/2023/07/SupportingOpenSourceAndOpenScienceInTheEUAIAct.pdf abgerufen

FreeSoftwareFoundation. (07. 11 2023). *https://www.gnu.org*. Von https://www.gnu.org/philosophy/free-sw.de.html#n1 abgerufen

Henderson, P., Mitchell, E., Manning, C. D., Jurafsky, D., & Finn, C. (07. 11 2023). *arXive*. Von https://arxiv.org/abs/2211.14946 abgerufen

Mitchell, M. a. (2019). Model cards for model reporting. *Proceedings of the conference on fairness, accountability, and transparency*, S. 220–229.

Solaiman, I. (2023). The gradient of generative AI release: Methods and considerations. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, S. 111–122.

Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (07.11 2023). *Universal and Transferable Adversarial Attacks on Aligned Language Models*. Von arXive: https://arxiv.org/abs/2307.15043 abgerufen

YouGov Deutschland GmbH (2023): Umfrage zu KI: KMU erkennen Vorteile, zögern aber noch beim Einsatz. Von https://www.ionos.de/newsroom/news/umfrage-zu-ki-kmu-erkennen-vorteile-zoegern-aber-noch-beim-einsatz/ abgerufen

Herausgeber

Prof. Dr. Volker Wittpahl
Institut für Innovation und Technik (iit)
in der VDI/VDE Innovation + Technik GmbH
Steinplatz 1, 10623 Berlin

Zitation

Weber, Sabine; Hanel, Juliane (2024).
sWas bedeutet Open Source für Künstliche Intelligenz (KI)? iit perspektive Nr. 69.
Institut für Innovation und Technik (iit),
Berlin.

Autor:innen

Dr. Sabine Weber
Tel.: +49 (0)89 51089630-208
E-Mail: sabine.weber@vdivde-it.de

Juliane Hanel
Tel.: +49 (0)30 310078-5480
E-Mail: juliane.hanel@vdivde-it.de

iit perspektive Nr. 69

Januar 2024
Layout: Poli Quintana
DOI: 10.23776/2023_09
Bildnachweise: vectorfusionart – stock.adobe.com

DOI: 10.23776/2023_09