HOCHSCHULE
FÜR FERNSEHEN UND
FILM MÜNCHEN

Verbot von ChatGPT? Pause von KI ?

Sylvia Rothe
HFF, Lehrstuhl für KI in der Medienproduktion

KI-Pause?

30.03.2023, 19:58 Uhr

KI-Stopp jetzt: KI-Entwicklung muss pausieren!

Halbes Jahr Pause, jetzt: Sonst droht der Menschheit Gefahr. Elon Musk, Apple-Gründer Steve Wozniak sowie viele weitere Unternehmer und Experten fordern dies in einem offenen Brief. Sind ihre Sorgen rund um "gigantische KI-Experimente" berechtigt?

Von  Dominic Holzer

Mit GPT-4 ist die Grenze erreicht, der Rubikon bald überschritten, finden sie: Über 1.000 Unterzeichnende fordern in einem offenen Brief die Entwicklung Künstlicher Intelligenz (KI) für sechs Monate zu pausieren. Der Menschheit drohe sonst ein zivilisatorischer Kontrollverlust.

Offener Brief (22.3.2023)

Initiator:
Non-Profit-Organisation Future of Life Institute
Berater u.a. Elon Musk, Apple-Gründer Steve Wozniak, …

Forderung:
Für mindestens sechs Monate sollen alle KI-Labore die Entwicklung von KI-Systemen stoppen, die leistungsfähiger sind als GPT-4

https://futureoflife.org/open-letter/pause-giant-ai-experiments/

leistungsfähiger als GPT-4

- Vergleichbar ?
- GPT-4 gefährlich? Einsatzfähig?
- GPT-4: nicht offen
- BLOOM, GPT-NeoX, Luminous, …

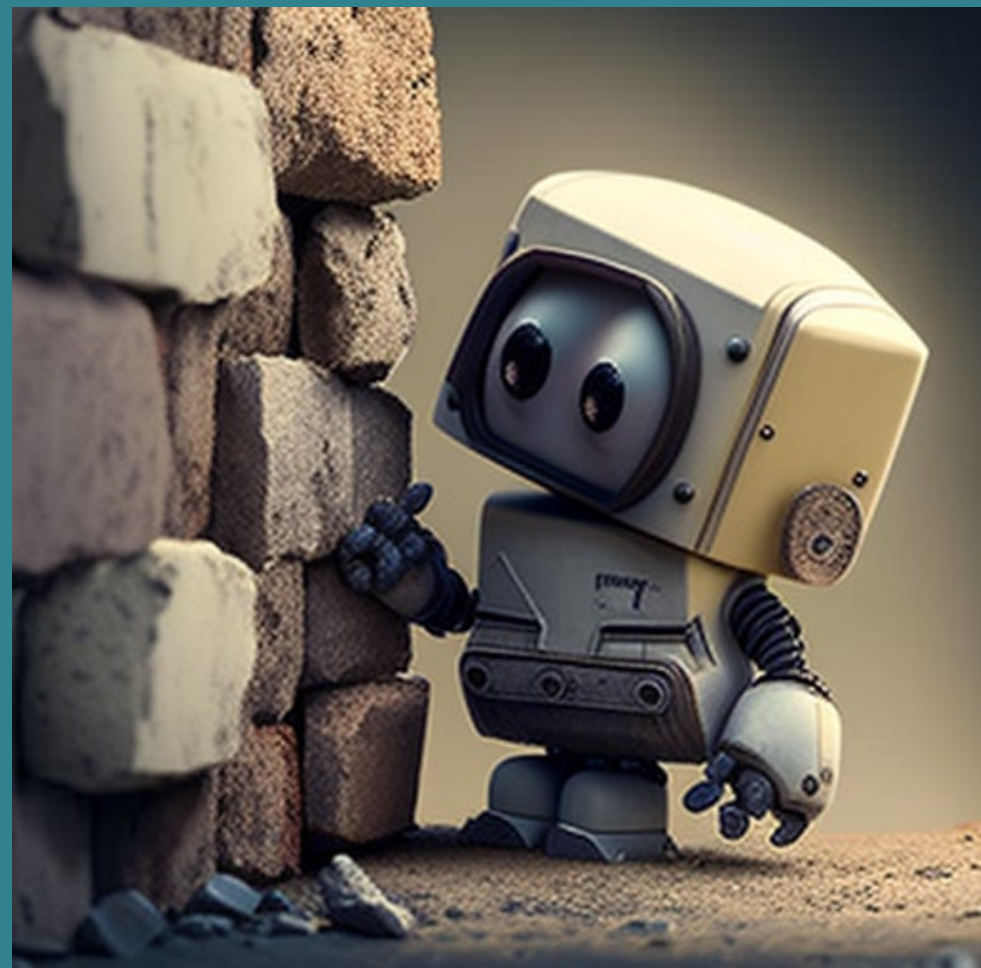
KI-Entwicklungen stoppen

- Umsetzbar? Kontrollierbar ?
- Welche Interessen?

Wettlauf der Systeme

- Google, Meta, …
- Anthropic
  - gegründet von von OpenAI-Aussteigern
  - plant für 5 Milliarden Dollar riesiges KI-Modell (zehnmal leistungsfähiger ist als die derzeit mächtigsten KI-Systeme, Training auf einem Rechencluster von "mehreren zehntausend GPUs")
  - Aussage: "We believe that companies that train the best 2025/26 models will be too far ahead for anyone to catch up in subsequent cycles.„
  - Unterzeichner des KI-Pause-Briefes
- Elon Musk: kaufte 10.000 GPUs für ein Twitter-KI-Projekt (https://www.businessinsider.com/elon-musk-twitter-investment-generative-ai-project-2023-4?op=1)

Verbot von ChatGPT?



Italien: Sperrung von ChatGPT

- Grund: Verstöße gegen den Daten- und Jugendschutz
- Anlass: Datenpanne im März
- Verbot, Nutzerdaten aus Italien zu verarbeiten

US-Verbraucherschutzbehörde ermittelt gegen OpenAI

- OpenAIs Markteinführung von GPT-4 verstoße US-amerikanisches Handelsrecht
- GPT-4 täusche und gefährde Menschen, sei voreingenommen und stelle ein Risiko für das Privatleben sowie die öffentliche Sicherheit dar
- Bedingungen für den Einsatz künstlicher Intelligenz :
  - Transparenz, Erklärbarkeit, Fairness
  - klare Verantwortlichkeit (Rechenschaftspflicht, Haftbarkeit im Schadensfall)

ChatGPT

Examples

"Explain quantum computing in simple terms" →

"Got any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →

Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests

Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

Limited knowledge of world and events after 2021

Send a message...

ChatGPT Mar 23 Version. Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts

OpenAI

2015 OpenAI LP als unabhängige Forschungseinrichtung gegründet

Ziel von OpenAI 2015:

- KI auf Open-Source-Basis entwickeln und zu vermarkten, dass sie der Gesellschaft Vorteile bringt und nicht schadet
- eine „freie Zusammenarbeit" mit anderen Institutionen und Forschern
- Patente und Forschungsergebnisse öffentlich zugänglich

seit 2019 ein gewinnorientiertes Start-up

Produkte: Dall-E, Whisper, GPT-3, ChatGPT,…

https://openai.com/charter

Privacy policy

Updated
April 7, 2023

We at OpenAI, L.L.C. (together with our affiliates, “OpenAI”, “we”, “our” or “us”) respect your privacy and are strongly committed to keeping secure any information we obtain from you or about you. This Privacy Policy describes our practices with respect to Personal Information we collect from or about you when you use our website and services (collectively, “Services”). This Privacy Policy does not apply to content that we process on behalf of customers of our business offerings, such as our API. Our use of that data is governed by our customer agreements covering access to and use of those offerings.

1. Personal information we collect

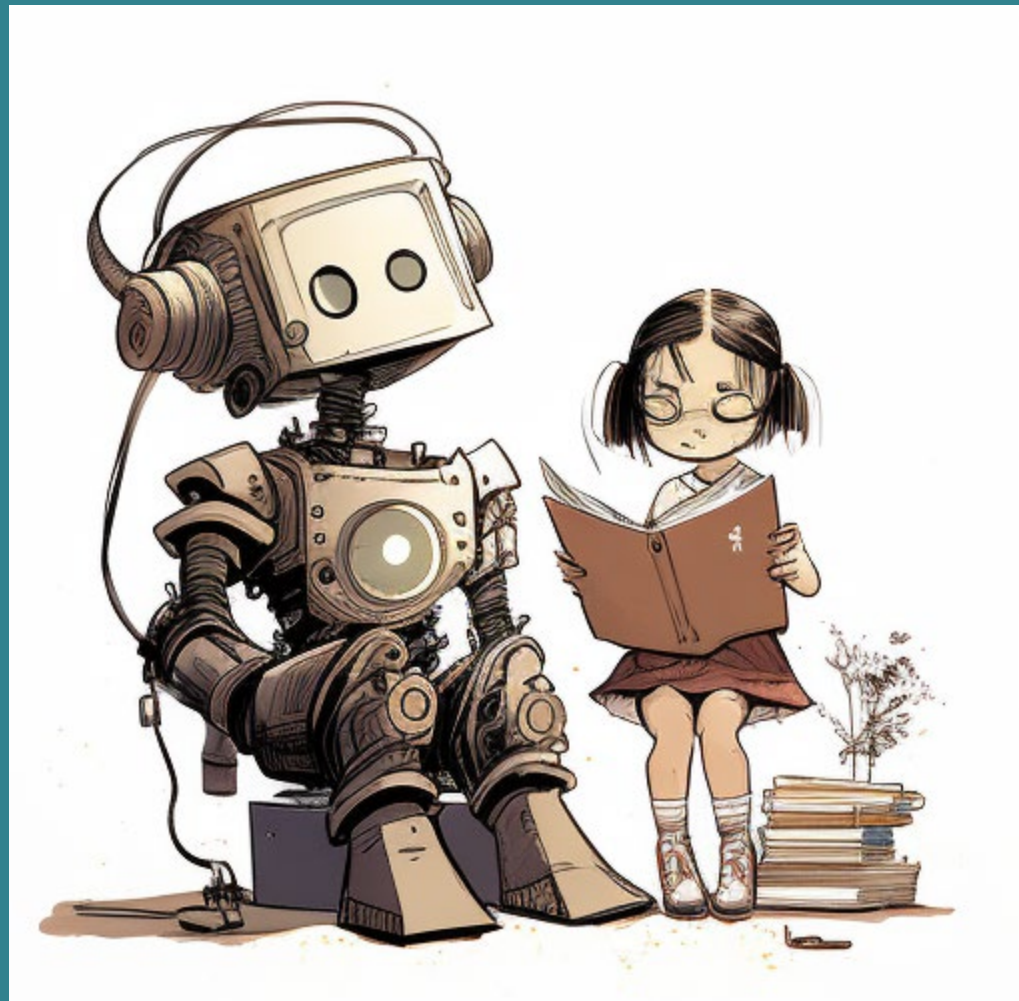
We collect information that alone or in combination with other information in our possession could be used to identify you (“Personal Information”) as follows:

Personal Information You Provide: We may collect Personal Information if you create an account to use our Services or communicate with us as follows:

- *Account Information:* When you create an account with us, we will collect information associated with your account, including your name, contact information, account credentials, payment card information, and transaction history, (collectively, “Account Information”).
- *User Content:* When you use our Services, we may collect Personal Information that is included in the input, file uploads, or feedback that you provide to our Services (“Content”).
- *Communication Information*: If you communicate with us, we may collect your name, contact information, and the contents of any messages you send (“Communication Information”).

https://openai.com/policies/privacy-policy

Was ist wichtig?



Was wir brauchen

- Rechtliche Grundlagen
- offene, für Forschende frei zugängliche KI-Modelle
- Transparenz
- Zusammenarbeit statt Wettlauf
- Keine Konzentration der KI-Modelle in großen Konzernen
- vorausschauende Überlegungen, Diskussionen, Forschung zu Anwendungen, Recht und Ethik

Umgang mit KI

- Festhalten an menschlichen Überprüfungen
- Regeln der Rechenschaftspflicht
- Möglichst offene KI-Modelle
- Vorteile von KI nutzen
- Debatten (z.B. rechtliche und ethische Aspkete)

HOCHSCHULE
FÜR FERNSEHEN UND
FILM MÜNCHEN

Aktuelle KI-Initiativen

LAION e.V. (Large-Scale Artificial Intelligence Network)

- Gemeinnützige Organisation
- Bereitstellung von Tools, Datensätzen und Modellen
- Sitz in Deutschland
- Ziele:
  - offene KI-Modelle
  - transparente Forschung
  - frei zugängliche Datensätze
  - kritische KI-Sicherheitsforschung, die der Allgemeinheit zugutekommt
  - technologische Unabhängigkeit von kommerziellen KI-Modellen großer Konzerne

https://laion.ai/ https://entwickler.de/machine-learning/laion-open-source-ai https://haveibeentrained.com/

LAION e.V. (Large-Scale Artificial Intelligence Network)

- statt Pause Beschleunigung der Forschung
- Einrichten eines gemeinsamen, internationalen Rechenclusters für groß dimensionierte offene Grundlagenmodelle (Large-Scale Foundation Models)
- Gemeinsame Erforschung von KI-Modellen
- Demokratisierung der KI-Forschung zu demokratisieren
- öffentlich finanzierter Supercomputer, um Open-Source-Modelle zu erstellen.

OpenAssistant Open-Source-ChatGPT

- LAION und YouTuber und KI-Influencer Yannic Kilcher
- Freiwillge erstellen Musterlösungen und bewerten die Antworten anderer
- Anschließend trainieren weitere Beteiligte auf dieser Grundlage Chatmodelle und veröffentlichen sie als Open Source.
- erste Work-in-Progress-Modelle als inoffizielle Demos ("OpenAssistant First Models are here: Open-Source ChatGPT„, 7.4.2023).
- Projektwebsite von OpenAssistant

LEAM (Large European AI Models)

Initiative des Bundesverbandes KI (KI Bundesverband) und führenden Vertretern aus Industrie und Forschung

- Sammlung und Erstellung umfassender Trainingsdatensätze
- Förderung exzellenter Forschung im Bereich KI
- Bereitstellung von Hyperscale-Infrastruktur
- Entwicklung von Organisationsstrukturen und Prozessen zur Etablierung eines kontinuierlichen Workflows zur Modellentwicklung und -verbesserung
- Integration der Modelle in das europäische Innovationsökosystem
- Entwicklung von Methoden, Benchmark-Datensätzen und Kriterien zur Sicherstellung ethischer Anforderungen und europäischer Werte

Die entwickelten Modelle sollen Open Source und für alle Marktteilnehmer frei zugänglich sind. Alle europäischen Sprachen sollen vollständig in die Modelle integriert werden.

https://leam.ai/

AI Act, EU KI-Verordnung

- https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:52021PC0206 Entwurf 21. April 2021, 6. Dezember 2022
- KI-Technologien werden in Kategorien sortiert (kein Risiko bis hohes Risiko)
- daran verschiedene Compliance- und Informationspflichten gekoppelt
- Verbot bei nicht akzeptablen Risiko (z.B. Social Scoring , Teile von biometrischer Videoüberwachung, subtile Verhaltensbeeinflussung)
- Wikipedia: „mit einer Einigung wird im Laufe des Jahres 2023 oder Anfang 2024 gerechnet, die Verordnung würde dann … zwei Jahre später Anwendung finden“

https://de.wikipedia.org/wiki/Gesetz_über_künstliche_Intelligenz

AI Act, EU KI-Verordnung

- Risikogruppen:
  - i) unannehmbares Risiko (z.B. Social Scoring): Verbot
  - ii) hohes Risiko (z.b.Personalmangement): umfassendes Qualitäts- und Risikomanagement
  - iii) geringes Risiko (z.b. Chatbots): Transparenzpflicht
  - iv) minimales Risiko (z.B. Spamfilter): keine Pflichten
- Bisher hochriskant:
  - *biometrische Identifizierung und Kategorisierung natürlicher Personen*
  - *Verwaltung und Betrieb kritischer Infrastrukturen*
- Generative KI hochriskant?

Ein risikobasierter Ansatz zur Regulierung

Je nach dem Typ der KI-Anwendung und ihrer Einstufung in eine Risikoklasse muss ihr Betreiber unterschiedliche hohe Schutz- und Transparenzanforderungen erfüllen.

Nicht akzeptables Risiko, z.B. Social Scoring	Verboten
Hohes Risiko, z.B. Bewerbungen, medizinische Geräte	Zulässig vorbehaltlich der Einhaltung hoher Compliance-Standards, Vorab-Folgeabschätzungen und -Zuverlässigkeitsprognosen
AI mit spezifischen Transparenzverpflichtungen	Zulässig, aber abhängig von Informations- und Transparenzverpflichtungen
Minimales oder kein Risiko	Zulässig ohne Beschränkungen

Quelle: Lucilla Sioli, A European Strategy for Artificial Intelligence

Hochrisiko-KI-Systemen

- Systeme, die erhebliche Risiken für die Gesundheit und Sicherheit oder die Grundrechte von Personen bergen.
- müssen Auflagen für vertrauenswürdige KI genügen und Konformitätsbewertungsverfahren unterzogen werden, bevor sie in Verkehr gebracht werden dürfen.
- Für einige KI-Systeme werden nur minimale Transparenzpflichten vorgeschlagen, insbesondere für den Einsatz von Chatbots oder „Deepfakes".

Hochrisikosysteme

- Höhere Auflagen. Dazu zählen unter anderem:

- *angemessene Risikobewertungs- und Risikominderungssysteme*
- *hohe Qualität der Datensätze, die in das System eingespeist werden, um Risiken und diskriminierende Ergebnisse so gering wie möglich zu halten*
- *Protokollierung der Vorgänge, um die Rückverfolgbarkeit von Ergebnissen zu ermöglichen*

Hochrisikosystem

1. Biometrische Identifizierung und Kategorisierung natürlicher Personen:

a) biometrische Echtzeit-Fernidentifizierung und nachträgliche biometrische Fernidentifizierung natürlicher Personen;

2. Verwaltung und Betrieb kritischer Infrastrukturen:

a) Sicherheitskomponenten in der Verwaltung und im Betrieb des Straßenverkehrs, der Wasser-, Gas-, Wärme- und Stromversorgung;

3. Allgemeine und berufliche Bildung:

a) Entscheidungen über den Zugang natürlicher Personen zu Einrichtungen der allgemeinen und beruflichen Bildung;

b) Bewertung von Schülern und Teilnehmern an für die Zulassung zu Bildungseinrichtungen erforderlichen Tests;

4. Beschäftigung, Personalmanagement und Zugang zur Selbstständigkeit:

a) Einstellung oder Auswahl natürlicher Personen, insbesondere das Sichten oder Filtern von Bewerbungen und das Bewerten von Bewerbern in Vorstellungsgesprächen oder Tests;

b) Entscheidungen über Beförderungen und Kündigungen von Arbeitsvertragsverhältnissen;

5. Zugänglichkeit und Inanspruchnahme grundlegender privater und öffentlicher Dienste und Leistungen:

a) Behördenbeurteilungen, ob natürliche Personen Anspruch auf öffentliche Unterstützungsleistungen und -dienste haben

b) Kreditwürdigkeitsprüfung und Kreditpunktebewertung natürlicher Personen,

c) Entsendung oder Priorisierung des Einsatzes von Not- und Rettungsdiensten, einschließlich Feuerwehr und medizinischer Nothilfe;

Hochrisikosystem

HOCHSCHULE FÜR FERNSEHEN UND FILM MÜNCHEN

6. Strafverfolgung:

a) individuelle Risikobewertungen natürlicher Personen

b) Lügendetektoren, Emotionserkennung

c) … g)

7. Migration, Asyl und Grenzkontrolle:

a) Lügendetektoren, Emotionserkennung

b) individuelle Risikobewertungen natürlicher Personen

c) … d)

8. Rechtspflege und demokratische Prozesse:

a) Unterstützung von Justizbehörden bei der Ermittlung und Auslegung von Sachverhalten und Rechtsvorschriften und bei der Anwendung des Rechts auf konkrete Sachverhalte

9.Generative KI ??? -> Überregulierung???? -> Zurückbleiben hinter China/USA < -- > Gesellschaftliche Verantwortung

Verbotene Systeme

- KI-Systeme, die Grundrechte, verletzen
- Praktiken, die ein erhebliches Potenzial haben, Personen zu manipulieren, indem sie auf Techniken zur unterschwelligen Beeinflussung zurückgreifen, die von diesen Personen nicht bewusst wahrgenommen werden
- Praktiken, die die Schwächen bestimmter schutzbedürftiger Gruppen wie Kinder oder Personen mit Behinderungen ausnutzen, um deren Verhalten massiv so zu beeinflussen, dass sie selbst oder eine andere Person psychisch oder physisch geschädigt werden könnten
- Bewertung des sozialen Verhaltens durch öffentliche Behörden („Social Scoring“)
- Einsatz von biometrischen Echtzeit-Fernidentifizierungssystemen in öffentlich zugänglichen Räumen für die Zwecke der Strafverfolgung bis auf wenige Ausnahmen

Forderungen der Anbieter

Allzweck-KI (General-Purpose-KI): verschiedenen Einsatzszenarien (Bild- oder Spracherkennung, - generierung, …)

Google, Microsoft, u.a.:

- Gleichgewicht der Verantwortung zwischen Nutzern, Anwendern und Anbietern
- vor allem der Anwender dafür verantwortlich, wenn mit den Programmen hochriskante Aufgaben erledigt werden

https://www.golem.de/news/ki-verordnung-hype-um-chatbots-bringt-eu-in-die-bredouille-2303-172461-3.html

Änderungsvorschläge (6.2.23)

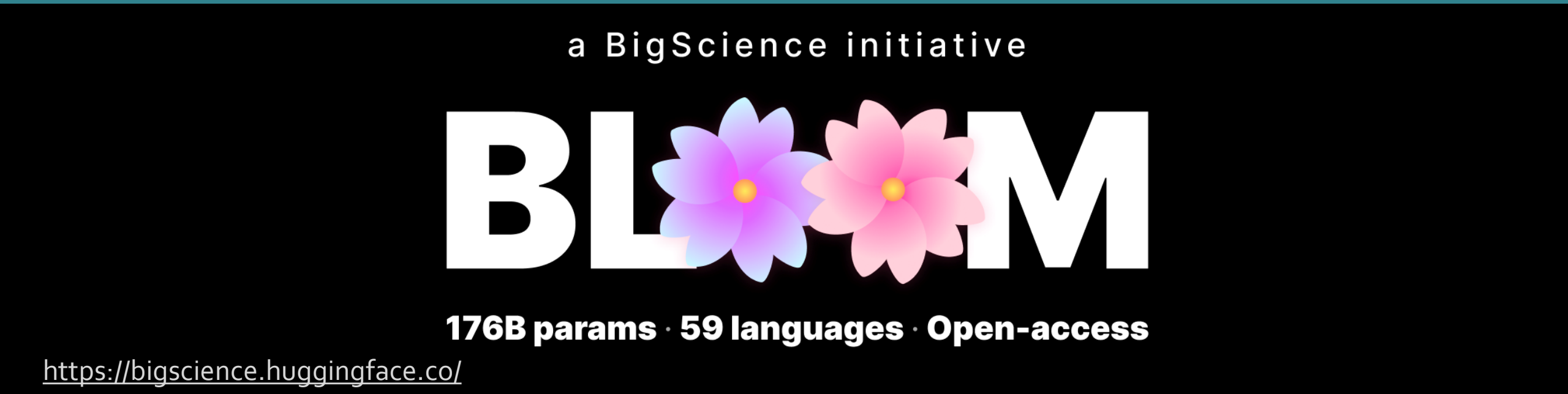
Neue Hochrisiko-KIs_:

- Systeme, die die Entwicklung von Minderjährigen beeinflussen könnten (z.B. Empfehlungssysteme in sozialen Netzwerken)
- generative KI-Systeme, die z.B. Texte erzeugen, die fälschlicherweise für menschengemacht gehalten werden könnten oder audiovisuelle Ergebnisse hervorbringen, die etwas aussagen, das nie stattgefunden hat.
- Ausnahmen:
  - Texte, wenn sie von einem Menschen überprüft wurden oder eine Person dafür rechtlich verantwortlich ist
  - audiovisuelle Inhalte in Kunstwerken

https://www.srd-rechtsanwaelte.de/blog/ki-verordnung-news-aktuelles/

Bloom (BigScience)

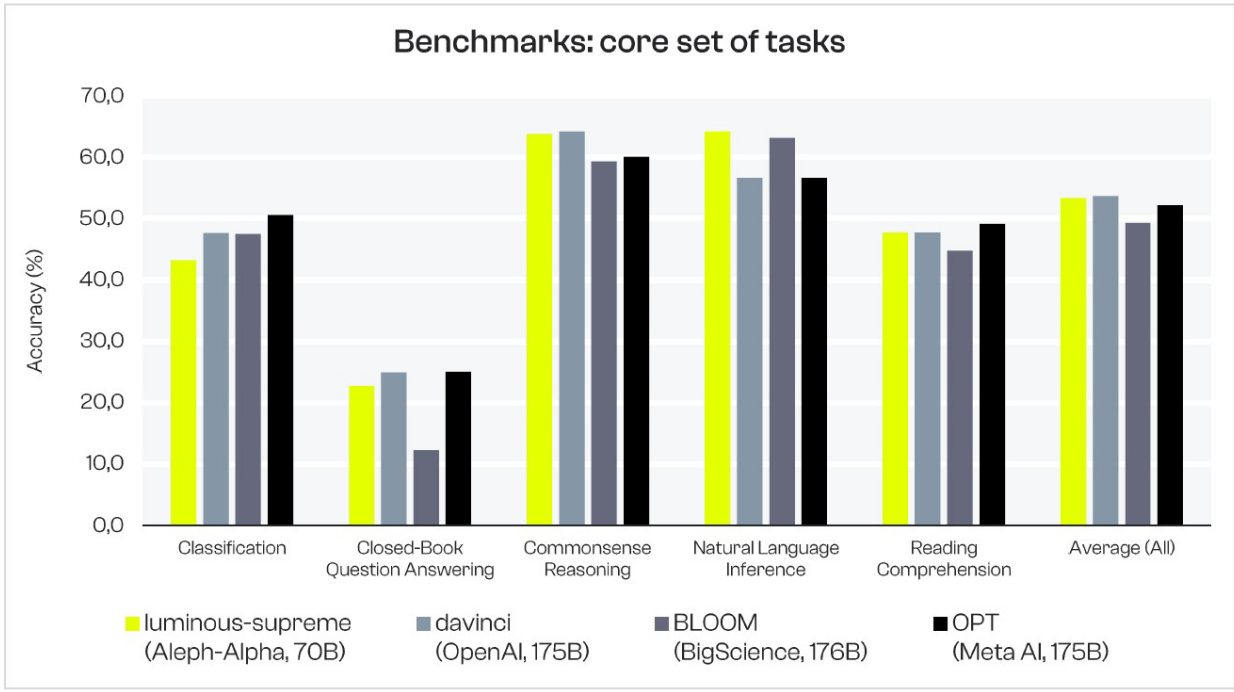
- **B**igScience **L**anguage **O**pen-science **O**pen-access **M**ultilingual
- Open Source
- frei verfügbar
- Textgenerierung in mehreren Sprachen



https://bigscience.huggingface.co/

Luminous (Aleph Alpha)

- Entwickelt von KI-Start-up in Deutschland
- Familie von Sprachmodellen
- Sprachverarbeitung im multimodalen Bereich: Erkennung und Beschreibung komplexer Bildinhalte



GPT-NeoX-20B (EleutherAI)

- Freigegeben Februar 2022
- Ursprung: Discord-Gruppe von KI-Interssierten
- Sprachmodell mit 20 Milliarden Parametern
- offen zugänglich

https://blog.eleuther.ai/announcing-20b/

Weitere frei zugängliche LLMs

GPT-NeoX-20B (EleutherAI)

https://blog.eleuther.ai/announcing-20b/

FLAN-T5 (Google)

https://huggingface.co/docs/transformers/model_doc/flan-t5

LLaMA (Large Language Model Meta AI, Facebook Research)

https://github.com/facebookresearch/llama