Übersetzung aus dem Englischen

Stellungnahme für den Bundestagsausschuss für Digitales

Öffentliche Anhörung „Generative Künstliche Intelligenz“

am Mittwoch, 24. Mai 2023, 14:30 – 16:30 Uhr

Catelijne Muller, ALLAI

Deutscher Bundestag
Ausschuss für Digitales

Ausschussdrucksache
zu 20(23)158

24.05.2023

1. Allgemeine Anmerkungen

Im laufenden Gesetzgebungsverfahren zum europäischen KI-Gesetz (AI Act, AIA) haben sich sogenannte Mehrzweck-KI-Systeme (General Purpose AI, GPAI) zu einem Streitthema entwickelt. Es wurden verschiedene Vorschläge lanciert, die sich zwischen dem vollständigen Ausschluss von GPAI aus dem Anwendungsbereich des AI Act und der Schaffung eines gesonderten Status für diese bewegten.

Auch die jüngsten Entwicklungen rund um die – ChatGPT, AutoGPT und BabyAGI zugrundeliegenden – großen Sprachmodelle (Large Language Models, LLM) wie GPT-3.5 und GPT-4, darunter auch mehrere offene Briefe und eine durch den italienischen Datenschutzbeauftragten durchgeführte Untersuchung zu ChatGPT, rückten das Thema der GPAI und auch der generativen KI (GenAI) sowie der „Foundation Models“ (FM) noch stärker in ein kritisches Licht.

1.1 Was sind General Purpose AI (GPAI), Generative KI und „Foundation-Modelle“?

Zunächst einmal ist zu erwähnen, dass aktuell noch diskutiert wird, was genau unter „General Purpose AI“ zu verstehen ist und worin die Unterschiede zwischen GPAI, generativer KI und Foundation-Modellen bestehen. Kürzlich schlug das Europäische Parlament zwei Definitionen vor – eine für General Purpose AI und eine für Foundation-Modelle. Es gibt auch noch den Aspekt generativer KI (GenAI), was zu weiterer Verwirrung führt. Die Mitgliedstaaten haben ihre Stellungnahme zum AIA im Dezember verabschiedet und definierten GPAI dabei als KI-System, das *„vom Anbieter zur Durchführung allgemein anwendbarer Funktionen wie Bild- und Spracherkennung, Audio- und Videoerstellung, Mustererkennung, Fragenbeantwortung, Übersetzung und anderen gedacht ist; ein General Purpose AI-System kann in vielfältigen Zusammenhängen zur Anwendung kommen und in eine Vielzahl anderer KI-Systeme integriert werden.“* Soweit für uns ersichtlich betrachtet das Europäische Parlament Foundation-Modelle als Untergruppe von GPAI und GenAI als eine Funktionalität. In dieser Reaktion werden wir diese drei Begrifflichkeiten unter GPAI/GenAI/FM zusammenfassen.

1.2 Einzelne Fehlerquellen mit großer Wirkung

Der Trend geht klar hin zu einer immer kleineren Anzahl immer allgemeinerer und oft sehr großer Modelle. Auch wenn das Verhalten dieser Modelle zu überzeugen vermochte, können sie auch unerwartet versagen (halluzinieren), bergen Voreingenommenheiten (Bias) und werden schlecht verstanden. Da diese Systeme im großen Maßstab eingesetzt werden,

können sie zu einzelnen Fehlerquellen werden, von denen aus Schäden (z. B. Sicherheitsrisiken, Diskriminierung, Ungerechtigkeiten) auf unzählige nachgelagerte KI-Anwendungen ausstrahlen.

Dies gilt ganz zu schweigen von den zahlreichen, durch dieses Modell aufgeworfenen rechtlichen und ethischen Fragen, so wie jene im Zusammenhang mit dem Datenschutz, geistigen Eigentumsrechten, dem Automatisierungsbias, der Manipulationsmacht, der Skalierung der Desinformation, der Kompetenzerosion, der potenziellen Arbeitsplatzverlagerung, dem Risiko der unkontrollierbaren Autonomie usw.

1.3 Benchmark-Datensätze

Neben „allgemeinen" KI-Modellen ist auch die Praxis der Verwendung sogenannter „Benchmark"-Datensätze, die das Rückgrat der Forschung und Entwicklung zum maschinellen Lernen bilden, weit verbreitet. Jüngste kritische Prüfungen dieser Datensätze förderten jedoch Voreingenommenheiten (Bias), schlechte Kategorisierungen und beleidigende Zuordnungen zutage. Selbst viele der fairsten Forscher:innen im Bereich maschinellen Lernens verwenden Datensätze „wie sie sind", ohne sie auf Vollständigkeit, Repräsentanz und Fairness insgesamt zu überprüfen.

1.4 Homogenität

Die oben beschriebenen Problemstellungen im Zusammenhang mit GPAI/GenAI/FM (bestehend aus immer weniger und immer allgemeiner werdenden Modellen und Benchmark-Datensätzen) lassen sich als „Problem der Homogenität" bezeichnen. Maschinelles Lernen führt seiner Natur nach im Vergleich zu menschlichen Entscheidungen zu homogeneren Entscheidungsprozessen. Wenn immer weniger Maschinen immer mehr Entscheidungen prägen würden, könnten Voreingenommenheiten und Fehler sich verstärken und quer durch die Gesellschaft verankert und verallgemeinert werden.

1.5 GPAI/GenAI/FM und der AIA – beabsichtigter Zweck und vernünftigerweise vorhersehbare Verwendung

Eines der Argumente zur Schaffung eines gesonderten Status für GPAI/GenAI/FM im AI Act besteht darin, dass Anbieter:innen von GPAI/GenAI/FM den Zweck, zu dem ihr System verwendet werden wird, nicht kennen, weshalb die Risikokategorie ihres Systems nicht im Vorfeld bestimmt werden kann.

In unserer Abhandlung „AIA in-depth #1 | Objective, Scope, Definition" stellen wir einen Ansatz vor, um dieses Thema anzugehen. Dieser ist in der Unionsgesetzgebung zur Produktsicherheit üblich. Der Ansatz besteht darin, den Begriff der „vernünftigerweise vorhersehbaren Verwendung" hinzuzufügen. Angesichts der potenziellen Auswirkungen dieser GPAI-/GenAI-/FM-Systeme ist es nicht unvernünftig, deren Anbieter:innen darum zu bitten, einen Versuch der Vorhersage der potenziellen Verwendungen ihres Systems zu unternehmen und dieses entsprechend zu kategorisieren. Mit anderen Worten: Wenn es vernünftigerweise vorhersehbar ist, dass ein GPAI-/GenAI-/FM-System als Hochrisiko-KI-System gemäß ANHANG II oder III des AI Act (oder als Bestandteil hiervon) eingesetzt wird, dann wird das GPAI-/GenAI-/FM-System selbst ebenfalls als Hochrisikosystem eingestuft. Im

Bewusstsein, dass nicht alle Verwendungen vorhersehbar sind, würde der Begriff nur jene Verwendungen umfassen, die vernünftigerweise vorhersehbar sind.

In den Allgemeinen Ansatz der Mitgliedstaaten wurde ein neues Kapitel über General Purpose AI (GPAI) aufgenommen, einschließlich dieser, wenn auch leicht abgewandelten, Auffassung der „vorhersehbaren Verwendung" an zwei Stellen in Artikel 4b:

1. KI-Systeme mit allgemeinem Verwendungszweck, die als Hochrisiko-KI-Systeme oder als Komponenten von Hochrisiko-KI-Systemen [...] verwendet werden können [...]

6. Bei der Erfüllung der in [...] genannten Anforderungen und Verpflichtungen

- ist jede Bezugnahme auf die Zweckbestimmung als Bezugnahme auf die mögliche Verwendung von KI-Systemen mit allgemeinem Verwendungszweck als Hochrisiko-KI-Systeme oder als Komponenten von Hochrisiko-KI-Systemen im Sinne von Artikel 6 zu verstehen;

Es wird auch angeregt, dass zu einem späteren Zeitpunkt von der Europäischen Kommission spezifische Anforderungen an GPAI/GenAI/FM gestellt werden sollten. Näheres zum Standpunkt des Rates zu GPAI/GenAI/FM finden Sie in unserem „AIA Policy Analysis | Council General Approach".

1.6 Für GPAI/GenAI/FM sollte ein höherer Standard gelten

Ein weiteres Argument zur Schaffung eines gesonderten Status für GPAI/GenAI/FM besteht darin, dass diese Systeme immer ein Re- oder Up-Training durchlaufen müssen, bevor sie in einem neuen Bereich zu einem bestimmten Zweck eingesetzt werden können (man denke an die Tumorerkennung im Gesundheitsbereich). Anbieter:innen von GPAI/GenAI/FM könnten in ihrem Compliance-Verfahren daher nie die Vielzahl nachgelagerter Anwendungen „vorhersehen", die einen solchen Re-Training-Prozess durchlaufen würden.

Zunächst behandeln die Datenanforderungen des AIA diese Problemstellung bereits auf schlaue Weise. Absatz 2 Buchstabe g ermöglicht „*die Ermittlung möglicher Datenlücken oder Mängel und wie diese Lücken und Mängel behoben werden können*." (siehe ANHANG I dieser Abhandlung). Hier verbleibt die Verantwortung zur Bereitstellung einer qualitativ hochwertigen, robusten und vertrauenswürdigen Kernfunktionalität bei den Anbieter:innen von GPAI/GenAI/FM, was bei Anwendungsfällen mit hohem Risiko auch die Verpflichtung zur ordnungsgemäßen Information aller nachgelagerten Nutzer:innen über mögliche Datenlücken oder Mängel einschließt.

Es ließe sich für GPAI/GenAI/FM sogar argumentieren, dass aufgrund ihrer potenziellen Verwendung in einer ganzen Bandbreite von Hochrisikobereichen (Gesundheitsversorgung, kritische Infrastruktur, Strafverfolgung) eher ein höherer, statt eines niedrigeren, Standards für sie gelten sollte. Tatsächlich besteht das Gesamtziel der Union bei sicherheits- und haftungsrechtlichen Rahmenbedingungen in der Gewährleistung, dass alle Produkte und Dienstleistungen, einschließlich jener, in die aufkommende digitale Technologien eingebunden sind, sicher, zuverlässig und konsequent funktionieren und dass Schäden effizient abgeholfen wird. Die EU verfolgt einen Ansatz, der sich von jenem in anderen Teilen der Welt unterscheidet, in denen die Verantwortung im Nachhinein bestimmt wird, was

häufig zu großen Haftungsforderungen führt. Man würde so auch mit dem Gesamtziel des AIA brechen, welches im Schutz der Gesundheit, der Sicherheit und der Grundrechte vor negativen Auswirkungen der KI besteht.

1.7 Eine Verschiebung der Verantwortung auf nachgelagerte Anbieter:innen wird Innovationen ersticken

Bei einer Begrenzung des Anwendungsbereichs des AIA für GPAI/GenAI/FM läuft man auch Gefahr, Innovationen vielmehr zu ersticken, denn sie zu fördern. Durch die Herabsetzung der Anforderungen oder das Setzen niedrigerer Standards auf Ebene der Anbieter:innen von GPAI/GenAI/FM würde sich die Verantwortung für die Einhaltung des AIA durch diese Systeme auf die „nachgelagerten" Nutzer:innen verlagern. Sie wären diejenigen, die die Anforderungen an Hochrisiko-KI erfüllen müssten, was vor allem für KMU und Mikrounternehmen eine zu hohe Belastung – oder vielleicht sogar technisch unmöglich – sein dürfte.

Selbst wenn Entwickler:innen von GPAI/GenAI/FM „nachgelagerten Nutzer:innen" bei den Formalitäten der AIA-Einhaltung helfen würden, so würden sie doch in eine Position völliger Abhängigkeit versetzt. Im Ergebnis könnte dies einerseits zu einer Begrenzung der Inanspruchnahme von GPAI-/GenAI-/FM-Systemen, sowie andererseits zu einer (weiteren) Konzentration der KI-Innovationskraft bei Entwickler:innen von GPAI/GenAI/FM führen.

Um zu verhindern, dass Entwickler:innen von GPAI/GenAI/FM in Verträgen oder AGB auch ihre Haftung gegenüber nachgelagerten Nutzer:innen beschränken oder gar ausschließen, hat das EP bereits eine Formulierung zur Vermeidung derartiger Vertragsbestimmungen angeregt.

1.8 Anforderungen für Hochrisiko-KI und GPAI/GenAI/FM

Bisher haben wir noch keinen Überblick gesehen, aus dem sich ergeben würde, welche Anforderungen für GPAI/GenAI/FM-Systeme angepasst werden müssen oder von Anbieter:innen von GPAI/GenAI/FM nicht erfüllt werden können. Daher haben wir in ANHANG I dieser Abhandlung die Anforderungen angesichts von GPAI/GenAI/FM vorläufig geprüft. Wir haben die Anforderungen um den zuvor erwähnten Begriff der „vernünftigerweise vorhersehbaren Verwendung" ergänzt, mit der Folge, dass jede Anforderung angesichts der „vernünftigerweise vorhersehbaren Verwendung" des GPAI-/GenAI-/FM-Systems betrachtet wird.

Diese vorläufige Beurteilung lässt darauf schließen, dass es bei den Anforderungen für Hochrisiko-KI (Titel III Kapitel 2 des AIA) nur wenige Aspekte gibt, deren Erfüllung für Anbieter:innen von GPAI/GenAI/FM aufgrund der Tatsache, dass sie nicht wissen, wie ihr System verwendet werden wird, theoretisch schwierig wäre. Tatsächlich kann eine Reihe von Anforderungen nur durch die Anbieter:innen der GPAI/GenAI/FM erfüllt (d. h. in das System eingebaut) werden, und nicht durch nachgelagerte Nutzer:innen. Wir möchten betonen, dass wir nicht berücksichtigt haben, ob eine Erfüllung der Anforderungen allgemein möglich ist. Wenn dem nicht so ist, wird das System tatsächlich nicht dem AIA entsprechen, und zwar unabhängig davon, wer dafür verantwortlich ist.

Im Vergleich zur vollen Verantwortung nachgelagerter Anbieter:innen, alle Anforderungen erfüllen zu müssen, liefert dies ein starkes Argument dafür, dass die aktuellen Anforderungen auch für Anbieter von GPAI/GenAI/FM gelten. Angesichts der aktuellen Entwicklungen rund um generative KI könnten vor allem im Zusammenhang mit geistigen Eigentumsrechten, Manipulation, maschineller Autonomie sowie potenziell auftauchendem Verhalten zusätzliche Anforderungen notwendig werden.

1.9 Haftung

Die Einbeziehung von GPAI/GenAI/FM in den AIA wird durch den jüngsten Vorschlag einer KI-Haftungsrichtlinie (in Verbindung mit dem Vorschlag zur Überarbeitung der Produkthaftungsrichtlinie) noch relevanter. Bei diesen Vorschlägen wird die Nichteinhaltung des AIA als Grund für die Annahme der Kausalität zwischen Anbieter:in und KI-System betrachtet. Klammert man die Anbieter:innen von GPAI/GenAI/FM aus dem Anwendungsbereich des AIA aus, so würden diese auch außer Reichweite der KI-Haftungsrichtlinie verschoben.

1.10 ChatGPT, AutoGPT, BabyAGI

In den letzten paar Monaten haben große Sprachmodelle (LLM) die Welt wie im Sturm erobert. Man hat schon vieles über sie gesagt und auf ihre Risiken wurden ausgiebig eingegangen. OpenAI selbst hat in seiner GPT-4-Systemkarte potenzielle Risiken beschrieben (und getestet). Das Modell zeigt die Tendenz, zu „halluzinieren“ (d. h. falsche Informationen zu liefern, einschließlich nicht existierender wissenschaftlicher Abhandlungen, falscher Anschuldigungen, fehlerhafter Berechnungen usw.) Ein australischer Bürgermeister hat OpenAI dafür verklagt, dass ChatGPT ihn fälschlicherweise der Bestechlichkeit und der Ableistung einer Haftstrafe beschuldigte. Von Fachleuten wird vor einer Überflutung des Internets mit Fake News und polarisierenden Inhalten gewarnt. Europol warnte vor einem Anstieg krimineller Aktivitäten wie z. B. Hacking, Cyberattacken und Phishing, welche mithilfe von ChatGPT wesentlich erleichtert werden können. Lehrkräfte mühen sich mit Schüler:innen ab, die ihre Hausaufgaben von ChatGPT erledigen lassen. Von Unternehmen wird die Nutzung von ChatGPT durch ihre Arbeitskräfte wegen Gefährdung ihres Geschäftsmodells untersagt. Und so geht es weiter.

1.11 Gesonderte Anforderungen für GPAI/GenAI/FM

OpenAI beschreibt auch das potenzielle Risiko der „autonomen Reproduktion“. Während seine Tests ergaben, dass GPT-4 (das ChatGPT zugrundeliegende große Sprachmodell [LLM]) auf Grundlage erster Versuche bei der Aufgabe der autonomen Reproduktion keine Wirkung zeigte, so wird dennoch angemerkt, dass für eine verlässliche Einschätzung riskanter emergenter Fähigkeiten von GPT-4 zusätzliche Tests notwendig seien.

In den letzten paar Wochen haben wir jedoch Versuche gesehen, die ein gewisses Maß an autonomer Reproduktion nachweisen. Computerwissenschaftler:innen haben mehrere Anwendungen auf GPT-4 aufgebaut, allen voran AutoGPT und BabyAGI, die anhand eines einzigen von menschlicher Hand definierten „Ziels“ dazu in der Lage sind, Folgeaufgaben selbst zu generieren und durchzuführen. Diese Systeme sind in der Lage, das Internet zu durchsuchen, ein Google-Konto zu eröffnen, einen Google-Drive-Ordner einzurichten, eine

Datei zu öffnen und dieser Datei Text hinzuzufügen, ohne dass der Mensch weiter eingreifen müsste. Insbesondere BabyAGI zeigte eine Form autonomer Reproduktion, wobei es ein von Menschen vorgegebenes Ziel in mehrere Unteraufgaben aufteilte, die sodann von verschiedenen GPT-4-Sprachmodellen, die es selbst angestoßen hatte, simultan ausgeführt wurden. Hier ist zu beachten, dass von den Computerwissenschaftler:innen selbst eingeräumt wird, dass für diese Systeme Sicherheitsvorkehrungen getroffen werden müssen.

1.12 KI-getriebene Manipulation

Kürzlich beging ein Belgier Selbstmord, nachdem er eine längere Unterhaltung mit einem Chatbot geführt hatte, der auf Grundlage eines großen Sprachmodells (LLM) betrieben wird. Seiner Frau zufolge nahm die Unterhaltung mit dem Chatbot eine verstörende Wendung und führte zum Selbstmord des Mannes. Ein anderes Unternehmen, das eine Chatbot-App betreibt, die intime Beziehungen mit den Nutzer:innen aufbaut, stellte fest, dass diese betrübt waren, nachdem der Grad an Intimität in den Unterhaltungen zurückgefahren worden war. Als Reaktion hierauf fügte das Unternehmen der App die Nummer der Selbstmord-Hotline hinzu.

Zurzeit fehlt es noch an einem ausreichenden Verständnis der – u. a. Chatbots innewohnenden – mächtigen Auswirkungen der KI-Manipulation oder einer hinlänglichen Befassung mit diesen. Die Eindämmung dieser Auswirkungen ist nicht durch die bloße Auferlegung von Transparenzmaßnahmen möglich. In unserer Abhandlung „AIA in-depth #2 | Prohibited AI Practices“ haben wir die Auffassung vertreten, dass der AIA eine großartige Möglichkeit der Befassung mit den Gesetzeslücken und den breiteren gesellschaftlichen Schäden, die eine KI-getriebene Manipulation mit sich bringen kann, bietet. Ein Verbot von KI-Praktiken, die auf Täuschung, wesentliche Verhaltensverzerrungen oder Ausnutzung der Schwachpunkte einer Person abzielen oder hierzu führen, würde sich gut in das weitere Ziel des AIA einfügen. Wir haben eine Änderung des in Art. 5 Buchstaben a und b enthaltenen Verbots angeregt, welche bereits zum Teil durch den Rat berücksichtigt wurde.

Uns ist bewusst, dass die Durchsetzung dieses Verbots herausfordernd sein wird, doch bei der Gesetzgebung gibt es viele Herausforderungen. Das hat uns auch bisher nicht davon abgehalten, Gesetze zu erlassen. Ein derartiges, eindeutiges Verbot wird andererseits eine großartige präventive Wirkung entfalten, die nicht zu unterschätzen ist.

Uns ist bewusst, dass dies bedeuten könnte, dass GPAI-/GenAI-/FM-Systeme immer die Anforderungen für Hochrisiko-KI werden erfüllen müssen, selbst wenn sie in Bereichen oder Anwendungen mit niedrigem Risiko eingesetzt werden. Wir möchten jedoch geltend machen, dass die Einhaltung dieser Anforderungen die Qualität, Zuverlässigkeit und Vertrauenswürdigkeit von GPAI-/GenAI-/FM-Systemen allgemein anheben wird, wodurch ein positiver Gesamttrend gesetzt wird.

Angesichts der jüngsten Entwicklungen rund um die generative KI und insbesondere große Sprachmodelle (LLM) könnten zusätzliche Sicherheitsvorkehrungen / Anforderungen für GPAI/GenAI/FM sogar notwendig sein.

2. Antworten auf die Fragen des Bundestagsausschusses für Digitales

Übersetzung aus dem Englischen

1) Die Regulierung generativer KI ist derzeit Gegenstand der Verhandlungen um den europäischen AI Act (AIA). Wie kann Ihrer Einschätzung nach generative KI wirksam im AIA einbezogen und reguliert werden und wie beurteilen Sie vorgeschlagene Differenzierungen innerhalb generativer KI zwischen „general purpose AI" und „foundation models"?

Siehe allgemeine Anmerkungen.

2) Generative KI bietet zahlreiche Anwendungsmöglichkeiten in den unterschiedlichsten Berufsständen und kann für Entlastungen am Arbeitsmarkt sorgen. Wie schätzen Sie die Potenziale und Risiken generativer KI für die Arbeitswelt ein und wo sehen Sie Regelungsbedarf?

Historisch gesehen hat schneller technologischer Wandel häufig zu Arbeitsplatzverlusten, aber auch zur Entstehung neuer Arbeitsplätze, geführt. Es stellt sich jedoch die Frage, wie groß die zeitliche Verzögerung zwischen dem Verlust und der Schaffung von Arbeitsplätzen ausfällt und wie die Vorteile des neu geschaffenen Wertes verteilt werden. Bei der aktuellen Kurve generativer KI-Technologie könnten Arbeitsplatzverluste sehr schnell vonstatten gehen. So werden jüngst zum Beispiel viele Copywriter:innen nur noch um Überprüfung KI-generierter Texte gebeten oder vollständig entlassen.

3) Inwieweit können sich Anwendungen aus staatlichen oder wirtschaftlichen Systemen, die nicht immer demokratische und freiheitliche Werte teilen, auf die europäische Gesellschaft auswirken und wie sollten die EU und Deutschland damit umgehen?

Der AI Act wird eine globale Reichweite entfalten, da er, unabhängig vom Ort ihrer Entwicklung, alle Systeme umfassen wird, die EU-Bürger:innen betreffen. Ähnlich wie im Falle der DSGVO wird auch er außerhalb des Unionsgebiets seine Wirkung entfalten.

5) Derzeit kursieren zahlreiche Vorschläge, um die regulatorischen Herausforderungen generativer KI-Anwendungen in den EU-Gesetzgebungsvorhaben für eine KI-Verordnung und eine KI-Haftungsrichtlinie passgenau zu verankern. Ist der risikobasierte Ansatz zur Regulierung generativer KI überhaupt geeignet oder braucht es z. B. eine systemische Risikoanalyse analog zum Risikoanalyse- und Minimierungsmechanismus im DSA?

Der risikobasierte Ansatz (siehe Antwort auf Frage 1) und eine systemische Risikoanalyse könnten auch gleichzeitig zum Einsatz kommen. Darüber hinaus sollen auch die Anforderungen des AI Act hinsichtlich eines Risikomanagementsystems sowie der Überwachung nach dem Inverkehrbringen der Analyse systemischer Risiken dienen. Das Europäische Parlament unterbreitet den Vorschlag, den Schutz der Demokratie, der Rechtsstaatlichkeit und der Umwelt als Ziele des AIA hinzufügen (hierzu zählen auch systemische Risiken).

6) Sind neue Phänomene und Fragestellungen im Hinblick auf einen negativen Einfluss von Anwendungen generativer KI auf den demokratischen Meinungsbildungsprozess zu erwarten? Wie lassen sich Medienfreiheit und Meinungsvielfalt im Zeitalter generativer KI rechtlich und politisch stärken, auch – aber nicht ausschließlich – im Hinblick auf die angemessene Vergütung von Journalist:innen, Künstler:innen und Kreativen? Wo sehen Sie möglichen Anpassungsbedarf etwa im Urheberrecht?

Übersetzung aus dem Englischen

Eines der Hauptrisiken generativer KI besteht in der Verbreitung von Falschinformationen und Deep Fakes im großen Maßstab, was zur Verbreitung von Fake News, Filterblasen, Desinformation und Propaganda führen kann und die Fähigkeit von Individuen zur Meinungsbildung und -entwicklung, zum Erhalt und zur Weitergabe von Informationen und Ideen beeinträchtigen und sich somit auf unsere Meinungsfreiheit und die Funktionsfähigkeit unserer Demokratie auswirken kann.

14) Welche Regeln braucht es aus Ihrer Sicht beim AI-Act für Generative KI, konkret was die Pflichten für Entwickler von Foundation-Modellen zur Informationsweitergabe innerhalb der Lieferkette angeht? Welche Vor- und Nachteile gehen mit solchen Pflichten einher? Ab welcher Schwelle sollten für Anwendungen, die auf Generativer KI basieren, die Hoch-Risiko-Regeln greifen, welche im AI-Act vorgesehen sind?

Ich begrüße die durch den IMCO- und den LIBE-Ausschuss des Europäischen Parlaments im neuen Artikel 28b des AIA vorgeschlagenen Verpflichtungen, die besagen, dass Anbieter:innen von Foundation-Modellen:

(a) durch geeignete Gestaltung, Erprobung und Analyse unter Beweis stellen, dass die Identifizierung, Reduzierung und Abmilderung vernünftigerweise vorhersehbarer Risiken für die Gesundheit, die Sicherheit, die Grundrechte, die Umwelt, die Demokratie und die Rechtsstaatlichkeit vor und während der Entwicklung mit geeigneten Methoden, so wie durch die Einbindung unabhängiger Fachleute sowie die Dokumentation nicht vermeidbarer Risiken nach der Entwicklung *[hier fehlt wahrscheinlich: „sichergestellt wird"; Anm. d. Ü.]*;

(b) nur Datensätze verarbeiten und integrieren, die geeigneten Daten-Governance-Maßnahmen für Foundation-Modelle, insbesondere Maßnahmen zur Prüfung der Angemessenheit der Datenquellen sowie möglicher Bias und geeigneten Abmilderungsmaßnahmen unterliegen;

(c) das Foundation-Modell so gestalten und entwickeln, dass während dessen Lebenszyklus bei Leistung, Vorhersehbarkeit, Interpretierbarkeit, Korrigierbarkeit, Sicherheit und Cybersicherheit ein geeignetes Niveau erreicht wird, das mittels geeigneter Methoden, so wie der Modellevaluierung unter Beteiligung von Fachleuten, der dokumentierten Analyse und der ausgiebigen Erprobung während der Konzeptionalisierung, Gestaltung und Entwicklung, geprüft wird;

(d) das Foundation-Modell unter Anwendung geltender Standards zur Reduzierung des Energieverbrauchs, des Ressourceneinsatzes und Abfalls sowie der Steigerung der Energieeffizienz und der Gesamteffizienz des Systems gestalten und entwickeln. Dies gilt unbeschadet bestehender unionsrechtlicher und einzelstaatlicher Rechtsvorschriften, und diese Verpflichtung gilt erst nach Veröffentlichung der in Artikel 40 genannten Standards. Sie sind so zu konzipieren, dass sie Möglichkeiten zur Messung und Protokollierung des Energie- und Ressourcenverbrauchs sowie, sofern technisch machbar, anderer durch Einsatz und Nutzung der Systeme in ihrem ganzen Lebenszyklus möglicherweise entstehenden Umweltauswirkungen bieten;

(e) erstellen umfassende technische Dokumentationen und verständliche Anwendungshinweise, um nachgelagerten Anbieter:innen die Erfüllung ihrer Verpflichtungen gemäß Artikel 16 und 28 Absatz 1 zu ermöglichen;

(f) richten ein Qualitätsmanagementsystem zur Gewährleistung und Dokumentation der Einhaltung dieses Artikels mit der Möglichkeit des Experimentierens bei der Erfüllen dieser Anforderung ein;

(g) registrieren jenes Foundation-Modell gemäß den Hinweisen in Anhang VIII Abschnitt C bei der in Artikel 60 bezeichneten EU-Datenbank;

(h) stellen den zuständigen einzelstaatlichen Behörden die in Absatz 1 Buchstabe c benannte technische Dokumentation für einen auf zehn Jahre nach Inverkehrbringen oder Inbetriebnahme ihres Foundation-Modells befristeten Zeitraum zur Verfügung;

Außerdem begrüße ich die Verpflichtung von Anbieter:innen von Foundation-Modellen, die in KI-Systemen zum Einsatz kommen, die, mit unterschiedlichen Autonomiegraden, speziell für die Generierung von Inhalten wie komplexem Text, Bildern, Audio oder Video bestimmt sind („generative KI“), sowie Anbieter:innen, die ein Foundation-Modell in ein generatives KI-System spezialisieren:

a) den in Artikel 52 Absatz 1 dargestellten Transparenzverpflichtungen zu entsprechen;

b) das Foundation-Modell so zu trainieren, und ggf. zu gestalten und entwickeln, dass angemessene Sicherheitseinrichtungen zum Schutz vor der Generierung von Inhalten, die gegen Unionsrecht verstoßen, gewährleistet werden, die dem allgemein anerkannten Stand der Technik entsprechen und die Grundrechte, einschließlich des Rechts auf freie Meinungsäußerung, nicht berühren;

c) unbeschadet einzelstaatlicher oder unionsrechtlicher Rechtsvorschriften zum Urheberrecht eine hinlänglich detaillierte Zusammenfassung der Verwendung urheberrechtlich geschützter Trainingsdaten dokumentieren und öffentlich zugänglich machen.

17) Inwiefern unterscheidet sich die Verteilung von Vor- und Nachteilen durch GPAI zwischen unterschiedlichen Bevölkerungsgruppen (sowohl innerhalb nationaler Gesellschaften als global betrachtet mit Blick auf den globalen Süden/Norden) aufgrund der nachfolgend aufgezählten Aspekte:

- *Unterschiedliche Zugangsmöglichkeiten zur Technologie (z.B. wegen unterschiedlicher technischer, materieller, bildungs- u.a. anderer Voraussetzungen)*
- *Unterschiedliche Repräsentanz in Trainingsdaten (z. B. Gesundheitsdaten von Frauen vs. Männern, von Weißen vs. PoC, afrikanische Sprachen vs. Englisch etc.)*
- *Unterschiedliche Betroffenheit durch stereotype Zuschreibungen und Diskriminierungen (z.B. aufgrund von Geschlecht oder Ethnie)*
- *Unterschiedliche Belastung durch den von KI-Systemen verursachten Ressourcenverbrauch, und wie wäre eine gerechtere Verteilung der Vor- und Nachteile erreichbar?*

Übersetzung aus dem Englischen

Wie auch bei allen anderen Technologien, ist General Purpose AI ein Abbild jener Gesellschaften, in denen sie entwickelt wird, und reproduziert tendenziell Diskriminierungs- und Unterdrückungsstrukturen jener Gesellschaften. GPAI wird zumeist von weißen, gebildeten Programmierern im kapitalistischen Interesse von Aktionär:innen entwickelt. In diesem Prozess erfahren die Interessen von Ökosystemen oder systematisch benachteiligten Gruppen keine gleichberechtigte Berücksichtigung. Oftmals führt dies dazu, dass KI benachteiligte Gruppen diskriminiert, und viele Clickworker im Globalen Süden werden innerhalb des Prozesses der Datenzuordnung für und des Trainings von KI-Modellen ausgebeutet. Technologien wie GPAI verfügen über ein emanzipatorisches Potenzial, Menschen von unnötiger und nicht lohnenswerter Plackerei zu befreien, bei ihrer aktuellen Kurve verwirklicht sie dieses Potenzial aber nicht. Aktuell kommen zumeist jene mit wirtschaftlicher und politischer Macht in den Genuss der Vorteile der KI.

18) Sollte generative KI als Mehrzweck-KI grundsätzlich als Hochrisiko-KI im Sinne der europäischen KI-Verordnung eingestuft werden, um höhere Standards zu erfüllen? Für wie sinnvoll/umsetzbar halten Sie Regulierungsoptionen für generative KI wie Transparenzpflichten zu Trainingsdaten und Trainingsprozessen, die Verpflichtung zum Risikoassessment durch Bereitsteller einer GPAI und dessen Veröffentlichung, sichtbare oder unsichtbare Kennzeichnungen von allen oder bestimmten KI-generierten Inhalten, das Recht auf Überprüfbarkeit der Diskriminierungsfreiheit und den Zugang für Forscher:innen und andere diskutierte Optionen?

Siehe allgemeine Anmerkungen.

Übersetzung aus dem Englischen

ANHANG I: Vorläufige Abschätzung der Anforderungen für Hochrisiko-KI im Anbetracht von GPAI-/GenAI-/FM-Systemen[i]

Artikel 9 (Risikomanagementsystem)

Das in Artikel 9 beschriebene Risikomanagementsystem als kontinuierlicher, sich wiederholender Prozess der Erkennung von Risiken für Gesundheit, Sicherheit und Grundrechte scheint auch für Anbieter:innen von GPAI/GenAI/FM ein recht vernünftiges System zu sein. Vor allem, wenn jene Risiken (über API) nachgelagerte Anwendungen beeinträchtigen können, würde in einem solchen System beschrieben werden, wie Risiken des betreffenden GPAI-/GenAI-/FM-Systems gemanagt werden.

Artikel 10 (Daten und Daten-Governance)

Viele, wenn nicht sogar alle der derzeitigen „GPAI-/GenAI-/FM-Systeme" sind datengetrieben, so dass die Anforderung einer ordnungsgemäßen Daten-Governance hier von zentraler Bedeutung zu sein scheint. Einige erwähnenswerte Aspekte: Absatz 2 Buchst. g berücksichtigt *„die Ermittlung möglicher Datenlücken oder Mängel und wie diese Lücken und Mängel behoben werden können",* was, wie oben erwähnt, eine Lösung des Problems des „Re- oder Up-Trainings" (wenn ein System für einen besonderen Zweck eines Re- oder Up-Trainings bedarf) darstellen könnte.

Absatz 3, in dem die Anforderungen an Trainings-, Validierungs- und Testdaten festgelegt werden, setzt sich aus zwei Teilen zusammen. Einem allgemeinen Teil, der besagt: *„Trainings-, Validierungs- und Testdatensätze müssen relevant, repräsentativ,[im bestmöglichen Maße] fehlerfrei und vollständig sein [und] Sie haben die geeigneten statistischen Merkmale [...]"*. Und einem spezifischen Teil, in dem spezielle Anwendungsfälle oder Bereiche gegebenenfalls eine Reihe zusätzlicher Anforderungen *„bezüglich der Personen oder Personengruppen, auf die das Hochrisiko-KI-System bestimmungsgemäß angewandt werden soll"* auslösen. Die Erfüllung des ersten Teils ist für Anbieter:innen von GPAI/GenAI/FM einfach. Gegebenenfalls, d. h. für vernünftigerweise bekannte Anwendungsfälle oder Bereiche ihres Systems, könnten sie selbst den zweiten Teil erfüllen.

Um die obigen Punkte widerzuspiegeln, könnte Absatz 4 ohne Weiteres angepasst werden: *Die Trainings-, Validierungs- und Testdatensätze müssen, soweit dies für den* ***vernünftigerweise bekannten oder vorhersehbaren Zweck*** *erforderlich ist [...].*

Artikel 11 (Technische Dokumentation)

Angesichts der Verantwortung der Anbieter:innen von GPAI/GenAI/FM erscheint diese Anforderung zugleich vernünftig und sogar wünschenswert.

Artikel 12 (Aufzeichnungspflichten)

Diese Anforderung zielt darauf ab, KI-Systeme so zu konzipieren und entwickeln, dass ihre Arbeitsweise nachvollziehbar ist. Sie zielt ausdrücklich nicht auf die Durchführung tatsächlicher Nachverfolgungsaktivitäten ab. Mit anderen Worten muss das System Aufzeichnungen und Protokollierung technisch ermöglichen. In der Tat wäre dies eine jener Anforderungen, deren Erfüllung für nachgelagerte Anbieter:innen unmöglich wäre, wenn die

Anbieter:innen von GPAI/GenAI/FM diese Funktionen nicht in das System eingebaut hätten. Somit wird diese Anforderung vor allem aus geschäftlicher Sicht für Anbieter:innen von GPAI/GenAI/FM in der Tat sehr relevant, da sie bedeuten würde, dass GPAI-/GenAI-/FM-Systeme ohne ordnungsgemäße Protokollierungsfunktionen nicht für Hochrisiko-KI verwendet werden.

Wir schlagen für Absatz 4 eine Textänderung vor: *Die Protokollierungsfunktionen der in Anhang III Absatz 1 Buchstabe a genannten Hochrisiko-KI-Systeme müssen zumindest* ***ermöglichen*** *(...):*

Artikel 13 (Transparenz)

Diese Anforderung zielt darauf ab, das KI-System so zu konzipieren und entwickeln, dass seine Funktionsweise hinreichend transparent ist. Dies erfordert Nutzungshinweise, eine Änderungsprotokollierung sowie technische Maßnahmen, um die Interpretation der Ergebnisse von KI-Systemen zu ermöglichen. Stellt man GPAI-/GenAI-/FM-Systeme von dieser Anforderung frei, so bleiben sie jene Black Box, die sie oft darstellen, und die Erfüllung dieser Anforderung wird nachgelagerten Anbieter:innen, die GPAI-/GenAI-/FM-Systeme als Komponente eines Hochrisiko-KI-Systems verwenden, extrem erschwert, wenn nicht gar unmöglich gemacht.

Zwei erwähnenswerte Aspekte:

Durch die Anbieter:innen von GPAI/GenAI/FM können, unter Ausnahme der in Artikel 14 Absatz 3 Buchstabe b beschriebenen und in Artikel 13 Absatz 3 Buchstabe d genannten „Maßnahmen zur Gewährleistung der menschlichen Aufsicht“ fast alle Unteranforderungen von Artikel 13 erfüllt werden.

Technische Aufsichtsmaßnahmen (gemäß Art. 14 Abs. 3 Buchst. a) können höchstwahrscheinlich nur im Kern des KI-Systems selbst, also dem GPAI-/GenAI-/FM-System, umgesetzt werden, und nicht im Nachhinein eingebaut werden.

Artikel 14 (Menschliche Aufsicht)

Von dieser Anforderung werden keine tatsächlichen Aktivitäten zur Gewährleistung der menschlichen Aufsicht vorgeschrieben, sie macht es nur zur Bedingung, dass das System so konzipiert ist, dass die Möglichkeit menschlicher Aufsicht gewährleistet ist. Wie zuvor unter Artikel 13 beschrieben können technische Aufsichtsmaßnahmen höchstwahrscheinlich nur auf Ebene der GPAI/GenAI/FM umgesetzt werden.

Stellt man GPAI-/GenAI-/FM-Anbieter:innen von dieser Anforderung frei, so würde die Last der Sicherstellung, dass das System wirksam von Menschen überwacht werden kann, nachgelagerten Nutzer:innen auferlegt werden. Sieht das GPAI-/GenAI-/FM-System diese Möglichkeit nicht vor, so kann sich dies als unmöglich erweisen.

Der einzige Aspekt, der wahrscheinlich nicht von den Anbieter:innen von GPAI/GenAI/FM erfüllt werden könnte, liegt in der Anforderung des 4-Augen-Prinzips gemäß Absatz 5.

Artikel 15 (Genauigkeit, Robustheit und Cybersicherheit)

Vor allem was Cybersicherheit anbelangt regen wir an, diese Sonderanforderung unabhängig von deren Risikoniveau zu einer Pauschalanforderung für alle KI-Systeme zu machen. Im Hinblick auf GPAI-/GenAI-/FM-Systeme können die Anforderungen bezüglich der Genauigkeit und Robustheit auch von GPAI-/GenAI-/FM-Anbieter:innen erfüllt werden, sofern der Begriff der „vernünftigerweise vorhersehbaren Verwendung" in Absatz 1 aufgenommen wird.

Ein erwähnenswerter Aspekt:

Für „dazulernende" KI-Systeme, was so verstanden werden kann, dass sie für eine bestimmte Verwendung ein „Re- oder Up-Training" erhalten, heißt es: *„Hochrisiko-KI-Systeme, die nach dem Inverkehrbringen oder der Inbetriebnahme weiterhin dazulernen, sind so zu entwickeln, dass auf möglicherweise verzerrte Ergebnisse, die durch eine Verwendung vorheriger Ergebnisse als Eingabedaten für den künftigen Betrieb entstehen („Rückkopplungsschleifen"), angemessen mit geeigneten Risikominderungsmaßnahmen eingegangen wird."* Der Artikel als solcher befasst sich bereits teilweise mit dem Problem, dass nicht mit Gewissheit bekannt ist, wie und wo das GPAI-/GenAI-/FM-System zum Einsatz kommen wird.

[i] Bei dieser Abschätzung wird nicht festgestellt, ob eine dieser Anforderungen aus technischer Sicht überhaupt erfüllbar ist. Kann eine Anforderung aufgrund der besonderen technischen Unzulänglichkeiten nicht erfüllt werden, entspricht das System – unabhängig davon, wer für diese Einhaltung verantwortlich ist – nicht dem AIA.