Regulierung von Large Language Models in DSA und AIA-E

Alexander Wehde ist studentische Hilfskraft am Lehrstuhl für Bürgerliches Recht, Informations- und Datenrecht bei Prof. Dr. Louisa Specht-Riemenschneider an der Rheinischen-Friedrich-Wilhelms-Universität Bonn und stellv. Vorstandsvorsitzender der Forschungsstelle für Rechtsfragen neuer Technologien sowie Datenrecht (ForTech) e.V.

Die Regulierung großer Sprachmodelle (engl. Large Language Models) ist diffizil, befindet sie sich doch im Spannungsfeld von Innovationsförderung und mannigfaltigen Befürchtungen über die genaue Verwendung sowie Ausgestaltung solcher Systeme. Letztere erwachsen dabei insbesondere aufgrund der Tatsache, dass Large Language Models in der Lage sind, Text zu erzeugen, der dem Leser den Eindruck vermitteln mag von Menschen geschriebener Text läge vor (sog. ELIZA-Effekt). Vor diesem Hintergrund stellt sich die Frage, inwieweit die angestrebte bzw. bereits verankerte digitalpolitische Regulierung in Gestalt des Digital Services Act (DSA) und des Komissionsentwurfs zum Artificial Intelligence Act (AIA-E) Large Language Models angemessen berücksichtigen, um einen ethischen und verantwortungsbewussten Einsatz dieser in Zukunft sicherzustellen.

I. Large Language Models

Large Language Models wie GPT-3, BERT oder Blender Bot 3 erfreuen sich dieser Tage großer Aufmerksamkeit. Die Bezeichnung dieser als „large" stellt dabei einen Rückbezug auf den sehr großen Datensatz anhand dessen das System trainiert wurde dar. Jener besteht oftmals aus Milliarden von Wörtern und Abfolgen dieser, was es den Large Language Models ermöglicht die Muster und Strukturen menschlicher Sprache verblüffend präzise zu erlernen und Text zu generieren, der dem eines Menschen ähnlich ist. Heute werden große Sprachmodelle daher häufig für die Verarbeitung von Sprach- oder Texteingaben verwendet, wie zB im Rahmen maschineller Übersetzungen oder zur Sprach- und Textgenerierung.

Technisch arbeiten Large Language Models häufig auf Basis sog. Feedforward Networks, die als Unterform künstlicher neuronaler Netzwerke beschrieben werden können (zur Architekur von Large Language Models etwa Vaswani et al., Attention Is All You Need, NIPS 2017). Als solche orientieren sich künstliche neuronale Netzwerke grundsätzlich am Erkenntnisstand der Arbeitsweise des menschlichen Gehirns und bestehen aus miteinander verbundenen künstlichen Neuronen, die Daten verarbeiten und verknüpfen können. Im Kontext von Large Language Models sind diese im künstlichen neuronalen Netzwerk in unterschiedlichen bzw. mehreren „Layern" organisiert, wobei der Input Layer die rohen Eingabedaten des Nutzenden erhält und der Output Layer die Ausgabevorhersage erstellt. Zwischen dem Input- und Output Layer liegen mehrere sog. Hidden Layers, die die Eingabedaten verarbeiten und das Ausgabeergebnis an den Output Layer weiterleiten. Die erzeugte Ausgabe im Output Layer basiert dabei maßgeblich auf den vorhandenen neuronalen Verknüpfungen in den Hidden Layers sowie deren erlernter Gewichtung untereinander bei gewissen Eingaben. Insbesondere der (meist automatisierten und algorithmusbasierten) Anpassung der Gewichtungen zwischen den neuronalen Verknüpfungen in den Hidden Layers kommt daher bei der Fehlerminimierung sowie dem Training von ausgegebenen Informationen, Vorhersagen oder Klassifikationen der Large Language Models entscheidende Bedeutung zu (zu den bekanntesten Optimierungsalgorithmen für Large Language Models gehören dabei Stochastic Gradient Descent, Adam, AdaGrad oder RMSProp).

In Verbindung mit großen Sets an Bild- oder Audiodaten können Large Language Models zudem für Text-in-Bild oder Text-in-Audio Transformationen eingesetzt werden, wie etwa die Anwendung Dall-E zeigt. Large Language Models sind daher nicht nur in ihrem (vermeintlich) primären Anwendungsbereich, der Sprach- und Textgenerierung, sondern auch vor dem Hintergrund sekundärer Anwendungsbereiche, zu denken.

II. Problemkreise von Large Language Models und deren Erzeugnissen

Bezüglich der Verwendung von Large Language Models können mehrere Problemkreise, die hier nur überblicksartig dargestellt werden können, identifiziert werden (umfassend betrachtend Weidinger et al., Ethical and social risks of harm from Language Models, 2021). Ein zentrales Problem liegt bereits darin, dass – trotz häufig überwiegend guter Text- oder Sprachausgaben – der generierte Text unsinnig sein oder grammatikalische Fehler enthalten kann. Dies kann passieren, wenn das Sprachmodell in den Trainingsdaten Muster gelernt hat, die im realen Sprachgebrauch nicht vorkommen, oder weil das System versucht, einen Text zu generieren, der zu lang oder zu komplex für seine derzeitigen Fähigkeiten ist.

Ein weiteres mögliches Problem ist, dass der generierte Text voreingenommene, beleidigende oder diskriminierende Inhalte enthalten kann. Dies geschieht zB dann, wenn das Modell einseitige Muster aus nicht-repräsentativen Trainingsdaten gelernt hat oder wenn es Text auf der Grundlage unvollständiger oder irreführender Eingaben generiert (Brown et al., Language Models are Few-Shot Learners, NeurIPS, 2020; Bender et al., On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, FAccT 2021, 610–623).

Darüber hinaus kann das Trainieren von Large Language Models, in Abhängigkeit zu seiner Größe, Architektur und verwendeter Hardware, viele Rechenressourcen und damit Energie erfordern, was im Kontext von Energieknappheit und Klimaneutralität ein Nachteil ihrer Verwendung und Verhältnismäßigkeit darstellen kann (Strubell et al., Energy and Policy Considerations for Deep Learning in NLP, MIT Technology Review 2019; zum rechtlichen Rahmen s. Rodi, HdB Klimaschutzrecht/Saurer et al., § 19 Rn. 18 ff.). Schließlich bestehen auch Bedenken, dass Large Language Models für haftungsrechtliche Probleme bei deren Primär- wie Sekundärnutzung sowie für Verstöße gegen Urheber-, Straf- oder Datenschutzrecht sorgen könnten (Hilgendorf et al, FAZ v. 9.1.2023; Papastefanou CR 2023, 1 (1 ff.)). So zeigten etwa Tests an einer Vorgängerversion von ChatGPT, dass diese unbeabsichtigt personenbezogenen Trainingsdaten preisgab (Carlini et al., Extracting Training Data from Large Language Models, UsenixSec, 2021).

III. Pflichtenkreis für Large Language Models nach dem DSA sowie dem AIA-E

Unmittelbar werden Large Language Models oder deren Erzeugnisse weder im Digital Services Act (DSA) noch im Entwurf des Artifical Intelligence Act (AIA-E) adressiert. Jedoch enthalten beide Regulierungen Ansätze, bestimmte KI-Systeme oder deren Erzeugnisse zu regulieren. Hierbei wählt der AIA einen risikobasierten Ansatz während der DSA spezifische KI-Systeme bzw. Inhalte auf Online-Plattformen und -Suchmaschinen in den Blick nimmt.

1. DSA und Large Language Models

Vor dem Hintergrund einer Regulierung von KI-Systemen findet sich im DSA ein durchaus weites Feld an Bestimmungen, wobei für Large Language Models im Folgenden insbesondere das Verbot von Dark Patterns in Art. 25 DSA sowie die spezielle Transparenzvorschrift für sog. Empfehlungssysteme in Art. 27 Abs. 1 DSA beleuchtet werden. Letztere werden dabei gem. Art. 3 lit. s DSA als „vollständig oder teilweise automatisierte Systeme definiert, die von einer Online-Plattform verwendet werden, um auf ihrer Online-Schnittstelle den Nutzern bestimmte Informationen vorzuschlagen oder diese Informationen zu priorisieren". Zwar vermögen Large Language Models nach etwaiger Texteingabe als zumindest teilweise automatisiert ablaufend betrachtet werden, jedoch basieren sie in der Praxis regelmäßig (noch) auf einer gewissen Mitwirkungshandlung des interagierenden Nutzers, welche in der Definition gerade nicht abgebildet wird. Dies macht die Anwendung von Art. 27 DSA auf Large Language Models zumindest holprig, da vielmehr vom Betreiber einer Online-Plattform aus gedacht wird. Eindeutiger könnten daher Large Language Models unter Art. 27 DSA fallen, bei denen der jeweilige Input Layer sich aus zuvor über die Online-Plattform erhobenen (personenbezogenen oder nicht-personenbezogenen) Daten „selbst bedient" und hierauf basierend etwa Werbung mit für den jeweiligen Nutzer besonders ansprechendem Werbetext schaltet.

In diesem Fall scheint eine Subsumtion von eingesetzten Large Language Models unter den Begriff der Empfehlungssysteme iSv Art. 3 lit. s DSA durchaus möglich, was Online-Plattformen gem. Art. 27 Abs. 1 DSA zunächst dazu verpflichten würde die Parameter auf welchen die (vorhergesagten) Empfehlungen basieren in den AGB zu erläutern. Auszuklammern wäre für den Betreiber einer Online-Plattform, welche solche Large Language Model nutzen, ferner Werbung, welche auf Profiling iSv Art. 4 Nr. 4 DS-GVO beruht und sich personenbezogener Daten iSv Art. 9 Abs. 1 DS-GVO bedient (Art. 26 Abs. 3 DSA). Inwiefern diese Regelungen auch für Online-Suchmaschinen gelten ist noch ungewiss (vgl. Dregelies MMR 2022, 1033 (1038)). Faktisch spielen Large Language Models im Bereich der Empfehlungssysteme jedoch noch keine beherrschende Rolle auf Online-Plattformen, nimmt hier doch auch das sog. Collaborative Filtering weiterhin eine entscheidende Rolle ein (Wei et al., Collaborative filtering and deep learning based recommendation system for cold start items, Expert Systems with Applications, 2017, S. 29–39).

Auch vom in Art. 25 DSA normierten Verbot von Dark Patterns auf Online-Plattformen, also „Praktiken, mit denen darauf abgezielt oder tatsächlich erreicht wird, dass die Fähigkeit der Nutzer, eine autonome und informierte Entscheidung oder Wahl zu treffen, erheblich verzerrt oder beeinträchtigt wird" (Erwägungsgrund 67 DSA) können Large Language Models miterfasst sein. So können Large Language Models, die zB auf einem Datensatz von Website-Popup-Nachrichten trainiert wurden und nun aufgefordert werden, einen Text für eine neue Popup-Nachricht zu erstellen, einen Text erzeugen, der manipulative Sprache verwendet, um die Benutzer dazu zu verleiten, auf eine Schaltfläche zu klicken oder persönliche Daten preiszugeben. Inwieweit dieses nun auch im DSA verankerte Verbot gegenüber den bestehenden Verboten diesbezüglich in der UGP-RL und der DS-GVO einen sinnvollen Ansatz darstellt, wird letztlich vor allem auch davon abhängen, inwiefern die Durchsetzung über Art. 25 DSA Vorteile mit sich bringt, welche die Mitgliedstaaten bei Verstößen eher dazu verleiten Sanktionen nach Art. 52 DSA zu erlassen.

Hinzutreten im DSA Bestimmungen für Erzeugnisse von automatisiert arbeitenden Bots auf Online-Plattformen, welche als systemische Risiken iSv Art. 34 DSA genauso anerkannt werden (Erwägungsgrund 84 DSA), wie rechtswidrige Inhalte, welche potenziell auch von Large Language

Models stammen können und entlang des zu etablierenden Notice-and-takedown-Verfahrens zu löschen oder zu sperren sind (Art. 6 Abs. 1 lit. b DSA). Zur Erforschung dieser systemischen Risiken, sein sie durch den Betreiber oder den Nutzern einer Online-Plattform bzw. -Suchmaschine gesetzt, birgt zudem Art. 40 DSA ein wichtiges Datenzugangsrecht für Behörden sowie die Forschung (hierzu Wehde MMR 2022, 827 (827 ff.)). Dieses mag zukünftig wichtige Detaileinblicke in das Funktionieren dieser Systeme liefern.

2. Kommissionsentwurf des Artifical Intelligence Act und Large Language Models

a) General-purpose AI systems (GPAI)

Unter den vom Kommissionsentwurf gewählten risikobasierten Regulierungsansatz im AI Act, der sich an sektoralen und einsatzbezogenen Verwendungen von KI-Systemen orientiert, können Large Laguage Models aufgrund ihrer oftmals multiplen use-cases nur schwer subsumiert werden (vgl. Art. 6 Abs. 2 AIA-E iVm Anhang III AIA-E). Kritik hieran forderte daher Nachbesserungen, welche auch eine Regulierung für sog. general-purpose AI systems (GPAI) vornimmt. Die meisten vorgeschlagenen Definitionen für GPAI's (vgl. Kompromissentwürfen des Europäischen Rates sowie die Stellungnahme des JURI-Ausschusses) beschreiben diese nun als KI-Systeme, die „allgemein anwendbare Funktionen" in „einer Vielzahl von Kontexten" erfüllen, womit Bezug auf die Zweckoffenheit von Large Language Models genommen wird.

Zwar stößt das Eingehen des Rates auf Large Language Models im vorgelegten Kompromissentwurf zum AIA-E grundsätzlich auf positives Echo, jedoch wird insbesondere die Definition der GPAI's als zu vage kritisiert sowie chilling effects in der Entwicklung von Open-Source GPAI's prognostiziert (Engler, The EU's attempt to regulate open-source AI is counterproductive, 2022). Gestützt wird letztere Kritik dabei auf den vorgeschlagenen Pflichtenkreis für Entwickler von GPAI's, der auf Anforderungen für KI-Hochrisikosystemen aufsetzt und dadurch nicht tragbare Haftungsrisiken für die häufig finanzschwächere Open-Source Community kreiere. Diese Einwände erscheinen berechtigt, greifen doch auch die Ausnahmetatbestände von diesem Pflichtenkreis nur für GPAI's, die ausschließlich Forschungszwecke verfolgen sowie, wenn ihre Entwickler den Missbrauch des Systems verhindern können. Die Ausnahmetatbestände greifen daher im aktuellen Ratsentwurf für die auch im gesamtgesellschaftlichen Interesse stehenden Open-Source GPAI's häufig nicht und lösen Haftungsrisiken aus.

b) Transparenzpflicht für Deepfakes bzw. entsprechenden Erzeugnissen von Large Language Models

Partiell lassen sich aber auch die im AIA-E bereits vorgesehenen Transparenzregelungen auf Large Language Models anwenden. Dies betrifft insbesondere Art. 52 Abs. 3 AIA-E, welcher für Nutzer eines KI-Systems, das Bild-, Ton- oder Videoinhalte erzeugt oder manipuliert, die wirklichen Personen, Gegenständen, Orten oder anderen Einrichtungen oder Ereignissen merklich ähneln und einer Person fälschlicherweise als echt oder wahrhaftig erscheinen würden („Deepfake") eine Offenlegungspflicht dieser Inhalte statuiert ohne die Art und Weise hierfür genau zu konkretisieren (krit. auf Grund der Nutzerfokussierung sowie der offengelassenen Art und Weise der Offenlegung Kumkar/Rapp ZfDR 2022, 199 (224); Hinderks ZUM 2022, 110 (118); s. hierzu auch die Änderungsvorschläge in der Stellungnahme des JURI-Ausschusses). Dabei ist zu beachten, dass sich Art. 52 Abs. 3 AIA-E in eine Reihe ähnlicher Normen im EU-Verhaltenskodex für Desinformation oder auch im Medienstaatsvertrag stellt, die als Offenlegungs- und

Kennzeichnungspflichten ebenfalls vielfach als auslegungsbedürftig und unzureichend konzipiert empfunden wurden (Kalbhenn ZUM 2021, 663 (670 f.)).

Während rein synthetische Texterzeugnisse nicht unter den Begriff eines Deepfakes iSv Art. 53 Abs. 3 AIA-E fallen, kann unter Zuhilfenahme von Large Language Models, als Teilsystem von Anwendungen wie Dall-E, entstandenes Bild-, Audio- oder Videomaterial ein Deepfake darstellen.

Vor dem Hintergrund unterschiedlicher Auffassungen darüber, was ein KI-System bzw. Deepfakes auszeichnet, ist der Regulierungsansatz jedoch auch hinsichtlich seiner Regulierungssubjekte nicht unumstritten. Zwar scheint Einigkeit darüber zu bestehen, dass ein Deepfake zumindest die Verwendung von KI-basierter Technologie und die Absicht zu täuschen in sich trägt, doch sind darüber hinaus gehende Charakteristiken zumeist mit praktischen Problemen verknüpft. So erscheinen insbesondere die Grenzen zwischen tiefgreifenden Fälschungen und simplen audiovisuellen Manipulationen schwierig zu ziehen (Paris/Donovan, Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence, 2019, S. 10 ff.).

Kritik trifft den Kommissionsentwurf ferner, da er für Art. 52 Abs. 3 AIA-E, also der Transparenz von Deepfake-Technologien, keine ausdrückliche Sanktion bei Nichteinhaltung normiert, was als schwacher Anreiz für die Einhaltung der Vorschrift angesehen wird. Zwar droht über die mitgliedstaatlichen Aufsichtsbehörden das Risiko, dass KI-Systeme mit begrenztem Risiko einer Neubewertung des Risikos über Art. 67 AIA-E ausgesetzt sind, doch stellt an dieser Stelle die enge Verbindung zwischen Deepfake-Erkennungs- und Deepfake-Erstellungssystemen eine Herausforderung bei der Durchführung der Risiko(neu)bewertungen dar (vgl. Anhang III Nr. 6 lit. c AIA-E, der Deepfake-Erkennungssysteme im Rahmen der behördlichen Strafverfolgung als KI-Systeme mit hohem Risiko aufführt, während Deepfakeerstellungssysteme keine Berücksichtigung finden).

Zuletzt scheint unklar, wer die Transparenzverpflichtung in Art. 52 Abs. 3 AIA-E durchsetzt bzw. kompetenzrechtlich vorgesehen ist zu entscheiden, welche Inhalte unter einen Deepfake fallen und wie die in Art. 52 Abs. 3 AIA-E beinhalteten Ausnahmen zu Zwecken der Ausübung der Meinungs-, Kunst- oder Wissenschaftsfreiheit auszulegen sind (hierzu u.a. Hinderks ZUM 2022, 110 (114)). Möglicherweise soll an dieser Stelle auch an Art. 69 AIA-E angeknüpft werden, der darauf abzielt, Verhaltenskodizes als Mittel zur freiwilligen Einhaltung der Anforderungen des AIA-E für nicht-hochriskante KI-Systeme zu etablieren.

IV. Zusammenfassung und Ausblick

Large Language Models liefern bereits heute teils beeindruckende Ergebnisse im Bereich der Text- und Sprachgenerierung, was menschliche Tätigkeiten sinnvoll unterstützen kann. Gerade neuere Large Language Models wie Dall-E zeigen jedoch, dass auch darüberhinausgehende Use-Cases denkbar sind. So sind in der Zukunft gerade bessere Text-in-Bild- oder Text-in-Audio-Anwendungen zu erwarten, die Texteingabedaten mit Bild- und Audiodaten verknüpfen. Auch scheinen Large Language Models vermehrt auf den Markt zu kommen, die ausschließlich dafür verwendet werden, synthetische Daten für das Training anderer maschineller Lernmodelle zu erzeugen (sog. data augmentation), was etwa in der medizinischen Forschung nützlich sein kann, wenn die realen Datenlage begrenzt ist.

Hinsichtlich der regulatorischen Berücksichtigung von Large Language Models im DSA und den aktuellen Entwürfen zum AI-Act finden sich im Ergebnis nur wenige Bestimmungen, welche diese oder deren Erzeugnisse direkt adressieren bzw. regulieren würden. Während im Rahmen Letzteren noch unklar scheint, inwiefern Large Language Models bzw. GPAI's überhaupt unmittelbar reguliert werden sollen, beinhaltet der DSA zwar partiell wichtige Ergänzungen und Transparenzinstrumente, bleibt jedoch aufgrund seines Bezugs auf Online-Plattformen und -Suchmaschinen limitiert. Auch deshalb erscheint es wenig verwunderlich, dass bereits heute weitergehende Regulierungsmöglichkeiten insbesondere für Erzeugnisse von Large Language Models diskutiert werden (spezifisch zu Deepfakes durch KI-Systeme etwa Kumkar/Rapp ZfDR 2022, 199 (225 ff.); Linardatos, LTO v. 27.7.2021), welche durch die im AIA-E angestrebte (Erwägungsgrund 2, S. 3 ff.) bzw. im DSA (Erwägungsgrund 9, S. 2) bestehende Sperrwirkung nationaler Gesetzgeber befeuert wird. Zwar besteht auch über diverse Selbstverpflichtungsmaßnahmen etwaiger Hersteller oder Nutzer von Large Language Models die Chance damit verbundene Risiken und Gefährdungen einzudämmen, jedoch vermögen diese einen rechtlichen Flickenteppich hervorzurufen, was für eine Regulierung dieser Systeme im AI-Act spricht. Hierbei erscheint nicht nur essentiell klar zwischen Herstellern, Nutzenden von Large Language Models in Sekundärmärken (wie etwa Plug-In-Herstellern) und Verbrauchern zu differenzieren, sondern auch transparenzschaffende Pflichten dieser Gruppen mitzudenken (vgl. Bertuzzi, EURACTIV v. 10.1.2023). Technisch sollten Trainingsdaten von Large Language Models daher bereits bei der Entwicklung konsequent mit den Quellenangaben, also Metadaten, angereichert sein, um eine menschliche Nachvollziehbarkeit und Kontrolle der letztendlich ausgegebenen synthetischen Erzeugnisse zu ermöglichen sowie technische Selbsterklärung zu fördern.