Jutta Jahnel/Reinhard Heil

KI-Textgeneratoren als soziotechnisches Phänomen

Ansätze zur Folgenabschätzung und Regulierung

Abstract: AI text generators enable the automated generation of high-quality text that we might consider to be meaningful and human-created content. It is claimed that they represent a „game changer“ for science, business and society, as they would have a fundamental impact on the way we write, think and work. For a realistic assessment of the socio-technical phenomenon of AI text generators, the specific contexts of usage and concrete applications are crucial. This is reflected in the current draft of the AI Act which proposes a risk-based approach for assessing different AI applications. Technology assessment (TA) considers ethical and social implications of technologies and develops guidance for policymakers and society from an interdisciplinary perspective. The risks of AI text generators lie mainly in the area of data protection, discrimination and the possible spread of misinformation. Furthermore, the malicious use of this technology causes new types of fraud and cyber risks. This contribution uses the classification of possible risks at different levels for individuals, organizations and society to systematize the necessary multidimensional regulatory measures. These include, in particular, transparency obligations to combat deception and manipulation and to enhance the ability for a critical evaluation of generated content. In addition, broader societal impacts on communication and democratic processes need to be addressed. In general, the regulation of AI text generators includes technology-neutral approaches to ensure that fundamental rights such as data protection and copyright are respected. In addition, specific AI rules and requirements for foundation models are being developed at the European level.

1 Einleitung

Mit neuen Methoden der Künstlichen Intelligenz (KI) ist die Erstellung und Veränderung von Medieninhalten wie Bilder, Texte, Audio- oder Videoinhalte einfach und kostengünstig möglich.[1] KI-basierte Chatbots erzeugen Texte, die sich von

1 Mariëtte van Huijstee et al., „Tackling Deepfakes in European Policy. Study. Panel for the Future of Science and Technology“, *Publications Office of the European Union*, 2021 (DOI: 10.2861/325063).

 https://doi.org/10.1515/9783111351490-021

Menschen geschriebenen Texten nur noch sehr schwer unterscheiden lassen.[2] Derartige KI-Textgeneratoren, wie bspw. ChatGPT,[3] zählen zu den anschaulichen Anwendungsbeispielen der sogenannten „generativen KI". Diese Systeme bieten enorme Chancen und viele Experten gehen von einem hohen disruptiven Potential aus, da sie die Art und Weise, wie wir denken, kommunizieren und arbeiten, erheblich verändern könnten.[4]

Der vorliegende interdisziplinäre Tagungsband beschäftigt sich mit den möglichen Konsequenzen durch computergenerierte Texte. Dazu zählen insbesondere Risiken, mit denen wir als Nutzer von Textgeneratoren oder auch als Empfänger von generierten Inhalten umgehen müssen.[5] Für die Entwicklung geeigneter Maßnahmen, die mögliche Risiken eindämmen sollen, ist es zunächst wichtig, die Funktionsweise und Eigenschaften der zugrundeliegenden großen Sprachmodelle zu kennen und die möglichen Folgen in konkreten Nutzungs- und Anwendungskontexten zu betrachten. Dabei ist die gesamte Wertschöpfungskette von generierten Texten und die dabei eingebundenen Akteure in den Blick zu nehmen. Neben den Entwicklern von Modellen für Textgeneratoren sind dies auch die Anwender der Modelle, die Programme wie ChatGPT oder Suchmaschinen betreiben, und auch die Nutzer dieser Anwendungen. Denn die Risiken von KI-Textgeneratoren liegen nicht nur in der Konzeption der Sprachmodellen selbst, sondern in der Art und Weise, wie solche Modelle und Systeme konkret genutzt werden. Dabei spielen in hohem Maße auch Risiken durch Missbrauch für böswillige Zwecke eine Rolle.

---

2 Stephen Wolfram, „What Is ChatGPT Doing … and Why Does It Work?", 2023, in [https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work] (Zugriff: 03.10.2023).

3 OpenAI, „ChatGPT", 2023, in [https://chat.openai.com] (Zugriff: 03.10.2023).

4 Ryan Morrison, „OpenAI's New Chatbot ChatGPT Could Be a Game-Changer for Businesses", *Tech Monitor*, 2022, in [https://techmonitor.ai/technology/ai-and-automation/chatgpt-openai-chatbot] (Zugriff: 03.10.2023). Mubin Ul Haque et al., *‚I think this is the most disruptive technology': Exploring Sentiments of ChatGPT Early Adopters using Twitter Data*, arXiv 2022 (DOI: 10.48550/arXiv.2212.05856).

5 Steffen Albrecht, „TAB-Hintergrundpapier Nr. 26: ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen", *Karlsruher Institut für Technologie-Bibliothek*, 2023 (DOI: DOI: 10.5445/IR/1000158070); Laetitia Ramelet, „ChatGPT – Themenpapier: ChatGPT – wenn die künstliche Intelligenz schreibt wie ein Mensch. Und was es dabei zu beachten gilt", *TA-SWISS*, 2023, in [https://www.ta-swiss.ch/chatgpt] (Zugriff: 03.10.2023); Europol, „ChatGPT - the Impact of Large Language Models on Law Enforcement", a Tech Watch Flash Report from the Europol Innovation Lab, Publications Office of the European Union, Luxembourg, 2023, in [https://www.europol.europa.eu/publications-events/publications/chatgpt-impact-of-large-language-models-law-enforcement] (Zugriff: 03.10.2023).

2 Begriffe und Funktionsweise von KI-Textgeneratoren

Zu den prominentesten Beispielen von KI-Textgeneratoren zählt ChatGPT, ein KI-Chatbot des Unternehmens OpenAI.[6] Ein Chatbot ist ein textbasiertes Dialogsystem, über das sich in natürlicher Sprache mit einem technischen System kommunizieren lässt. Herzstück solcher KI-Chatbots sind Sprachmodelle, sogenannte „Large Language Models" (LLM). Derartige KI-Textgeneratoren basieren auf dem Maschinellen Lernen mit neuronalen Netzen, dem sogenannten Deep Learning. Die Sprachmodelle, sogenannte „Generative Pre-trained Transformer" (GPT), werden anhand einer großen Anzahl von Texten vortrainiert.[7]

Textgeneratoren wie ChatGPT arbeiten auf Grundlage von Wortfolgestatistiken,[8] d. h. sie berechnen ihre Ausgabe anhand der Wahrscheinlichkeit, mit der im zum Trainieren genutzten Textkorpus ein Wort auf das andere folgt. OpenAI nutzt zusätzlich das sogenannte bestärkende Lernen aus menschlichem Feedback („Reinforcement Learning from Human Feedback"). Menschen bewerten die Ausgaben des Chatbots und deren Feedback wird zum Feintuning des Modells genutzt.

Linguistisch plausible Texte von Textgeneratoren beantworten Fragen somit nicht auf Grundlage von Semantik, Wissen oder ethischer Reflexion, sondern anhand von Wahrscheinlichkeiten.[9] Einige Autoren bezeichnen Sprachmodelle deshalb auch als „unreliable narrator […] untethered to the truth"[10] oder als „Stochastic Parrots"[11]. Die generierten Texte besitzen keine Faktentreue und können sogar frei erfundene Inhalte oder Quellen enthalten (sog. „Halluzinieren"[12]). Da die Funktionsweise von Sprachmodellen und den darauf beruhenden Textgeneratoren nur

6 OpenAI 2023; Mohammad Aljanabi/ChatGPT, „ChatGPT: Future Directions and Open possibilities", *Mesopotamian Journal of Cyber Security*, 2023, 16–17 (DOI: 10.58496); Sascha Lobo, „ChatGPT: Das machtvollste Instrument, das je vom Menschen geschaffen wurde", *DER SPIEGEL*, 2023, in [https://t1p.de/32pzj] (Zugriff: 03.10.2023).

7 Morrison 2022.

8 Vgl. hierzu Wolfram 2023.

9 Viriya Taecharungroj, „‚What Can ChatGPT Do?' Analyzing Early Reactions to the Innovative AI Chatbot on Twitter", *Big Data and Cognitive Computing*, Bd. 7, 2023 (DOI: 10.3390/bdcc7010035).

10 Robert Dale, „GPT-3: What's It Good for?", *Natural Language Engineering*, Bd. 27, 2021, 113–118 (DOI: 10.1017/S1351324920000601).

11 Emily M. Bender et al., „On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, 610–623 (DOI: 10.1145/3442188.3445922).

12 Ziwei Ji et al., „Survey of Hallucination in Natural Language Generation", *ACM Computing Surveys*, Bd. 55, 2023, 1–38 (DOI: 10.1145/3571730).

den wenigsten Anwendern bekannt sein dürfte, ist das Risiko groß, dass es zu falschen Erwartungen und Fehlnutzungen kommt. Dies kann dazu führen, dass die Ausgaben von Textgeneratoren falsch bewertet werden bzw. ihnen zu sehr vertraut wird. Deshalb wird insbesondere vor Anwendungen in sensiblen Feldern wie im Gesundheitsbereich oder im Finanzbereich gewarnt.[13] Da Sprachmodelle außerdem mit einer großen Anzahl an Texten vortrainiert werden, die auch unausgewogene, ethisch problematische oder gar falsche Aussagen beinhalten können, sind die Resultate grundsätzlich unter Vorbehalt zu rezipieren. So wurde insbesondere von Verzerrungen, Vorurteilen und rassistischen sowie sexistischen Bemerkungen durch das Sprachmodell GPT-3 berichtet.[14]

3 Risikodimensionen

Für einen verantwortungsvollen Umgang mit KI-Textgeneratoren ist es von großer Bedeutung zu verstehen, welche Risiken mit ihnen verbunden sind und ob ihr Einsatz gar Grund- und Menschenrechten gefährdet. Wir verstehen im Folgenden Risiko als Kombination der Wahrscheinlichkeit des Auftretens eines Schadens und der Schwere dieses Schadens, wobei eine Einzelperson, eine Vielzahl von Personen oder eine bestimmte Personengruppe beeinträchtigt werden kann (vgl. Artikel 3 (1) Nr. 1a und 1b des KI-Gesetzes).[15] Da Risiken kontextabhängig sind, muss einerseits das Risiko durch die Art und Weise, wie diese Systeme konzipiert sind, andererseits aber auch die konkreten Anwendungen dieser Systeme betrachtet werden (Erwägungsgrund 58 a des KI-Gesetzes). Somit sind unterschiedliche Akteure im Lebenszyklus von generierten Inhalten zu berücksichtigen. Die Entwickler von Sprachmodellen, auch als Basismodelle bezeichnet, haben eine besondere Rolle, da sie für die Konzeption des Lern- oder Trainingsprozesses und die Auswahl der dafür

13 Jon Christian, „Magazine Publishes Serious Errors in First AI-Generated Health Article“, *Neoscope*, 2023, in [https://futurism.com/neoscope/magazine-mens-journal-errors-ai-health-article]. (Zugriff: 13.09.2023).

14 Luciano Floridi/Massimo Chiriatti, „GPT-3: Its Nature, Scope, Limits, and Consequences“, *Minds and Machines*, Bd. 4, 2020, 681–694 (DOI: 10.1007/s11023–020–09548–1); Dale 2021; Chris Stokel-Walker/Richard Van Noorden, „What ChatGPT and Generative AI Mean for Science“, *Nature*, Bd. 614, 2023, 214–216 (DOI: 10.1038/d41586–023–00340–6).

15 Europäische Kommission, *Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union*, Brüssel 2021, in [https://t1p.de/yqlrn] (Zugriff: 03.10.2023) in der Fassung der Abänderungen des Europäischen Parlaments vom 14. Juni 2023 (COM(2021)0206 – C9–0146/2021–2021/0106(COD)), 2023, in [https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_DE.html] (Zugriff: 03.10.2023).

verwendeten Daten verantwortlich sind. Werden diese Basismodelle jedoch als Dienstleistung über einen entsprechenden Zugang an Betreiber weiterer KI-Systeme bereitgestellt, zum Beispiel als Teil von Suchmaschinen, dann sind insbesondere die Betreiber dieser KI-Systeme für mögliche Folgen des generierten Outputs in die Pflicht zu nehmen. Denn es ist davon auszugehen, dass die Betreiber den spezifischen Verwendungskontext und die daraus resultierenden potentiellen Risiken besser kennen als die Entwickler der Modelle. Schließlich können die Nutzer dieser KI-Systeme die generierten Inhalte in sozialen Medien an weitere Empfänger verbreiten und dadurch weitreichende systemische Risiken auslösen.

Die Risiken von Textgeneratoren und Basismodellen umfassen im Allgemeinen finanzielle oder psychologische Folgen einer möglichen Diskriminierung oder Missachtung des Datenschutzes. Weiterhin sind neben den traditionellen „Safety-Risiken“ auch „Security“-Aspekte zu berücksichtigen, d. h. Risiken durch intentionale Bedrohungen und durch kriminellen Missbrauch.[16] Aufgrund des breiten sektorübergreifenden Anwendungspotentials von Sprachmodellen sind deren mögliche Risiken jedoch nicht umfassend bekannt und abschätzbar. Im Folgenden wird in Anschluss an Luhmann eine Kategorisierung in drei verschiedenen Ebenen (Mikro-, Meso- und Makro-Ebene) vorgenommen.[17] Diese Risikoeinteilung wurde schon an anderer Stelle als hilfreiche Systematik zur Identifizierung vertrauensbildender Maßnahmen digitaler Systeme angewendet.[18] Dazu werden die Risiken für Individuen auf der Mikro-Ebene, die Risiken für Organisationen auf der Meso-Ebene und die gesamtgesellschaftlichen Risiken auf der Makro-Ebene betrachtet.

3.1 Risiken auf individueller Ebene (Mikro-Ebene)

In diese Dimension fallen Risiken durch die Verletzung von grundlegenden Rechten natürlicher Personen, also zum Beispiel von Persönlichkeitsrechten oder Datenschutzrechten. Ein Schaden bezieht sich auf private Interessen und kann materieller oder immaterieller Art sein, einschließlich physischer oder psychischer Schäden. Aufgrund fehlenden Wissens über die Funktionsweise von Sprachmodellen besteht zunächst das Risiko, dass es zu falschen Erwartungen, Bewertungen und damit zu Fehlnutzungen derartiger Systeme kommen kann. So wird beispielsweise über einen Versuch in der Gesundheitsberatung berichtet, in dem das

16 Terje Aven, „A unified framework for risk and vulnerability analysis covering both safety and security“, *Reliability Engineering & System Safety*, Bd. 6, 2007, 745–754 (DOI: 10.1016/j.ress.2006.03.008).
17 Niklas Luhmann, *Soziale Systeme. Grundriß einer allgemeinen Theorie*, Frankfurt a. M. 1984.
18 Anna Hornik et al., „Die Zukunft des Vertrauens in digitale Welten. Studie Kurzfassung“, beauftragt vom BMBF, 2022.

Sprachmodell GPT3 hochriskante Empfehlungen gab.[19] Zu den Risiken, die sich aus dem Trainingsprozess von Sprachmodellen ergeben, zählen mögliche Diskriminierungen von Personen oder Personengruppen. So konnten u. a. in einer Studie zu frühen Sprachmodellen implizite Vorurteile gegenüber Menschen mit Behinderungen aufgezeigt werden.[20]

Weitere Risiken auf der individuellen Ebene ergeben sich durch die illegale Nutzung personenbezogener Daten ohne Erlaubnis der Betroffenen. Nutzer können außerdem selbst Datensicherheitsverletzungen verursachen, wenn diese bspw. durch Fragen und Aufforderungen im Chatverlauf (meist unwissentlich) personenbezogene Daten preisgeben. Da Textgeneratoren auch zur Imitation eines spezifischen Sprachstils von Personen verwendet werden können, eignen sich die resultierenden Texte besonders effektiv zur Täuschung, für betrügerische und kriminelle Handlungen, Diskreditierungen sowie zur Beeinflussung menschlichen Verhaltens („Social Engineering").[21]

Schließlich werden Textgeneratoren tiefgreifende Veränderungen der Arbeit von hochqualifizierten Berufsgruppen wie Programmierer:innen, Journalist:innen oder auch kreativ Tätigen auslösen.[22] Hier besteht das Risiko weniger in der vollständigen Verdrängung menschlicher Arbeit, sondern der Veränderung von Arbeitsbedingungen durch zunehmenden Konkurrenzdruck, beschleunigte Arbeitsprozesse und wachsenden Stress.[23]

3.2 Risiken auf Organisationsebene (Meso-Ebene)

Zu dieser Risikodimension zählen insbesondere Sicherheitsrisiken und finanzielle Risiken für Unternehmen und Wirtschaft. Auch hier besteht das Risiko einer unbeabsichtigten Datensicherheitsverletzung durch die Preisgabe von Firmengeheimnissen im Chatverlauf. Schulungen der Arbeitnehmer hinsichtlich einer kompetenten und rechtskonformen Anwendung sind unabdingbar. Es müssen dabei

19 Ryan Daws, „Medical Chatbot Using OpenAI's GPT-3 Told a Fake Patient to Kill Themselves", *AI News*, 2020, in [https://t1p.de/afcv3] (Zugriff: 03.10.2023).

20 Pranav Narayanan Venkit et al., „A Study of Implicit Bias in Pretrained Language Models against People with Disabilities", *ACL Anthology, Proceedings of the 29th International Conference on Computational Linguistics*, 2022, 1324–1332, in [https://aclanthology.org/2022.coling-1.113] (Zugriff: 03.10. 2023).

21 Europol 2023.

22 Tyna Eloundou et al., *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*, arXiv 2023 (DOI: 10.48550/arXiv.2303.10130).

23 Patrick Dax, „Warum ChatGPT so schnell kein Job-Killer ist", *futurezone*, 2023, in [https://t1p.de/jl5hz] (Zugriff: 03.10.2023).

auch sicherheitsrelevante Strategien gegen mögliche böswillige Cyberangriffe, beispielsweise durch Manipulationen und Täuschungen mit nachgemachten Sprach- und Textstilen erlernt werden. Berücksichtigt man zusätzlich das steigende Ausmaß und die Geschwindigkeit der automatisierten Generierung irreführender Inhalte, erwarten Experten:innen insbesondere ein hohes finanzielles Risiko durch authentisch erscheinende Phishing-E-Mails und andere Betrugsfälle. Da Textgeneratoren nicht nur zur Erstellung von Texten, sondern auch zur Generierung von Programm-Codes in der Softwareentwicklung eingesetzt werden, stellt sich die Frage, wie sich dies auf die Qualität von Softwareprodukten auswirkt.[24]

In einer Literaturstudie zu den Risiken von KI für Organisationen wurden neben konventionellen Angriffen völlig neuartige Bedrohungen wie Nutzerdatendiebstahl und Angriff auf firmeneigene Trainingsdaten identifiziert. Insbesondere Sabotage und Spionage wurden als Angriffszwecke aufgeführt, die auch im Bereich der generativen KI relevant sein können.[25] Dabei ist zu berücksichtigen, dass nicht nur Unternehmen, sondern auch deren Personal und Kunden betroffen sein können.

3.3 Risiken für die Gesellschaft (Makro-Ebene)

Durch die Zunahme von authentisch wirkenden Texten wird das Risiko von Des- und Misinformation verstärkt und dadurch auch Schutzgüter von öffentlichem Interesse, wie der Schutz der Demokratie oder der Rechtsstaatlichkeit, gefährdet.[26] Häufig genannte Beispiele sind der Missbrauch für Propagandazwecke, zur Manipulation von Wahlen oder zur Rekrutierung oder Finanzierung terroristischer Aktivitäten.[27] Da KI-Chatbots immer schwerer von Menschen zu unterscheiden sind, ist fraglich, ob und inwiefern eine unabhängige politische Willensbildung mittels ausgewogener Informationen überhaupt noch möglich ist. Weiterhin kann ein Bias in den für das Trainieren von Sprachmodellen verwendeten Datensätzen soziale Ungleichheiten verstärken oder einzelne Personengruppen diskriminieren. Da zu erwarten ist, dass sich generierte Inhalte immer mehr durchsetzen werden, ist eine zunehmende Standardisierung von Meinungen und Wissen sowie eine Erosion von kultureller und linguistischer Diversität zu befürchten. Letztlich kann

24 Ramelet 2023.

25 Yisroel Mirsky et al., The Threat of Offensive AI to Organizations, arXiv 2021, in [http://arxiv.org/abs/2106.15764] (Zugriff: 03.10. 2023).

26 Greg Noone, „This Is How GPT-4 Will Be Regulated“, *TechMonitor*, 2023, in [https://techmonitor.ai/technology/ai-and-automation/this-is-how-gpt-4-will-be-regulate] (Zugriff: 03.10.2023).

27 Europol 2023; Bender et al. 2021.

auch eine Veränderung weiterer menschlicher Fähigkeiten wie Empathie oder soziale Kompetenz nicht ausgeschlossen werden, was weitreichende Auswirkungen auf die Art und Weise unseres Zusammenlebens haben könnte.[28]

Auf der Makro-Ebene ergeben sich auch ökologische Risiken, da die Entwicklung und der Einsatz großer Sprachmodelle mit einem hohen Stromverbrauch verbunden ist.[29] Nur wenige marktbeherrschende BigTech-Unternehmen können außerdem die ökonomischen Voraussetzungen für die notwendigen Rechenleistungen aufbringen und dominieren zunehmend die gesellschaftlich relevante Forschung und Entwicklung von Sprachmodellen.

Notwendige Regulierungsmaßnahmen stellen auf der Makro-Ebene eine besondere Herausforderung dar, da sie über den weitgehend individualrechtlichen Ansatz der rechtlichen Regulierung hinausgehen und gleichzeitig weitere Grundrechte wie Meinungs- oder Kunstfreiheit abgewogen werden müssen.[30]

4 Regulierung

Die mit der Anwendung von Sprachmodellen verbundenen mehrdimensionalen Risiken sollten durch adäquate Regulierungsmaßnahmen adressiert werden. Es existieren bereits unterschiedliche technikneutrale Regulierungsansätze auf EU und nationaler Ebene im zivilrechtlichen Bereich der Antidiskriminierung insbesondere nach Maßgabe der Europäischen Menschenrechtskonvention,[31] im Bereich des Datenschutzes[32] und des Urheberrechts,[33] aber auch durch das aktuelle Cybersicherheitsrecht.[34] Diese Regulierungen berücksichtigen überwiegend die Mi-

28 Morrison 2022.

29 David Patterson et al., *Carbon Emissions and Large Neural Network Training*, arXiv 2021 (DOI: 10.48550/arXiv.2104.10350).

30 Josh A. Goldstein et al., *Generative Language Models and Automated Influence Operations. Emerging Threats and Potential Mitigations*, arXiv 2023 (DOI: 10.48550/arXiv.2301.04246).

31 Europäische Union, *Charta der Grundrechte der Europäischen Union (GRCh)*, 2012, in [https://dejure.org/gesetze/GRCh] (Zugriff: 16.08.2023).

32 Europäische Union, *Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates – vom 27. April 2016 – zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung)*, 2016, in: [https://t1p.de/1a5b1] (Zugriff: 16.08.2023).

33 Bundesministerium der Justiz, *Gesetz über Urheberrecht und verwandte Schutzrechte (Urheberschutzgesetz)*, 2021, in [https://www.gesetze-im-internet.de/urhg/BJNR012730965.html] (Zugriff: 16.08.2023).

34 *Europäisches Parlament und Rat, Verordnung (EU) 2019/881 des Europäischen Parlaments und des Rates vom 17. April 2019 über die ENISA (Agentur der Europäischen Union für Cybersicherheit)*

kro- und Meso-Ebene der Risiken durch Sprachmodelle, hinken aber der rapiden technologischen Entwicklung oft hinterher. Unklarheit besteht insbesondere bei der Anwendung des Urheberrechtes auf geschützte Trainingsdaten. Aktuell erfolgt in der EU auch ein Trilogprozess zur Entwicklung einer technikspezifischen Regulierung (Gesetz über Künstliche Intelligenz[35]). In dem aktuellen Textentwurf des Europäischen Parlamentes werden Sprachmodelle oder Textgeneratoren unter den Begriffen „Basismodell" und „Generative KI" eingeführt. Da es bislang der einzige rechtliche Regulierungsansatz speziell zu Sprachmodellen und Textgeneratoren darstellt, wird dieses Gesetz im Folgenden detailliert betrachtet.

Die EU-Kommission hat am 21. April 2021 den weltweit ersten Vorschlag eines Rechtsrahmens für KI vorgelegt. Seitdem hat das geplante Gesetz über KI das Gesetzgebungsverfahren der EU weiter durchlaufen und Änderungen durch den Rat der EU und das Europäische Parlament erfahren. Insbesondere die am 14. Juni 2023 verabschiedete Verhandlungsposition des Parlaments sieht wesentliche Änderungen am ursprünglichen Entwurf der Kommission vor.[36] Mit diesem Gesetz soll ein wirksames und verbindliches Regelwerk für KI-Systeme eingeführt werden. Die allgemeinen Grundsätze werden in Artikel 4a (1) des KI-Gesetzes aufgeführt und umfassen folgende Punkte:

- Menschliches Handeln und menschliche Aufsicht
- Technische Robustheit und Sicherheit
- Privatsphäre und Datenqualitätsmanagement
- Transparenz
- Vielfalt, Diskriminierungsfreiheit und Fairness
- Soziales und ökologisches Wohlergehen.

Gewählt wurde ein risikobasierter Ansatz, bei dem Art und Inhalt der Vorschriften

> auf die Intensität und den Umfang der Risiken zugeschnitten werden, die von KI-Systemen ausgehen können. Es ist daher notwendig, bestimmte unannehmbare Praktiken im Bereich der künstlichen Intelligenz zu verbieten und Anforderungen an Hochrisiko-KI-Systeme und Verpflichtungen für die betreffenden Akteure sowie Transparenzpflichten für bestimmte KI-Systeme festzulegen.[37]

---

und über die Zertifizierung der Cybersicherheit von Informations- und Kommunikationstechnik und zur Aufhebung der Verordnung (EU) Nr. 526/2013 (Rechtsakt zur Cybersicherheit), 2019, in [http://data.europa.eu/eli/reg/2019/881/oj/deu] (Zugriff: 03.10.2023).

35 Europäische Kommission, *Gesetz über Künstliche Intelligenz*, 2021.

36 Die nachfolgenden Inhalte und Zitate beziehen sich, soweit nicht anders angegeben, allesamt auf den aktuellen Vorschlag des Europäischen Parlaments zum Gesetz über Künstliche Intelligenz, 2023.

37 Erwägungsgrund 14 des KI-Gesetzes.

KI-Technologien werden im KI-Gesetz in unterschiedliche Risikoklassen eingeteilt und die vorgeschriebenen Maßnahmen und Verpflichtungen daran angeknüpft. Für bestimmte, in Anlage III des Gesetzes, näher spezifizierte Bereiche von sogenannten Hochrisikosystemen werden beispielsweise Konformitätsprüfungen notwendig.

Große vortrainierte Sprachmodelle können als Basismodelle für unterschiedliche Zwecke eingesetzt werden. Das Risiko wird meist erst durch die konkrete Nutzung des Endverbrauchers bestimmt. Dieser dynamische Anwendungskontext erschwert die erforderliche Risikocharakterisierung und damit die Entscheidung, ob in einem spezifischen Fall ein sogenanntes Hochrisiko-KI-System vorliegt.[38]

Weiterhin gibt es im Trilog-Prozess eine Debatte zur Differenzierung zwischen „KI-Systemen mit allgemeinem Verwendungszweck (General Purpose AI)“ und „Basismodellen (Foundation Models)“. Der Begriff Basismodell umfasst hoch entwickelte KI-Modelle, einschließlich verschiedener generativer KI-Systeme wie ChatGPT, GPT-4, Bard oder Stable Diffusion (vgl. Artikel 3 (1) Nummer 1 c des KI-Gesetzes). Im Gegensatz dazu ist der vom Europäischen Rat im Dezember 2022 eingeführte Begriff „Allzweck-KI-System“ unschärfer (vgl. Artikel 3 (1) Nummer 1 d des KI-Gesetzes).

Ausschließlich Basismodelle sollen im Rahmen des KI-Gesetzes angemessenen und spezifischeren Anforderungen und Verpflichtungen unterliegen.

> Vortrainierte Modelle, die für eine enger gefasste, weniger allgemeine und begrenztere Reihe von Anwendungen entwickelt wurden und nicht an ein breites Spektrum von Aufgaben angepasst werden können, wie z. B. einfache Mehrzweck-KI-Systeme, sollten für die Zwecke dieser Verordnung nicht als Basismodelle betrachtet werden, da sie besser interpretierbar sind und ihr Verhalten weniger unvorhersehbar ist.[39]

Der aktuelle Vorschlag für Basismodelle unterscheidet nun drei unterschiedliche Ebenen:

- Mindeststandards für Entwickler von Basismodellen und eine zusätzliche Transparenzpflicht bezüglich des Einsatzes dieser Modelle in generativen KI-Systemen (Art. 28 b [4] in Verbindung mit Art. 52 [1] des KI-Gesetzes)
- Spezifische Hochrisiko-Regeln für Anbieter, die das KI-System erheblich verändern, dies gilt nur für generative KI-Modelle, die in konkreten Hochrisiko-Anwendungsfällen eingesetzt werden;
- Regeln für die Zusammenarbeit entlang der KI-Wertschöpfungskette

38 Natali Helberger/Nicholas Diakopoulos, „ChatGPT and the AI Act“, *Internet Policy Review,* Bd. 12, 2023, in [https://policyreview.info/essay/chatgpt-and-ai-act] (Zugriff: 03.10.2023).

39 Erwägungsgrund 60 g des KI-Gesetzes.

Zu den Mindeststandards der Basismodelle zählen beispielsweise Datenschutz, Nichtdiskriminierung, Energieeffizienz, Qualitätsmanagement oder Cybersicherheitsverpflichtungen, die von den Entwicklern dieser Modelle umzusetzen sind, während die spezifischen Hochrisiko-Regeln für Anbieter und professionelle Nutzer gelten sollten. Das KI-Gesetz reguliert somit einerseits die Basismodelle per se anhand von Mindeststandards, aber auch deren Kennzeichnung und Anwendung.

In einer Stellungnahme zum gegenwärtigen Vorschlag wurde neben den Transparenzpflichten der Anbieter generativer Modelle auch Pflichten für Nutzer gegenüber den Empfängern vorgeschlagen, um die Verbreitung von Fake News und Fehlinformationen zu bekämpfen und derartige systemische Risiken aufzufangen.[40] Das KI-Gesetz weist hinsichtlich gesellschaftlicher Risiken durch Misinformation auf die Wichtigkeit der Entwicklung von KI-Kompetenz hin; insbesondere sollen „die Anbieter und Nutzer von KI-Systemen in Zusammenarbeit mit allen einschlägigen Interessenträgern die Entwicklung ausreichender KI-Kompetenzen bei Menschen aller Altersgruppen, einschließlich Frauen und Mädchen, in allen Bereichen der Gesellschaft fördern".[41] Unter KI-Kompetenz werden alle Fähigkeiten und Kenntnisse verstanden, die es Anbietern, Nutzern und Betroffenen ermöglichen, „KI-Systeme in Kenntnis der Sachlage einzusetzen sowie sich der Chancen und Risiken von KI und möglicher Schäden, die sie verursachen kann, bewusst zu werden und dadurch ihre demokratische Kontrolle zu fördern."[42] Eine weitere Konkretisierung erfolgt hier jedoch nicht.

Zum Umgang mit Misinformation ist der Co-Regulierungsansatz zwischen staatlichen und nichtstaatlichen Akteuren im Bereich der Inhaltemoderation von Texten durch das Gesetz über Digitale Dienste[43] heranzuziehen. Dieses Gesetz ist ein prominentes Beispiel zur Regulierung der Verteilung von problematischen Inhalten durch Plattformen mittels eines „Notice and Action"-Verfahrens: Plattformen sind nur dann von einer Haftung für rechtswidrige Inhalte freigestellt, solange sie davon keine Kenntnis haben oder nach Kenntniserlangung unverzüglich reagieren und die Inhalte entfernen. Sie werden allerdings nicht dazu verpflichtet, die Inhalte

40 Philipp Hacker/Andreas Engel/Marco Mauer, „Regulating ChatGPT and Other Large Generative AI Models", *ACM Conference on Fairness, Accountability, and Transparency*, 2023, 1112–1123 (DOI: 10.1145/3593013.3594067). Deutscher Bundestag, „Anhörung zum Thema ‚Generative Künstliche Intelligenz'", 24.05.2023, in [https://t1p.de/zmgrt] (Zugriff: 03.10.2023).

41 Siehe Erwägungsgrund 9 b des KI-Gesetzes.

42 Erwägungsgrund 9 b des KI-Gesetzes.

43 Europäische Kommission, *The Digital Services Act Package*, 2021, in [https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package] (Zugriff: 03.10.2023); Conseil de l'Union européenne, *Regulation (EU) of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC*, 15.06.2022, in [https://data.consilium.europa.eu/doc/document/ST-9342-2022-INIT/x/pdf] (Zugriff: 03.10.2023).

aktiv zu überwachen. Zusätzlich müssen die Betreiber sehr großer Plattformen (u. a. Facebook, Twitter) evaluieren, welche systemischen Risiken für Schutzgüter wie die Grundrechte der Nutzer oder für demokratische Wahlen von ihren Diensten ausgehen, gegebenenfalls Gegenmaßnahmen ergreifen und darüber Bericht erstatten.

Eine ähnliche Regelung zur Evaluierung von systemischen Risiken durch Anbieter von Sprachmodellen wird aktuell von mehreren Expert:innen gefordert. Unter den Anwendungsbereich des Gesetzes über Digitale Dienste fallen ausschließlich sehr große Plattformen im Bereich der sozialen Medien, so dass zwar die Verteilung illegaler durch Menschen generierte Inhalte dem „Notice and Action"-Verfahren unterliegt. Die Risiken der mit Sprachmodellen generierten Texte per se bleiben aber unbeachtet. Um dieser Regelungslücke zu begegnen, ist vorgeschlagen worden, einige der Verpflichtungen des Gesetzes über Digitale Dienste auch auf die Entwickler generativer KI auszuweiten. Dabei käme ein analoger Melde- und Aktionsmechanismus nach dem Gesetz über Digitale Dienste in Frage (vgl. Artikel 16 DSA). Die Nutzer von Sprachmodellen hätten dann die Möglichkeit, problematische Inhalte zu markieren und zu melden. Diese Meldungen könnten dann an die Entwickler und/oder Bereitsteller der Modelle weitergeleitet werden.[44]

Neben der rechtlichen Regulierung unterziehen sich Entwickler großer Sprachmodelle, wie z. B. OpenAI, auch freiwilligen Selbstverpflichtungen.[45] Hier werden u. a. Anforderungen für die Nutzung im Gesundheitsbereich, als Coaching und im Finanz- und Nachrichten-Bereich gestellt. Weiterhin beziehen sich Richtlinien, Leitlinien und Verhaltenscodizes auf die Problematik systemischer Risiken auf der Makro-Ebene. Im Verhaltenscodex gegen Desinformation verpflichten sich unterzeichnende Unternehmen zu Maßnahmen gegen Desinformation.[46] Hier werden insbesondere die Verbesserung der Medienkompetenz und Vertrauenswürdigkeit digitaler Inhalte und die Kooperation mit Faktenprüfer genannt. Die Unterzeichner sprechen sich für Instrumente zur Bewertung der Herkunft und Authentizität digitaler Inhalte durch Nutzer aus. Diese Instrumente und Maßnahmen sollen die Resilienz demokratischer Gesellschaften erhöhen.

Die Risiken von Sprachmodellen bezüglich generierter Misinformation werden im Gegensatz zu den Risiken auf der Mikro- und Meso-Ebene aktuell hauptsächlich durch freiwillige Maßnahmen und Selbstverpflichtungen aufgefangen. Es ist unklar, ob diese Maßnahmen ausreichen oder eine rechtliche Regulierungslücke besteht.

---

44 Philipp Hacker/Andreas Engel/Marco Mauer, *Regulating ChatGPT and other Large Generative AI Models*, arXiv 2023 (DOI: 10.48550/arXiv.2302.02337); Deutscher Bundestag 2023.

45 Open AI 2023.

46 Europäische Kommission, Strengthened Code of Practice on Disinformation, 2022, in [https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation] (Zugriff: 03.10.2023).

5 Ausblick und Empfehlungen

Durch die Anwendung von KI-Textgeneratoren kommen neue Fragen zur Zukunft unserer Kommunikation, unseres Arbeitens und unseres Zusammenlebens auf. Zur Beantwortung dieser Fragen benötigen wir ein breites und interdisziplinäres Verständnis der Funktionsweise und der damit verbundenen komplexen Risiken durch Textgeneratoren.

Gegenwärtig ist das rechtliche Umfeld rund um KI eher von Unsicherheit und Unklarheit geprägt. Die Regulierung von Textgeneratoren stellt zudem eine besondere Herausforderung dar, weil es eine Vielzahl von Akteuren im Lebenszyklus generierter Inhalte gibt. Neben den Entwicklern und Anbietern von Sprachmodellen sind dies auch die Anwender von Sprachmodellen und Betreiber von Programmen zur Textgenerierung, sowie die Nutzer der Textgeneratoren und die Social Media Plattformen, über die die generierten Inhalte verteilt werden können. Bei der Entwicklung geeigneter Maßnahmen sollte ein breites Spektrum von Stakeholdern einbezogen werden, um eine möglichst umfassende und ausgewogene Sichtweise zu gewährleisten, damit Risiken auf der Mikro-, Meso- und Makro-Ebene gleichermaßen berücksichtigt werden.