Leopoldina
Nationale Akademie
der Wissenschaften

2024 | Diskussion Nr. 34

Generative KI – jenseits von Euphorie und einfachen Lösungen

Judith Simon | Indra Spiecker gen. Döhmann | Ulrike von Luxburg

Impressum

Herausgeber
Deutsche Akademie der Naturforscher Leopoldina e. V.
– Nationale Akademie der Wissenschaften –
Präsident: Prof. (ETHZ) Dr. Gerald H. Haug
Jägerberg 1, 06108 Halle (Saale)

Redaktion
Christina Hohlbein, Dr. Sebastian Wetterich, Dr. Charlotte Wiederkehr,
Dr. Matthias Winkler
Nationale Akademie der Wissenschaften Leopoldina
Kontakt: politikberatung@leopoldina.org

Lektorat
Jürgen Schreiber, Textkuss – Werkstatt für Sprache und Struktur, Halle (Saale)

Gestaltung und Satz
Klötzner Company Werbeagentur GmbH, Hamburg

Druck
Druck-Zuck GmbH
Seebener Str. 4
06114 Halle (Saale)

DOI
https://doi.org/10.26164/leopoldina_03_01226

Bibliografische Information der Deutschen Nationalbibliothek
Die deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie, detaillierte bibliografische Daten sind im Internet unter https://portal.dnb.de abrufbar.

Zitiervorschlag
Simon, J., Spiecker gen. Döhmann, I. & von Luxburg, U. (2024): Generative KI – jenseits von Euphorie und einfachen Lösungen. Diskussion Nr. 34, Halle (Saale): Nationale Akademie der Wissenschaften Leopoldina.

Redaktionsschluss
September 2024

Generative KI – jenseits von Euphorie und einfachen Lösungen

Judith Simon | Indra Spiecker gen. Döhmann | Ulrike von Luxburg

Inhalt

1 Einleitung

Seit der Veröffentlichung des Textgenerators und Chatbots ChatGPT im Jahr 2022 durch das US-amerikanische Softwareunternehmen OpenAI hat generative Künstliche Intelligenz (KI) die digitale Welt im Sturm erobert; KI-Anwendungen sind nun allgemein zugänglich und vielfältig einsetzbar. Allein ChatGPT erreichte innerhalb von nur zwei Monaten etwa 100 Millionen Nutzerinnen und Nutzer. Zudem finden Tools zur automatischen Erstellung von fotorealistischen Bildern und Videos wie Midjourney, Dall-E, Gemini oder jüngst Sora ebenfalls bereits in der Breite Verwendung. Die genannten Anwendungen können mittels generativer KI auf Knopfdruck Texte, Bilder oder Videos von erstaunlich hoher Qualität erstellen. Diesen Siegeszug befördert hat die unmittelbare Nutzbarkeit der entsprechenden Tools durch den freien Zugang über das Internet und einfache Interfaces; Nutzerinnen und Nutzer benötigen nahezu keine Vorkenntnisse und nur wenige technische Voraussetzungen, um in Sekundenschnelle Antworten auf Fragen aller Art zu erhalten oder Texte, Bilder und Videos zu generieren.

Grundlage und Kern generativer KI ist die Kapazität, auf Basis erlernter Muster aus vielfältigen Daten verschiedenster Herkunft und Qualität neue sprachliche oder visuelle Produkte zu erstellen. Dies geschieht auf der Basis von Korrelationen oder Wahrscheinlichkeiten, nicht aber auf Grundlage eines echten Verständnisses.

ChatGPT ist ein sogenanntes Large Language Model, das mit riesigen Textmengen trainiert wird: Webseiten, Bücher, Artikel, Songtexte, Posts, Tweets, Kommentare oder andere Meinungsäußerungen – sprich: mit sämtlichen Texterzeugnissen, die im Internet zu finden sind. Das Training besteht konkret darin, basierend auf Sprachmustern, welche aus diesen Daten gelernt wurden, für vorgegebene Satzfragmente jeweils das nächste Wort vorherzusagen. Hierfür analysiert ChatGPT zunächst den Kontext des betreffenden Satzes mithilfe statistischer Verfahren und gibt anschließend ein nach Wahrscheinlichkeitsmaßstäben berechnetes Nachfolgewort aus. Auf diese Weise kann ChatGPT Fragen Wort

für Wort statistisch plausibel beantworten und neue Texte produzieren. Eine entsprechend generierte Aussage kann korrekt sein, sie muss es aber nicht – und oft sind Antworten des Chatbots tatsächlich falsch. ChatGPT kreiert dann sinnvoll klingende Sätze, die inhaltlich jedoch frei erfunden („halluziniert") sind. Das Large Language Model versteht folglich weder die Inhalte, die es analysiert, noch jene, die es ausgibt – es hat lediglich gelernt, welche Wörter und Zeichen in bestimmten Kontexten üblicherweise aufeinanderfolgen und überträgt diese Muster auf die eigene Textproduktion.

Generative Modelle zur Erzeugung von Bildern und Videos funktionieren ähnlich: Sie werden mit riesigen Mengen an Bild- und Textmaterial vor allem aus dem Internet trainiert, um den Zusammenhang zwischen sprachlicher Beschreibung und Bild zu erlernen. Neue Bilder oder Videos erstellen entsprechende Tools dann dadurch, dass sie aus einem Startbild, das nur zufälliges Rauschen enthält, ein Bild oder Video „herausschärfen", das zur sprachlichen Anweisung passt.

All diesen Formen generativer KI ist eines gemeinsam: Sie erzeugen („generieren") mediale Produkte, die es so zuvor noch nicht gegeben hat. Insofern sind KI-generierte Inhalte etwas anderes als bloße Reproduktionen bereits existierender Texte, Bilder oder Videos. Neben der Breite ihrer Anwendbarkeit ist es vor allem diese Fähigkeit generativer KI, die immer wieder fasziniert; denn mit wenigen Klicks lassen sich originale, sprachlich einwandfrei klingende Texte und realistisch aussehende Bilder erstellen. Dass entsprechende Bilder und Texte KI-generiert sind, fällt uns als Nutzerinnen und Nutzern – wenn überhaupt – oft erst auf den zweiten Blick auf. Dadurch entstehen zum einen unterschiedliche Formen der Täuschung – sowohl über die Tatsache, dass wir mit einer KI interagieren, in Bezug auf die produzierten Materialien, aber auch hinsichtlich der Fähigkeiten, die der KI zugesprochen werden. Zum anderen ist zu berücksichtigen, dass neu generierte Texte und Bilder vollständig auf Trainingsdaten basieren, also auf der zunächst wahllosen und unkritischen Abschöpfung des online verfügbaren Text- und Bildmaterials. Das aber heißt, dass KI-Anwendungen auch alle in den Trainingsdaten zum Ausdruck kommenden Wertvorstellungen bis hin zu Vorurteilen, Verzerrungen und Limitierungen reproduzieren. Die hier nur kurz angerissenen Aspekte gilt es im Folgenden noch näher zu beleuchten.

Die vielfältigen Nutzungsmöglichkeiten und die potenziellen ökonomischen Gewinne, die der Einsatz von KI im Allgemeinen verspricht, aber auch die ethischen und regulatorischen Herausforderungen, die mit diesem Einsatz verbunden sind, wurden bereits vielfach adressiert.[1] Generative KI wirkt als Katalysator, der die Chancen, aber auch die Risiken von KI-Technologien noch verschärft – gerade aufgrund der elementaren Rolle von Sprache und Bildern für das menschliche Miteinander: Während Sprache das zentrale Medium für menschliche Kommunikation und Informationstransfer darstellt, sind Bilder und Videos von entscheidender Bedeutung für Fragen der Evidenz, aber auch für die Vermittlung und Erzeugung von Emotionen.

Ein besonders problematischer Aspekt generativer KI-Tools wie ChatGPT, der bisher nur unzureichend thematisiert worden ist, betrifft darüber hinaus deren Einbindung in zahlreiche Anwendungen, für die sie nicht explizit kuratiert worden sind. In vielen Ländern wie etwa Brasilien, China, Estland oder den USA werden beispielsweise bereits Anwendungen für die Rechtsprechung entwickelt, die die richterliche Entscheidungsfindung auf KI-Basis unterstützen oder sogar übernehmen sollen.[2] Eine durch KI beeinflusste Rechtsprechung basierte dann allerdings, wenn sie nicht sehr gezielt entwickelt wird, auf verschiedenartigen Daten, Wertvorstellungen, rechtlichen Erfahrungen und Vorurteilen, die zu Trainingszwecken aus verschiedensten Quellen, auch dem Internet, abgeschöpft wurden – und nicht auf einem der jeweiligen Rechtsordnung gemäßen Fundament an rechtlichen Werten und Interessen.

Es besteht daher die Notwendigkeit, die realen und drängenden Gefahren, die mit der Verwendung generativer KI-Systeme einhergehen, in den Blick zu nehmen und deren verantwortungsvolle Entwicklung und Nutzung sicherzustellen. Dazu gilt es zunächst, sich nicht von Scheindebatten über Singularität, das Ende der Menschheit durch KI, die Auflösung des Arbeitsmarkts, eine erforderliche Rechtspersönlichkeit oder Behauptungen über ein vorgebliches Bewusstsein von Chatbots ablenken zu lassen. Wie intuitiv KI-Systeme auch immer aufgebaut

1 Zum Beispiel High-Level Expert Group on Artificial Intelligence (2019a, 2019b, 2020a, 2020b); Datenethikkommission (2019); Deutscher Ethikrat (2023); Leopoldina, acatech, Akademienunion (2021); Orwat (2019), Spiecker gen. Döhmann & Towfigh (2023).

2 Wissenschaftliche Dienste des Deutschen Bundestages (2021).

sein mögen: Sie verfügen bislang weder über Verständnis noch über Bewusstsein – stattdessen erkennen sie Muster in großen Datenmengen und lernen, Zeichen und Muster zu rekombinieren und diese in neuen Zusammenhängen zu reproduzieren. Dabei handelt es sich, wie oben bereits erläutert, um rein statistische Anwendungen. Die Produktion von Text, Bildern oder Videos erfolgt also ohne jedes Verständnis der Inhalte oder Zusammenhänge – auch wenn dies Nutzerinnen und Nutzern so erscheinen mag.

Im Folgenden werden sieben Problemfelder skizziert, die zwar nicht ausschließlich im Zusammenhang generativer KI auftreten, die sich hier aber besonders deutlich zeigen:

1. Notwendigkeit und Angemessenheit von De-Biasing;
2. Möglichkeiten und Grenzen von erklärbarer KI;
3. Risiko der vierfachen Täuschung;
4. Macht und Machtasymmetrien;
5. KI als kritische Infrastruktur;
6. Verantwortungsdiffusion und Kontrolldefizite sowie
7. Verfügbarkeit und Offenheit von KI-Systemen.

2 Ethische und gesellschaftliche Herausforderungen durch generative KI

2.1 KI als konservatives Instrument und die Rolle von De-Biasing

Eine grundlegende Herausforderung bei der Entwicklung und Anwendung datenbasierter KI-Systeme besteht hinsichtlich ihrer Genauigkeit, Neutralität und Objektivität. Auch wenn solche Systeme durchaus die Möglichkeit eröffnen, Entscheidungsprozesse zu verbessern, so führen die stets im Datenmaterial enthaltenen Wertungen und Annahmen zu systematischen Verzerrungen. Zahllose Beispiele liefern Belege für solche Verzerrungen – sogenannte Biases – und die hieraus resultierende Diskriminierung von Personen oder gesellschaftlichen Gruppen.[3] Das ist besorgniserregend, aber nicht überraschend, denn datenbasierte KI-Systeme sind zwangsläufig konservativ: Wenn nicht aktiv gegengesteuert wird, spiegeln entsprechende Systeme die gesellschaftlichen und kulturellen Verhältnisse ihrer Datenbasis und somit auch die hier zum Ausdruck kommenden Wertvorstellungen einschließlich vorhandener Ungleichheiten und Ungerechtigkeiten wider. Fließen diese dann in KI-generierte Prognosen und Entscheidungen mit ein, werden Verhältnisse der Vergangenheit in die Zukunft hinein fortgeschrieben und schließlich zementiert – versteckt in scheinbar neutralen Algorithmen. Eine KI auf der Grundlage von Daten aus dem Jahr 1950 würde also auch im Jahr 2024 Entscheidungen nach den Maßstäben von 1950 treffen oder vorschlagen.

Wie komplex das Problem angesichts generativer KI-Technologien tatsächlich ist, zeigen die jüngsten Debatten um das KI-System Gemini, dem KI-Assistenten von Google.[4] Googles KI-Assistent hatte unter anderem die Gründerväter der USA, die in der Realität allesamt weiß und männlich waren, diversifiziert dargestellt, also mit unterschiedli-

3 Zum Beispiel Angwin et al. (2016).

4 Frankfurter Allgemeine Zeitung, 28.02.2024.

cher Hautfarbe und Geschlecht. Die Konsequenz war eine hitzige Debatte, die eine grundlegende Frage offenlegt: Sollen Texte, Bilder und Videos gesellschaftliche Realitäten mit allen Ungleichheiten und Ungerechtigkeiten reproduzieren, oder sollte hier aktiv gegengesteuert werden – beispielsweise durch sogenannte De-Biasing-Metriken? Die Frage lässt sich ganz offensichtlich nicht kategorisch klären, denn ihre Beantwortung hängt sowohl von konkreten Wertvorstellungen als auch vom Kontext ab. Geht es beispielsweise um die Darstellung historischer Begebenheiten oder um die Illustration aktueller Werbung? Normative (Vor-)Entscheidungen in der Entwicklung und beim Training generativer KI sind gleichwohl unvermeidbar; sie werden aber kaum explizit adressiert. Die Folgen sind Intransparenz und ein fehlendes Bewusstsein bei vielen Nutzerinnen und Nutzer.

Methodische Entscheidungen in der Entwicklung von KI-Modellen haben also oftmals ethische und politische Implikationen. Wichtig zu bedenken ist dabei: Sowohl der Verzicht auf De-Biasing als auch die Entscheidung für das De-Biasing, aber auch die Wahl spezifischer De-Biasing-Methoden oder Fairness-Metriken sind ethisch und politisch von enormer Bedeutung. Für solche Entscheidungen braucht es neben technisch-mathematischer also auch politisch-ethische Expertise. Daher sollten entsprechende Entscheidungen nicht ausschließlich Entwicklerinnen und Entwicklern überlassen werden. Sie erfordern insbesondere bei wirkmächtigen und tiefgreifenden KI-Systemen vielmehr einen breiten gesellschaftlichen und interdisziplinären Diskurs sowie die Partizipation der potenziell von Diskriminierung betroffenen Personengruppen.

2.2 Erklärbare KI zwischen überhöhter Hoffnung und Täuschung

Ein weiteres Problemfeld bilden die technisch inhärente mangelnde Transparenz, Nachvollziehbarkeit und Kontrolle von KI-Systemen.[5] Dieses Problem zeigt sich bei generativen KI-Modellen in besonderem Maße: Für Nutzerinnen und Nutzer ist nicht nachzuvollziehen, wie generative KI-Tools ihre Texte erzeugen, auf welche Trainingsdaten sie

5 Bordt et al. (2022); Crawford (2024).

zurückgreifen, und inwiefern die zur Verfügung gestellten Informationen tatsächlich korrekt und somit verlässlich sind. Transparenz und Nachvollziehbarkeit sind aber aus verschiedenen Gründen dringend geboten. Zum einen, um die Qualität des Outputs beurteilen und modellimmanente Annahmen bis hin zur systematischen Verzerrung von Tatsachen (Bias) oder zur Diskriminierung von Personen und Gruppen nachweisen zu können (siehe hierzu 2.1). Zum anderen, um Prognosen deuten und Entscheidungen begründen zu können, sowie ferner, um den verschiedenen Formen der Täuschung (siehe hierzu 2.3) entgegenzuwirken.

Eine scheinbare Lösung für das hier beschriebene Problem liefert die sogenannte erklärbare KI (Explainable AI). In diesem noch jungen Forschungsfeld werden Verfahren entwickelt, die KI-generierte Vorschläge oder Entscheidungen im Nachhinein begreiflich machen sollen. Betrachtet man als Beispiel ein KI-Entscheidungsverfahren zur bankinternen Beurteilung von Kreditanträgen, könnte sich der Fall wie folgt darstellen. Falls das KI-Modell einen Kreditantrag ablehnt, würde das Erklärsystem diese Entscheidung anschließend begründen: „Herr Schmidt erhält den Kredit nicht, weil er zu alt ist und sein Einkommen zu gering ausfällt." Zu bedenken ist bei entsprechenden Anwendungen, wie eine solche Erklärung (nicht) zustande kommt. Bei komplizierten KI-Entscheidungsverfahren, insbesondere im Kontext neuronaler Netze, ist es aufgrund ihrer Komplexität technisch nämlich unmöglich, den einen wahren Entscheidungsgrund anzugeben – denn einen solchen gibt es gar nicht. Stattdessen versuchen Erklärverfahren im Nachhinein plausible Gründe zu finden. Allerdings gibt es verschiedene Möglichkeiten, solche Erklärungen im Nachhinein zu erzeugen; und obwohl all diese Möglichkeiten für sich gesehen technisch plausibel sind, führen sie in der Praxis oft zu ganz unterschiedlichen Erklärungen – weshalb die tatsächlichen Gründe oft unklar bleiben.[6]

Darüber hinaus hat sich gezeigt, dass sich die gängigen Erkläralgorithmen manipulieren lassen. So ließe sich für den Fall der Kreditentscheidung beispielhaft ein diskriminierungsbasiertes Entscheidungssystem mit einem Erklärsystem kombinieren, das stets allgemein nachvollziehbare, unanfechtbare Begründungen generiert – die über

6 Bordt et al. (2022).

die tatsächlich entscheidungsrelevanten Eigenschaften der antragstellenden Person allerdings schweigen.[7] In einem solchen Fall wäre die absurde Konsequenz, dass die Erklärung nicht der Transparenz diente, sondern im Gegenteil ein probates Mittel zur Täuschung wäre. Besonders kritisch ist in diesem Zusammenhang, dass sich selbst durch umfangreiche technische Prüfung von außen nicht aufdecken lässt, ob eine Erklärung tatsächlich sinnvoll ist oder ob sie zielgerichtet manipuliert wurde.[8] Angesichts der rechtlichen Vorgaben, die Transparenz und Erklärbarkeit von KI-Vorgängen fordern, ist das ein ernst zu nehmendes Problem;[9] denn im juristischen Kontext ergeben Erklärungen nur dann Sinn, wenn sie auch überprüfbar sind – was technologisch aber eben bislang nicht zu leisten ist.

Wie schwierig es ist, die Entscheidungsfindung von KI-Systemen transparent zu machen, zeigt sich noch deutlicher beim Versuch, die Ergebnisse generativer KI-Modelle zu erklären. Große Sprachmodelle entscheiden basierend auf dem jeweiligen Kontext, wie der konkrete Text zu gestalten ist. Um zu erklären, warum ein Sprachmodell einen bestimmten Text generiert, müsste man also zunächst aufzeigen, auf welchen Kontext sich das Modell im konkreten Fall bezieht; dieser hängt unter anderem vom aktuellen Prompt und dem bereits generierten Text ab. Aktuell gibt es aber keinen technologischen Ansatz, der die konkrete Kontextbeziehung eines KI-generierten Textes von außen nachzeichnen könnte. Und es ist schwer vorstellbar, dass sich ein solcher Ansatz überhaupt finden lässt. Um das Problem zu umgehen, arbeiten Wissenschaftlerinnen und Wissenschaftler gegenwärtig daran, dass große Sprachmodelle ihre Vorgehensweisen und Ergebnisse selbst erklären. Die charmante Idee dahinter ist: Das Sprachmodell wisse selbst am besten, warum es einen bestimmten Text generiert hat. Entsprechende Erklärungen hören sich auch oft schlüssig an – sie sind aber ebenfalls nicht verlässlich.[10] Denn wie oben bereits dargelegt, besitzt ein Sprachmodell weder ein semantisches Verständnis von Texten noch ein Verständnis der eigenen Funktionsweise. Es produziert lediglich statistisch plausible

7 Sharma et al. (2024).

8 Bordt et al. (2022).

9 Europäische Datenschutz-Grundverordnung (kurz: EU-DSGVO); Europäische KI-Verordnung (kurz: EU-KI-VO).

10 Tanneru et al. (2024); Turpin et al. (2023).

Texte. Die hier enthaltenen Aussagen können wahr oder erfunden sein; und es ist technologisch nicht möglich, zwischen diesen beiden Kategorien zuverlässig zu unterscheiden. Um es mit den Worten des US-amerikanischen Philosophen Harry Frankfurt[11] zu sagen, produzieren solche Systeme „Bullshit" – sie lassen die Grenze zwischen Wahrheit und Lüge hinter plausibel klingenden Texten verschwinden.[12] Auf diese Weise tragen erklärbare KI-Systeme im schlimmsten Fall also noch zur Täuschung bei, anstatt diese zu vermeiden. Juristisch gesehen bieten entsprechende Ansätze somit kein zuverlässiges Beurteilungsinstrument.

Was folgt nun aus diesen Einsichten? Verlässliche Transparenz kann durch Erklärungsverfahren weder bei KI-Systemen im Allgemeinen noch bei generativer KI im Besonderen im Nachhinein hergestellt werden. Es gilt also zu entscheiden, bei welchen Anwendungen die unauflösbare Intransparenz generativer KI mit Blick auf Nutzen und Risiken hinzunehmen ist. Im Fall von Sprachmodellen betrifft das einfache Anwendungen wie die Ausführung von Routineaufgaben, bei denen Fehler entweder nicht kritisch oder aber schnell zu entdecken und leicht zu korrigieren sind, wie zum Beispiel das Ausformulieren einer E-Mail aufgrund von Stichworten oder das Erstellen eines Bewerbungsschreibens auf der Basis von Lebenslauf und Ausschreibung. In welchen Zusammenhängen ist Transparenz hingegen unverzichtbar? Hier sind unter anderem Anwendungen im juristischen Kontext zu nennen – beispielsweise ein KI-System, das umfangreiche Straf- oder Zivilprozessakten zusammenfasst oder auf Basis früherer Gerichtsentscheidungen der Richterschaft Entscheidungsvorschläge unterbreitet. In diesem Fall sind Systeme der generativen KI mit höchster Vorsicht zu genießen, denn weder die KI-generierten Resultate selbst noch die KI-generierte Erklärung ihres Zustandekommens stellten angesichts der vorangehenden Erläuterungen tatsächlich eine verlässliche Hilfe dar. Zwar scheint die Idee der menschlichen Aufsicht in diesem Zusammenhang zumindest grundsätzlich ein mögliches Korrektiv zu sein; aber wer sollte im juristischen Alltag die KI-Zusammenfassung einer umfangreichen Strafprozessakte überprüfen? Und wenn eine solche Prüfung ohnehin dem Menschen überlassen wäre, wäre der KI-basierte Effizienzgewinn dann nicht schon wieder verloren?

11 Frankfurt (2005).

12 Hicks et al. (2024).

Erklärbare KI ist somit ein interessantes und wichtiges Forschungsfeld, das gleichwohl nicht mit Erwartungen überfrachtet werden sollte. Bei komplexen Entscheidungs- oder Vorhersagemodellen helfen Erklärungen *ex post* nicht, weil diese nicht die tatsächlichen Prozesse nachvollziehen, sondern lediglich plausible Annäherungen liefern. In spezifischen Situationen und Zusammenhängen, in denen uns die Erklärbarkeit von Ergebnissen und ihrem Zustandekommen unverzichtbar ist, müssen daher Methoden eingesetzt werden, deren Vorhersagemodelle weniger mathematisch komplex, dafür aber direkt von Menschen interpretierbar sind – auch wenn diese möglicherweise zu weniger geeigneten Ergebnissen führen. Ein Beispiel für solche einfachen Vorhersagemodelle sind Entscheidungsbäume.

2.3 Das Problem der vierfachen Täuschung

Wie oben bereits dargelegt, können generative KI-Modelle mittlerweile Texte, Bilder und Videos in sehr hoher Qualität produzieren, womit die Gefahr einer Täuschung von Nutzerinnen und Nutzern einhergeht. Konkret lassen sich in diesem Zusammenhang mindestens vier verschiedene Dimensionen der Täuschung identifizieren, was ein zentrales Problem bei der Anwendung generativer KI darstellt.[13]

Zunächst betrifft das die Interaktion zwischen Nutzerin oder Nutzer und Chatbot, sofern unklar bleibt, dass es sich bei Letzterem nicht um einen Menschen, sondern um ein KI-System handelt. Die Frage, ob man mit einem Chatbot oder mit einem menschlichen Gegenüber interagiert, ist bereits in der klassischen Kundenberatung und bei der Moderation in sozialen Medien relevant – sehr viel mehr jedoch gilt das für besonders sensible Zusammenhänge, beispielsweise in der Psychotherapie.

13 Messeri & Crockett (2024); Deutscher Bundestag, Ausschuss für Bildung, Forschung und Technikfolgenabschätzung, Ausschussdrs. 20(18)108b, 21. April 2023. Expertengespräch zum Thema ChatGPT, Prof. Dr. Judith Simon, Universität Hamburg; https://www.bundestag.de/resource/blob/944448/004ca2f7a9fcf586a07113c6ba72b689/20-18-108b-Simon-data.pdf [letzter Zugriff: 20.08.2024].

Täuschungspotenzial besteht ferner hinsichtlich der Fähigkeiten generativer KI-Modelle. Denn auch wenn gegenwärtig verfügbare KI-Systeme weder ein semantisches Verständnis noch Bewusstsein besitzen, kann dies Nutzerinnen und Nutzern so erscheinen – und das selbst dann, wenn diese wissen, dass sie mit einem technischen System interagieren. Das zeigen schon frühe Erfahrungen mit der Software ELIZA,[14] die der deutsch-US-amerikanische Informatiker Joseph Weizenbaum in den 1960er-Jahren entwickelt hat, ebenso wie aktuelle Berichte zur Interaktion von Nutzerinnen und Nutzern mit ChatGPT: Menschen neigen ganz offensichtlich dazu, einem Chatbot Verständnis, Empathie oder Bewusstsein zuzusprechen, wenn dieser plausibel kommuniziert. Eine solche Zuschreibung menschlicher Fähigkeiten sagt zwar nichts über die tatsächlichen Funktionsweisen und Fähigkeiten der Maschine aus – wohl aber über die menschliche Tendenz zur Anthropomorphisierung von Technik.

Die dritte Dimension des Täuschungsproblems betrifft die Ebene der Resultate generativer KI-Systeme, die mit mannigfaltigen Gefahren einhergehen können: von Fake News und Deep Fakes zu Propaganda-, Verleumdungs- und Mobbingzwecken bis zur kriminellen Nutzung gefälschter Stimmen, um Angehörige zu betrügen oder in Gerichtsprozessen Beweiskraft vorzugaukeln. Täuschung und Manipulation sind gewiss keine neuen Phänomene – Qualität, Einfachheit, die weitgehende technologische und technische Voraussetzungslosigkeit sowie die Geschwindigkeit, mit der Texte, Bilder oder Videos heute hergestellt und weiterverbreitet werden können, eröffnen jedoch eine völlig neue Dimension möglichen Missbrauchs. Bei allen Chancen, die sich durch generative KI-Modelle ergeben, bergen ChatGPT und andere Systeme somit eine reale Gefahr für unsere demokratische Grundordnung, da grundlegende Prozesse der Information und Kommunikation schnell, einfach und nachhaltig gestört werden können und Beweis- und Glaubwürdigkeit keine verlässlichen Kategorien mehr darstellen. Diesen Umstand machen aktuelle Beispiele aus Wahlkämpfen in der Slowakei, den USA und der Europäischen Union deutlich.[15]

14 Weizenbaum (1966, 2023).

15 Zum Beispiel Wired, 03.10.2023

Täuschen können sich Nutzerinnen und Nutzer schließlich auch über die Funktionsweise generativer KI, sobald diese in andere Systeme integriert ist. Besonders ausgeprägt zeigt sich das Problem dort, wo die Bereitstellung existierender Informationen mit der Generierung neuer Informationen vermischt wird. Denn auch wenn online verfügbare Informationen einen unterschiedlichen Wahrheitswert aufweisen, also sowohl wahr als auch falsch sein können, so besteht doch ein gewaltiger Unterschied darin, ob ein Klick den Link zu bereits vorhandener Information öffnet oder ob er neue Informationen generiert.

Der Prozess der Zusammenstellung und Wiedergabe von existenten Informationen und der Prozess der Produktion neuer Information, die bisher getrennt waren, werden vermischt. Mittlerweile wird generative KI in immer mehr Anwendungen, Softwaresysteme und Tools eingebaut – vom PDF-Reader über E-Mail-Programme und Internet-Suchmaschinen bis hin zu ganzen Softwarepaketen wie dem KI-Assistenten Microsoft Copilot. Für die Nutzerinnen und Nutzer wird es daher zunehmend schwer, zwischen existierenden und neu generierten Inhalten zu unterscheiden, was die Einordnung und Bewertung der Qualität und Herkunft von Informationen zusätzlich erschwert.

2.4 Machtfragen und Machtasymmetrien

Eine weitere Herausforderung für die Entwicklung und Anwendung generativer KI besteht angesichts möglicher Machtasymmetrien, insbesondere bei Verwendung und Auswertung personenbezogener Daten. Aus der Verarbeitung sowohl ursprünglicher, frei verfügbarer Trainingsdaten als auch gezielt eingegebener Nutzerdaten ergeben sich vielfältige Gefahren für die Privatsphäre, Autonomie und Selbstständigkeit von Nutzerinnen und Nutzern. Gerade im Kontext von Scoring oder personalisierten Angeboten nutzen Unternehmen und Staat schon jetzt vielfältige Daten, um Nutzerinnen und Nutzern spezifisch zugeschnittene Angebote zu machen. Auch wenn dies hilfreich sein kann, so können sie ihr überlegenes Wissen über Einzelpersonen oder Gruppen auch gegen jene einsetzen. Dies geschieht beispielsweise im Kontext sogenannter Dark Patterns, mittels derer Entscheidungen unterschwellig beeinflusst werden, oder zur einseitigen Ausgestaltung von Vertragsbedingungen oder Zugang zu staatlichen Leistungen, weil Präferenzen und Zwänge

der Nutzerinnen und Nutzer ausgenutzt werden. Derartige Missbrauchsmöglichkeiten nehmen unter den Bedingungen von KI enorm zu. Die Ergebnisse entsprechender Datenanalysen sind keineswegs immer zutreffend, geschweige denn normativ wünschenswert. Machtasymmetrien zwischen denen, die mithilfe von KI-Systemen beurteilen, und jenen, die von KI-Systemen beurteilt werden, verstärken sich weiter. Weil Nutzerinnen und Nutzer von KI-Modellen nicht wissen, was aus ihrem Verhalten gefolgert wird, laufen Selbstschutzstrategien meist ins Leere. Dies betrifft alle Nutzerinnen und Nutzer von KI-Systemen, besonders aber gesellschaftlich marginalisierte Individuen und Gruppen.

Darüber hinaus stellen sich im Kontext der Entwicklung und Anwendung generativer KI-Modelle verschärft Fragen des geistigen Eigentums und des Urheberrechts, da die Modelle (vielfach) auf Basis von Daten trainiert werden, deren Urheberinnen oder Urheber weder über die Verwendung dieser Daten informiert wurden noch entsprechende Kompensationsleistungen erhalten haben – geschweige denn am KI-generierten Mehrwert beteiligt werden. Wie erste Gerichtsverfahren über die Verletzung von Urheberrechten im KI-Kontext zeigen, geht es aber auch ganz generell um die Frage des Zugriffsrechts bei Daten, die insbesondere im Internet technisch leicht verfügbar sind und zur Entwicklung von KI einseitig genutzt werden.

Ein überzeugendes Konzept, wie diejenigen, auf deren Daten der Erfolg einer KI-Anwendung beruht, individuell und gesamtgesellschaftlich beteiligt werden können, wodurch Chancen und Risiken fair geteilt würden, fehlt allerdings bislang. Gleiches gilt für die Frage, wie die Entscheidungshoheit von Urheberinnen und Urhebern über die Weitergabe und Verwendung ihrer Daten durch generative KI-Systeme zu gewährleisten ist.

2.5 KI als kritische Infrastruktur

Eine Besonderheit generativer KI-Modelle besteht wie oben bereits dargelegt hinsichtlich ihrer Einsatzvielfalt: Insbesondere große Sprachmodelle werden nicht nur direkt über das Webinterface von Nutzerinnen und Nutzern verwendet, sondern in zunehmendem Maße auch in eine Vielzahl von Produkten, Prozessen und Dienstleistungen integriert.

Dies geschieht über die Einbindung sogenannter Foundation Models.[16] KI-Systeme im Allgemeinen sowie generative KI- und Sprachmodelle im Besonderen bilden so in zweierlei Hinsicht eine Form kritischer Infrastruktur: zum einen angesichts ihrer Bedeutung für vielfache andere Einsätze, die KI-Systeme unumgehbar machen und eine große Abhängigkeit kreieren, zum anderen aber auch angesichts ihrer Unsichtbarkeit.

Hieraus ergeben sich eine Reihe von Risiken: Der Einbau generativer KI in verschiedene Softwaresysteme führt zunächst zum bereits oben formulierten Problem der vierten Täuschung (siehe hierzu 2.3) – also zur Verwischung der Grenze zwischen zuvor bereits existenten und neu generierten Inhalten. Darüber hinaus ergeben beziehungsweise verstärken sich Probleme, die unter dem Begriff der algorithmischen Monokultur[17] bereits für andere algorithmische Systeme diskutiert wurden; sowohl für die Entwicklung als auch für den Einsatz von KI-Anwendungen gibt es zahlreiche Belege, dass komplette Foundation Models oder einzelne Komponenten, insbesondere Trainingsdaten, aber auch Software-Packages geteilt werden. Als Folge dieser gemeinsamen Nutzung kann es zu einer Homogenisierung des Outputs dieser Modelle kommen. Das bedeutet, dass verschiedene Anwendungssysteme – beispielsweise Text- oder Bildgeneratoren – möglicherweise sehr ähnliche Ergebnisse erzeugen, weil sie auf Basis derselben Daten und KI-Modelle trainiert wurden. In der Folge können sich Fehler, aber auch Biases und möglicherweise daraus resultierende Diskriminierung in den Ergebnissen dann systematisch auswirken,[18] was allerdings unbemerkt bleibt, da die Foundation Models in den verschiedenen Anwendungen unsichtbar bleiben. Vor diesem Hintergrund wird deutlich, warum insbesondere für Foundation Models hohe Anforderungen an Transparenz und Qualität gelten müssen.

16 Dabei handelt es sich um Basismodelle, die mittels Maschinellen Lernens mit einer großen Datenmenge trainiert werden und für unterschiedliche nachgelagerte Anwendungen genutzt werden können. Dazu gehören auch Große Sprachmodelle (Large Language Models).

17 Kleinberg & Raghavan (2021).

18 Bommasani et al. (2022a, 2022b).

2.6 Verantwortungsdiffusion und Kontrolldefizit

Die vielen verschiedenen Akteure und Einflüsse im Lebenszyklus generativer KI-Systeme, der von der Entwicklung über die Integration in andere Anwendungssysteme bis zur Nutzung entsprechender Modelle reicht, führen zu einer enormen Verantwortungsdiffusion in diesem Bereich. Welche Einflüsse das konkret sind, welche Akteure auf Arbeitsweise und Resultate generativer KI-Modelle in welcher Weise einwirken, ist nach außen hin aber zumeist nicht erkennbar – und soll es bisweilen auch gar nicht sein, wie das Phänomen der sogenannten Deep Fakes deutlich macht. So bleibt allerdings unklar, wer die Verantwortung für die positiven wie negativen Effekte dieser Technologie trägt und wem die politischen, rechtlichen, sozialen und wirtschaftlichen Folgen wirksam zugeschrieben werden können – eine Gefahr für Demokratie und Rechtsstaat.

Dies gilt im Fall der generativen KI umso mehr, als eben nicht nur die handelnden Personen und Institutionen bisweilen unbekannt bleiben, sondern auch die technologischen Hintergründe und die Funktionsweise unzugänglich sind. Ob und von wem ein Bild oder Text erstellt beziehungsweise verändert wurde und mit welcher Zielrichtung dies geschehen ist, lässt sich in der Regel nicht mehr ermitteln. Kontrolle von Entscheidungen, wie sie durch generative KI angeleitet und zum Teil übernommen werden, braucht aber einen Zugriff auf relevante Parameter wie Inhalt, Zwecksetzung, zugrunde liegende Wertvorstellungen, Selektionskriterien, Autorschaft und Entscheidungsvorgaben, egal, ob sie von Privatpersonen, Presse, staatlichen Aufsichtsbehörden, Gerichten oder anderen durchgeführt wird. In der Folge müssen Nutzerinnen und Nutzer dem jeweilig verwendeten KI-Modell und dem entsprechenden Anwendungssystem vertrauen, wenn sie diese nutzen wollen – ohne dass ein solches Vertrauen tatsächlich gerechtfertigt wäre; rechtliche und ethische Regeln laufen somit ins Leere.

Verstärkt wird dieses Verantwortungs- und Kontrolldefizit durch drei weitere Umstände: Erstens kann KI Menschen und Institutionen bei Entscheidungen in unterschiedlicher Weise unterstützen, was die Zuschreibung von Verantwortlichkeit erschweren kann. So kann sie als Hilfsinstrument herangezogen werden, wobei die eigentliche Entscheidung der Mensch trifft, z. B. bei der Erstellung eines Anschreibens.

Stärkeres Gewicht erhält die KI, wenn die Entscheidung grundsätzlich KI-basiert erfolgt und der Mensch lediglich im Ausnahmefall eingreifen und die KI-Entscheidung verändern kann. Am gravierendsten ist der Kontrollverlust allerdings dann, wenn der Mensch gar keine Möglichkeit hat, die KI-basierte Entscheidung zu ändern, etwa bei Einbau generativer KI in andere Anwendungen, Softwaresysteme und Tools. Letzteres wird, gerade beim Einsatz von KI in komplexen autonomen Systemen, als erforderlich für deren Funktionalität angesehen.

Der zweite Umstand ergibt sich daraus, dass aus nachfolgenden Entscheidungen grundsätzlich nicht zu schließen ist, welche Informationen auf welche Weise in diese eingeflossen sind. Somit aber ist nicht nachzuvollziehen, wie eine KI-Anwendung – sofern überhaupt erkennbar ist, dass eine solche eingesetzt wurde – Informationen ermittelt, bewertet, sortiert und verarbeitet hat und nach welchen normativen Kriterien dies erfolgt ist. Insofern ist allerdings auch nicht zu beurteilen, ob das KI-generierte Ergebnis nach Maßgabe korrekt zustande gekommen ist. Die fortwährende Anpassung von KI im großen Maßstab, die zweifellos ihre besondere technologische Stärke ist, lässt technische Lösungen zur begleitenden Prüfung oder Nachkontrolle kaum vorstellbar erscheinen.

Verschärft wird das Kontrolldefizit drittens schließlich durch die technisch bedingte Unmöglichkeit, Entscheidungen einer KI-Anwendung zu reproduzieren, um diese der Kontrolle zuzuführen. Das bereits oben geschilderte Problem der Herstellung von Transparenz ist damit eng verbunden.

Die weitreichende Integration generativer KI in zahlreiche Produkte, Anwendungen, Dienstleistungen und Tools wirkt wegen der Vielzahl der Beteiligten einer klaren Verantwortungszuschreibung, der Einhaltung rechtlicher Regeln und ethischer Standards sowie einer effektiven Kontrolle durch Nutzerinnen und Nutzer oder staatliche Organe somit entgegen.

2.7 Verfügbarkeit und Offenheit von KI

Die Notwendigkeit sowie die Grenzen von Transparenz und Nachvollziehbarkeit ziehen Fragen in Bezug auf die spezifischen Modalitäten der Verfügbarkeit und Offenheit generativer KI-Systeme nach sich.

Offene Systeme, insbesondere Open-Source-Lösungen, bieten verschiedene Vorteile: Zunächst betrifft das die bessere Überprüfbarkeit. Während die Arbeitsweise proprietärer Systeme in der Regel im wirtschaftlichen Interesse der Eigentümerinnen und Eigentümer beziehungsweise shareholderorientiert konzipiert sein wird, womit Intransparenz sowie politische und gesamtgesellschaftliche Risiken verbunden sind, bleiben offene Systeme transparenter; sie ermöglichen eine externe Kontrolle und die Prüfung der verwendeten Daten, Methoden und Modelle. Gleichwohl ist vollständige Transparenz auch im Open-Source-Segment nicht immer herzustellen, da manche Trainingsdaten aus Datenschutzgründen nicht veröffentlicht werden dürfen.

Darüber hinaus versprechen Open-Source-Lösungen gemeinhin verlässliche allgemeine Zugänglichkeit, das heißt, dass die entwickelten Systeme auch in Zukunft uneingeschränkt zur Verfügung stehen werden. So wäre beispielsweise auszuschließen, dass die öffentliche Verwaltung ein sprachmodellbasiertes Anwendungssystem entwickeln und die Nutzung des zugrunde liegenden Sprachmodells zu einem bestimmten Zeitpunkt eingeschränkt oder kostenintensiv werden würde. Das Beispiel des KI-Tools AlphaFold 3 zeigt, dass genau das im Segment proprietärer Systeme auch tatsächlich passieren kann.[19] Zudem ermöglichen Open-Source-Systeme auch die Entwicklung von Anwendungen bei geringer Nachfrage, also dort, wo kein (oder geringes) wirtschaftliches Herstellerinteresse besteht; das gilt zum Beispiel bei Sprachmodellen für Sprachen, die nur von wenigen Menschen gesprochen werden.

Nichtsdestotrotz haben Open-Source-Systeme im Bereich der generativen KI auch viele Nachteile, denn auch eine missbräuchliche Verwendung der Technologie wird durch die Öffentlichkeit des Quellcodes stark erleichtert. Jede Person, Gruppe oder Institution kann diesen Quellcode nutzen und für ihre Zwecke verändern, etwa um Fake News oder Hasskommentare in großer Zahl zu generieren. Zudem hängen die

19 AlphaFold ist ein KI-Tool zur Erforschung von Proteinen, das von der Firma DeepMind, einer Google-Tochter, entwickelt wurde und weltweit in der Forschung eingesetzt wird; in der neuesten Version von Mai 2024 hat DeepMind den kostenlosen Zugang der öffentlichen Forschung plötzlich stark eingeschränkt und stellt den Source Code des Tools nicht mehr frei zur Verfügung, sodass nicht überprüft werden kann, was genau im Hintergrund passiert.

langfristige Verfügbarkeit und Sicherheit natürlich davon ab, ob die jeweilige Open-Source-Lösung nachhaltig betrieben wird.

Die Vor- und Nachteile offener KI-Modelle öffentlich zu diskutieren und angemessen abzuwägen, ist von entscheidender Bedeutung, um künftig ausbalancierte, praktikable und demokratieverträgliche Technologielösungen bereitzustellen. Die gegenwärtige Konstellation eines einfachen, gleichwohl nur scheinbar freien Zugangs zu opaken, proprietären Systemen dürfte die schlechteste Möglichkeit der Verwendung von generativer KI-Technologie darstellen; denn so werden frei verfügbare Wissens- und Nutzerdatenbestände ohne Einwilligung abgeschöpft und anschließend nach intransparenten, privatwirtschaftlich motivierten Regeln und Vorgaben zur Verfügung gestellt, ohne dass dieser Prozess gesellschaftsverträglich, ökonomisch ausgewogen und rechtlich abgesichert gesteuert oder begleitet werden könnte.

3 Fazit

Ziel des vorliegenden Diskussionspapiers ist es, einen realistischen Blick auf Chancen und Risiken in der Entwicklung und Anwendung generativer KI zu werfen – und sich damit sowohl den utopischen Heilsversprechen als auch den dystopischen Warnungen, die die gegenwärtige Debatte prägen, entgegenzustellen. Der Fokus der hier formulierten Überlegungen richtet sich auf die in der öffentlichen Diskussion noch nicht ausreichend reflektierten, gleichwohl schwerwiegenden Gefahren für Individuen, Demokratie, Wirtschaft und Gesellschaft. Diese Gefahren sind zum Teil in der Funktionslogik der Technologie selbst angelegt – etwa die Nichterklärbarkeit, die Nichtkontrollierbarkeit, die Nichtneutralität und die Nichtobjektivität generativer KI-Technologie; zu anderen Teilen entstehen sie erst in der konkreten Anwendung oder aufgrund des Zusammenspiels zwischen Mensch und Technik in unterschiedlichen Kontexten und organisationalen Rahmenbedingungen – etwa die Verantwortungsdiffusion, die Manipulierbarkeit oder die Täuschung über die Leistungsfähigkeit. Viele dieser Gefahren werden durch den European Artificial Intelligence Act und andere Gesetzeswerke[20] nicht oder nicht ausreichend adressiert, sind also gegenwärtig nicht Bestandteil normativer Leitplanken generativer KI-Entwicklung in Deutschland und Europa.

Dabei tut eine solche Regulierung gerade not. Denn die inhärente Wertung bei der Datensammlung, -aufbereitung und -sortierung und bei der Ausgestaltung von Entscheidungen in generativen KI-Systemen führt dazu, dass diese Systeme eben keine objektiven, neutralen und dem Menschen grundsätzlich überlegenen Entscheidungen produzieren würden. Vielmehr ist jede generative KI stets das Abbild der ihr zugrunde liegenden Trainingsdaten sowie der Ziele und Zwecke ihrer Entwicklung, die in ihre Funktionslogik untrennbar eingewebt sind.

20 Zum Beispiel General Data Protection Regulation, European Data Act, das Urheber- oder das Verfahrensrecht für die öffentliche Verwaltung.

Diese aber sind weder der KI selbst noch den entsprechenden Anwendungssystemen anzusehen und entziehen sich somit der Kontrolle und Regulierung durch herkömmliche Verfahren, Institutionen und Normen.

Wer generative KI verwendet, rezipiert Informationen auf Basis spezifischer Wertvorstellungen anderer, ohne diese in der Regel bewusst zu reflektieren. Diese Tatsache öffnet den Raum für Täuschung und Manipulation unter dem Deckmantel einer scheinbar überlegenen und vermeintlich neutralen, objektiven Technologie. Damit aber werden die Grundfesten unserer Weltwahrnehmung und unserer sinneserfahrungsbasierten Urteilskraft nachhaltig erschüttert. Text und Bild verlieren schließlich ihre Beweiskraft.

Vor dem Hintergrund laufender Diskussionen um die Entwicklung vertrauenswürdiger KI in Europa und eines damit verbundenen Wettbewerbsvorteils ist es geboten, Maßnahmen zur Schadensvermeidung zu forcieren. Dies betrifft beispielsweise Methoden, die Verzerrungen aufdecken oder minimieren (De-Biasing) und die Transparenz der Technologie erhöhen (Explainable AI).[21] Gleichwohl gilt es vor überhöhten Erwartungen an die Entwicklung entsprechender Technologien und Tools zu warnen, weil auch solche Entwicklungsmethoden an Grenzen stoßen oder sogar Risiken bergen, wie beispielsweise die Ausnutzung offener Quellcodes für die großflächige Verbreitung von Deep Fakes im Fall von Open-Source-Lösungen zeigt.

Die in diesem Diskussionspapier behandelten Aspekte sind selbstverständlich nicht abschließend zu verstehen. Kritisch zu sehen sind so beispielsweise die oftmals prekären Arbeitsbedingungen – insbesondere in Ländern des Globalen Südens – bei der Entwicklung und Anwendung zahlreicher KI-Modelle sowie der hohe Energie- und Ressourcenverbrauch beim Training der Modelle, beim Abarbeiten der vielen Eingabebefehle und bei der Nutzung der verschiedenen KI-basierten Tools.[22] Für all diese Fälle gibt es keine einfachen, allgemeingültigen Antworten oder Lösungen. Dennoch und gerade deswegen ist es notwendig, diese Aspekte in all ihrer Ambivalenz offenzulegen, um sie einer kritischen öffentlichen Debatte zugänglich zu machen.

21 Asghari et al. (2022).

22 Crawford (2024).

Literatur

Asghari, H., Birner, N., Burchardt, A., Dicks, D., Faßbender, J., Feldhus, N., Hewett, F., Hofmann, V., Kettemann, M. C., Schulz, W., Simon, J., Stolberg-Larsen, J., Züger, T. (2022). *What to explain when explaining is difficult? An interdisciplinary primer on XAI and meaningful information in automated decision-making.* Berlin: Alexander von Humboldt Institute for Internet and Society. https://doi.org/10.5281/zenodo.6375784

Angwin, J., Larson, J., Mattu, S., Kirchner, L. (2016). Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*, 23.05.2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [letzter Zugriff: 20.08.2024].

Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., Liang, P. (2022a). Picking on the same person. Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems*, 35, 3663–3678. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/17a234c91f746d9625a75cf8a8731ee2-Paper-Conference.pdf [letzter Zugriff: 31.07.2024].

Bommasani, R., Hudson, D. A. ..., Liang, P. (2022b). On the opportunities and risks of foundation models. *arXiv preprint*, 12.07.2022, arXiv:2108.07258v3. https://doi.org/10.48550/arXiv.2108.07258

Bordt, S., Finck, M., Raidl, E., von Luxburg, U. (2022). Post-hoc explanations fail to achieve their purpose in adversarial contexts. *FAccT '22. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 891–905. https://doi.org/10.1145/3531146.3533153

Crawford, K. (2024). Generative AI's environmental costs are soaring. And mostly secret. *Nature*, 626, 693. https://doi.org/10.1038/d41586-024-00478-x

Datenethikkommission der Bundesregierung (2019). *Gutachten der Datenethikkommission*. Berlin: Bundesministerium des Innern, für Bau und Heimat, Bundesministerium der Justiz und für Verbraucherschutz. URL: https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf;jsessionid=DEDC3EDADAB4D9F958616C80372741D2.live861?__blob=publicationFile&v=7 [letzter Zugriff: 10.07.2024].

Deutscher Ethikrat (2023). *Mensch und Maschine. Herausforderungen durch Künstliche Intelligenz* (Stellungnahme). Berlin: Deutscher Ethikrat. URL: https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf [letzter Zugriff: 12.05.2024].

Frankfurt, H. G. (2005). On bullshit. Princeton: Princeton University Press.

Frankfurter Allgemeine Zeitung, 28.02.2024. *Wir fragen eine Ethikerin. Welche Gesellschaft soll Gemini abbilden?* Von Hendrik Wieduwilt. URL: https://www.faz.net/pro/d-economy/kuenstliche-intelligenz/wir-fragen-eine-ethikerin-welche-gesellschaft-soll-gemini-abbilden-19550166.html [letzter Zugriff: 01.08.2024]

Hicks, M. T., Humphries, J., Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, 26, 38. https://doi.org/10.1007/s10676-024-09775-5

High-Level Expert Group on Artificial Intelligence (2019a). *Ethics guidelines for trustworthy AI.* Brussels: European Commission. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [letzter Zugriff: 20.08.2024].

High-Level Expert Group on Artificial Intelligence (2019b). *Policy and investment recommendations for trustworthy AI*. Brussels: European Commission. URL: https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence [letzter Zugriff: 20.08.2024].

High-Level Expert Group on Artificial Intelligence (2020a). *Assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment*. Brussels: European Commission. URL: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment [letzter Zugriff: 20.08.2024].

High-Level Expert Group on Artificial Intelligence (2020b). *Sectoral considerations on the policy and investment recommendations for trustworthy AI*. Brussels: European Commission. URL: https://futurium.ec.europa.eu/en/european-ai-alliance/document/ai-hleg-sectoral-considerations-policy-and-investment-recommendations-trustworthy-ai [letzter Zugriff: 20.08.2024].

Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22), e2018340118. https://doi.org/10.1073/pnas.2018340118

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627, 49–58. https://doi.org/10.1038/s41586-024-07146-0

Nationale Akademie der Wissenschaften Leopoldina, acatech – Deutsche Akademie der Technikwissenschaften, Union der deutschen Akademien der Wissenschaften (2021). *Digitalisierung und Demokratie.* Halle (Saale). https://doi.org/10.26164/leopoldina_03_00348

Orwat, C. (2019). *Diskriminierungsrisiken durch Verwendung von Algorithmen.* Baden-Baden: Nomos. URL: https://www.antidiskriminierungsstelle.de/SharedDocs/downloads/DE/publikationen/Expertisen/studie_diskriminierungsrisiken_durch_verwendung_von_algorithmen.pdf?__blob=publicationFile&v=3 [letzter Zugriff: 01.03.2024].

Sharma, R., Redyuk, S., Mukherjee, S., Sipka, A., Vollmer, S., Selby, D. (2024). X Hacking. The threat of misguided AutoML. *arXiv preprint*, 12.02.2024, arXiv:2401.08513v2. https://doi.org/10.48550/arXiv.2401.08513

Spiecker gen. Döhmann, I., & Towfigh, E. V. (2023). *Das Allgemeine Gleichbehandlungsgesetz und der Schutz vor Diskriminierung durch algorithmische Entscheidungssysteme. Bestandsaufnahme und Herausforderungen.* https://www.antidiskriminierungsstelle.de/SharedDocs/downloads/DE/publikationen/Rechtsgutachten/schutz_vor_diskriminierung_durch_KI.html?nn=305458 [letzter Zugriff: 02.10.2024].

Tanneru, S. H., Agarwal, C., Lakkaraju, H. (2024). Quantifying uncertainty in natural language explanations of large language models. Proceedings of the 27th International Conference on Artificial Intelligence and Statistics. *Proceedings of Machine Learning Research*, 238, 1072–1080. URL: https://proceedings.mlr.press/v238/harsha-tanneru24a/harsha-tanneru24a.pdf [letzter Zugriff: 20.08.2024].

Turpin, M., Michael, J., Perez, E., Bowman, S. R. (2023). Language models don't always say what they think. Unfaithful explanations in chain-of-thought prompting. *NIPS '23. Proceedings of the 37th International Conference on Neural Information Processing Systems*, 3275, 74952–74965. URL: https://dl.acm.org/doi/10.5555/3666122.3669397. [letzter Zugriff: 31.07.2024].

Weizenbaum, J. (1966). ELIZA. A computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery*, 9, 36–45. https://doi.org/10.1145/365153.365168.

Weizenbaum, J. (2023). *Die Macht der Computer und die Ohnmacht der Vernunft* (16. Auflage). Frankfurt am Main: Suhrkamp.

Wired, 03.10.2023. *Slovakia's election deepfakes show AI is a danger to democracy. Fact-checkers scrambled to deal with faked audio recordings released days before a tight election, in a warning for other countries with looming votes.* Von Morgan Meaker. URL: https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/. [letzter Zugriff: 21.08.2024].

Wissenschaftliche Dienste des Deutschen Bundestags (2021). *Künstliche Intelligenz in der Justiz: Internationaler Überblick.* WD 7 - 3000 - 017/21. URL: https://www.bundestag.de/resource/blob/832204/6813d064fab52e9b6d54cbbf5319cea3/WD-7-017-21-pdf-data.pdf. [letzter Zugriff: 06.09.2024].

Mitwirkende

Autorinnen

Prof. Dr. Judith Simon	Lehrstuhl für Ethik in der Informationstechnologie, Universität Hamburg
Prof. Dr. Indra Spiecker gen. Döhmann	Lehrstuhl für das Recht der Digitalisierung, Universität zu Köln
Prof. Dr. Ulrike von Luxburg ML	Professur für Theorie des Maschinellen Lernens, Eberhard Karls Universität Tübingen

Die mitwirkenden Wissenschaftlerinnen wurden entsprechend der veröffentlichten „Regeln für den Umgang mit Interessenkonflikten in der wissenschaftsbasierten Beratungstätigkeit der Nationalen Akademie der Wissenschaften Leopoldina“ verpflichtet, Tatsachen zu benennen, die geeignet sein können, zu Interessenkonflikten zu führen. Außerdem wird auf die vorliegenden Regeln verwiesen.

Wissenschaftliche Mitarbeit und Koordination

Dr. Sebastian Wetterich	Nationale Akademie der Wissenschaften Leopoldina
Dr. Charlotte Wiederkehr	Nationale Akademie der Wissenschaften Leopoldina
Dr. Matthias Winkler	Nationale Akademie der Wissenschaften Leopoldina

ML – Mitglied der Leopoldina

Weitere Veröffentlichungen aus der Reihe „Leopoldina Diskussion“

Nr. 33: Vernetzte Notfallvorsorge für Kulturgüter. Eine Umfrage unter den Notfallverbünden Deutschland – 2023

Nr. 32: Ein öffentlicher Dialog zur Fortpflanzungsmedizin – 2023

Nr. 31: Den kritischen Zeitpunkt nicht verpassen. Leitideen für die Transformation des Energiesystems – 2023

Nr. 30: Organisatorische Voraussetzungen der Notfallvorsorge für Kulturgüter – 2022

Nr. 29: Die rechtlichen Grundlagen der Notfallvorsorge für Kulturgüter – 2022

Nr. 28: Ärztliche Aus-, Weiter- und Fortbildung – für eine lebenslange Wissenschaftskompetenz in der Medizin – 2022

Nr. 27: Nutzen von wissenschaftlicher Evidenz – Erwartungen an wissenschaftliche Expertise – 2021

Nr. 26: Neuregelung des assistierten Suizids – Ein Beitrag zur Debatte – 2021

Nr. 25: Ansatzpunkte für eine Stärkung digitaler Pandemiebekämpfung – 2021

Nr. 24: Globale Biodiversität in der Krise – Was können Deutschland und die EU dagegen tun? – 2020

Nr. 23: Spuren unter Wasser – Das kulturelle Erbe in Nord- und Ostsee erforschen und schützen – 2019

Nr. 22: Übergewicht und Adipositas: Thesen und Empfehlungen zur Eindämmung der Epidemie – 2019

Nr. 21: Wie sich die Qualität von personenbezogenen Auswahlverfahren in der Wissenschaft verbessern lässt: Zehn Prinzipien – 2019

Nr. 20: Gemeinsam Schutz aufbauen – Verhaltenswissenschaftliche Optionen zur stärkeren Inanspruchnahme von Schutzimpfungen – 2019

Nr. 19: Die Bedeutung von Wissenschaftlichkeit für das Medizinstudium und die Promotion – 2019

Diese und weitere Diskussionspapiere der Leopoldina stehen kostenfrei unter folgendem Link zum Download zur Verfügung:
www.leopoldina.org/publikationen/stellungnahmen/diskussionspapiere

Deutsche Akademie der Naturforscher Leopoldina e. V.
– Nationale Akademie der Wissenschaften –

Jägerberg 1
06108 Halle (Saale)
Tel.: (0345) 472 39-867
E-Mail: politikberatung@leopoldina.org

Berliner Büros:
Reinhardtstraße 16
10117 Berlin

Unter den Linden 42
10117 Berlin

Die 1652 gegründete Deutsche Akademie der Naturforscher Leopoldina ist mit ihren rund 1.700 Mitgliedern aus nahezu allen Wissenschaftsbereichen eine klassische Gelehrtengesellschaft. Sie wurde 2008 zur Nationalen Akademie der Wissenschaften Deutschlands ernannt. In dieser Funktion hat sie zwei besondere Aufgaben: die Vertretung der deutschen Wissenschaft im Ausland sowie die Beratung von Politik und Öffentlichkeit.

Die Leopoldina tritt auf nationaler wie internationaler Ebene für die Freiheit und Wertschätzung der Wissenschaft ein. In ihrer Politik beratenden Funktion legt die Leopoldina fachkompetent, unabhängig, transparent und vorausschauend Empfehlungen zu gesellschaftlich relevanten Themen vor. Sie begleitet diesen Prozess mit einer kontinuierlichen Reflexion über Voraussetzungen, Normen und Folgen wissenschaftlichen Handelns.

www.leopoldina.org