Ethische Perspektiven auf Große Sprachmodelle am Beispiel von Trainingsdatenqualität

Jana Hecktor

Zusammenfassung

Die Entwicklung sogenannter Sprachmodelle hat durch ihre Einbindung in generative KI-Anwendungen in den letzten Jahren eine gesteigerte wissenschaftliche und gesellschaftliche Relevanz erhalten. Während sich ihre Einsatzfelder stetig erweitern und sie zunehmend Teil des Alltags vieler Menschen werden, haben sich parallel diverse kritische Perspektiven herausgebildet. Insbesondere die potenziell fehlerhaften, halluzinierten sowie diskriminierenden Systemausgaben stehen hierbei oftmals im Zentrum der Diskussion. Der vorliegende Beitrag widmet sich einer spezifischen Perspektive auf diese Problemfelder: einer ethisch orientierten Betrachtung der Qualität von Trainingsdaten als möglichem Ansatz, um konkrete Risiken von Sprachmodellen abzuschwächen. Hierfür werden der Leitfaden für Datenqualität in KI-Systemen des Bundesamts für Sicherheit in der Informationstechnik sowie das Glossar des Forschungsprojektes KITQAR herangezogen, um einen Überblick zu möglichen Kriterien der Trainingsdatenqualität zu gewinnen. Anschließend werden einzelne Kriterien wie Diversität oder Repräsentativität exemplarisch in den Blick genommen, um zu prüfen, inwieweit sie für das Training großer Sprachmodelle produktiv Anwendung finden können und welche Überlegungen insbesondere aufgrund möglicher „Trade-offs" zwischen einzelnen Werten aus einer ethisch geleiteten Perspektive zu berücksichtigen sind.

1. Einleitung

In den letzten Jahren haben generative KI-Systeme für einen enormen Anstieg in der Verbreitung und Nutzung von KI-Anwendungen gesorgt. Die Art und Weise, in der ChatGPT und ähnliche Systeme genutzt werden, ist dabei durchaus vielfältig und sorgt neben großem Interesse und hohen

Nutzendenzahlen[1] auch immer wieder zu kritischen oder warnenden medialen Berichterstattungen. In seinen diversen medialen Auftritten spricht auch Sam Altman, der CEO von OpenAI, immer wieder von Herausforderungen im Umgang mit dem von ihm (mit) entwickelten System. Er verweist unter anderem darauf, dass Nutzende dem System ohne Notwendigkeit höfliche Floskeln wie „Danke" und „Bitte" kommunizieren und auf diese Weise den Strom- und Rechenbedarf des Systems ‚unnötig' in die Höhe treiben (vgl. Reyes 2025). Zeigt dies die Herausforderungen von OpenAI mit möglicherweise unerwartetem Verhalten seitens der Nutzer:innen umzugehen, lassen sich die kritischen Ebenen des Systems aus ethisch angeleiteter Perspektive in der Regel vor allem in den Auswirkungen für Nutzer:innen selbst finden. So wird ChatGPT inzwischen auch für unzählige, teils sehr private, sensible, intime oder komplexe Aufgaben genutzt und dabei als eine Art Suchmaschine und Ratgeber in Einem verstanden. Nutzende fragen das System nach medizinischen Diagnosen, lassen sich einem viralen Trend auf den sozialen Medien folgend die eigene Zukunft imaginieren, verwandeln das System in einen virtuellen Partner, oder nutzen Sprachmodelle für Therapiesitzungen (vgl. Afshar 2024; Ayre/Cvejic/McCaffery 2025; Lee 2024; OpenAI 2025a; OpenAI 2025b; Raile 2024).[2] Auf diese Weise dringen Systeme wie ChatGPT in höchst vulnerable, intime und sehr private Bereiche ihrer Nutzenden ein. Ein entscheidender Grund für den Erfolg lässt sich vermutlich darin finden, dass Nutzende mit dem System via Sprache interagieren können. Mit Blick auf medizinische Diagnosen vermutet Sam Altman, dass ebenso entscheidend ist, dass Nutzende ihre Antworten direkt und schnell haben wollen. Dies sei ihnen scheinbar wichtiger als gesichert akkurate Antworten zu erhalten. In seinem Podcast warnt Altman im Wissen um all dies vor den Gefahren, die von der Nutzung ausgehen können und betont, dass ChatGPT keine Wahrheiten ausgebe, sondern stattdessen Wortnachbarn berechne. Wenngleich manche hinter diesen Warnungen einen wohl durchdachten Coup des CEOs sehen, zeigt es doch auch zwei Dinge sehr deutlich: 1. (Sprach-)Modelle wie ChatGPT[3] werden vielfältig und ubiquitär für viele, teils sehr private,

1 In seinem TED-Talk im April 2025 sprach Sam Altman davon, dass etwa 10 Prozent der Weltbevölkerung ChatGPT regelmäßig nutzen würden (vgl. TED 2025).

2 In gewisser Weise führen Systeme wie ChatGPT damit eine Art der Nutzung fort, wie sie bereits in dem 1966 von Joseph Weizenbaum entwickelten System ELIZA angelegt war.

3 Obgleich ChatGPT in der Regel als Sprachmodell diskutiert wird, setzt sich das System aus verschiedenen technischen Programmierungen zusammen.

Zwecke genutzt; 2. Die Art und Weise wie Sprachmodelle in Anwendung gebracht werden, kann durchaus problematisch sein.[4]

Zu den häufigsten Anwendungen großer Sprachmodelle gehören das Übersetzen von Texten, das Erstellen von textbasierten Inhalten, das Zusammenfassen von Informationen sowie das Beantworten konkreter Fragen auf Basis individueller Prompts.[5] Diese Fülle an Möglichkeiten hat dazu geführt, dass Sprachmodelle in ganz unterschiedlichen Feldern Anwendung finden. Unternehmen, Plattformen und Institutionen implementieren Sprachmodelle auf ihren Webseiten, in ihren Callcentern oder als Assistenz in Form von KI-Agenten und Chatbots, die Kundenanfragen beantworten, Support leisten und allgemeine Informationen bereitstellen können. In vielen Berufen finden Sprachmodelle Anwendung, um bei textbasierten Aufgaben zu unterstützen – E-Mails werden vorformuliert, Pressemitteilungen übersetzt oder Textbausteine ganz von KI geschrieben (vgl. Heesen et al. 2023). Diese Möglichkeiten der Textarbeit haben insbesondere im Bildungsdiskurs bereits einige Diskussionen ausgelöst und zu der Entwicklung vielfältiger Handreichungen geführt, die eruieren, auf welche Weise KI-Systeme im Bildungssektor genutzt werden sollten oder dürfen (vgl. Scheiter et al. 2025). Sprachmodelle werden dabei nicht allein für jene Arten der Textarbeit genutzt, die formalen Logiken folgt, sondern durchaus auch für mit Kreativität verbundene Arbeitsprozesse in den Einsatz gebracht. Exemplarisch sei an dieser Stelle auf die Games-Branche verwiesen, in der Große Sprachmodelle inzwischen unter anderem dazu genutzt werden, Charakterkonzepte zu brainstormen oder sich Anregungen für Umwelten und narrative Elemente ausgeben zu lassen (vgl. Shaker et al. 2016; Galotta et al. 2024; Thompson 2024; Yannakakis/Togelius 2024). Viele dieser Möglichkeiten lassen sich auch abseits des beruflichen oder bildungstheoretischen Kontextes im privaten Alltag nutzen. Große Sprachmodelle können dabei helfen, eine E-Mail an den Stromanbieter zu schicken, die Nachrichten des Airbnb-Gastgebers zu übersetzen oder die eigene Charaktererstellung für Pen & Paper unterstützen.

4 ChatGPT steht dabei an dieser Stelle stellvertretend für viele andere ähnliche Systeme, die als generative KI basierend auf komplexen Sprachmodellen Text- oder Bildeingaben verarbeiten, um anschließend Text- oder Bildausgaben zu produzieren.

5 Ein wesentlicher Unterschied zu der übergeordneten Kategorie der generativen KI liegt in dem Umfang des medialen Inputs und Outputs. Während Große Sprachmodelle sich auf sprachbasierte Daten konzentrieren, umfasst generative KI auch andere Medienformen wie die Bildgenerierung oder das Verarbeiten von Code.

All dies zeigt, wie vielfältig sich Sprachmodelle in verschiedene Diskurse eingeschrieben haben und wie viel Potential für zukünftige Anwendungen in ihnen liegt. Die unbestreitbare Omnipräsenz der Systeme und ihre Wirkmacht innerhalb heterogener gesellschaftlicher Prozesse führen dazu, dass auch die negativen Dimensionen der Systeme gesamtgesellschaftliche Konsequenzen haben können. Nicht überraschend haben sich daher in den letzten Jahren auch kritische ethische Perspektiven auf diese KI-Systeme eröffnet.[6] Davon abgesehen, dass selbst der Firmenchef von OpenAI auf die kritischen Dimensionen des eigenen Systems hinweist, wurde in der medialen Berichterstattung sowie innerhalb wissenschaftlicher Untersuchungen bereits vielfach deutlich gemacht, dass die Generierung sowie die Nutzung von Sprachmodellen negative Konsequenzen haben können. Dies beginnt bei der Ausgabe diskriminierender Inhalte, dem Generieren von Desinformationen oder Halluzinationen und geht hin zu dem fehlenden Schutz vulnerabler Nutzer:innen sowie Problemen des Copyrights und des Datenschutzes (vgl. Chun 2021; Heesen et al. 2021; Mehrabi et al. 2021; Loh 2024).

Bei einer Medientechnologie mit derart breit gefächerter sowie allgegenwärtiger Nutzung können auch die negativen Dimensionen der Systeme große Wirkmacht entfalten. Einen ethisch angeleiteten Blick auf die negativen Dimensionen Große Sprachmodelle zu werfen, bedeutet daher sowohl die gesamtgesellschaftlichen Herausforderungen zu adressieren als auch sich den Risiken der konkreten und individuelle Nutzung zu widmen. Der vorliegende Text möchte dabei insbesondere die inhärenten Logiken der Systeme als Ausgangspunkt konkreter Herausforderungen in den Blick nehmen. Hierfür soll sich exemplarisch den Trainingsdaten der Modelle als entscheidendes Element der Funktionalität der Systeme gewidmet werden.

6 An dieser Stelle sei darauf verwiesen, dass eine ethische Betrachtung grundsätzlich eine Abwägung von Werten und Gütern bedeutet und nicht zwingend negative Perspektiven zur Folge hat. KI-Anwendungen können durchaus positive Folgen haben oder für produktive Zwecke eingesetzt werden, beispielsweise indem Inhalte einfacher an individuelle Bedürfnisse angepasst werden, was wiederum zu einer gesteigerten Inklusion gesellschaftlicher Minderheiten führen kann.

2. Ethische Perspektiven auf Sprachmodelle

Große Sprachmodelle sind ein entscheidender Teilbereich der Entwicklung generativer KI. Viele an den Begriff KI herangetragenen kritischen Perspektiven lassen sich daher auch auf Sprachmodelle übertragen.[7] Eine Dimension, die in den letzten Jahren zunehmend Aufmerksamkeit erhalten hat, sind die ökologischen Folgen der Systeme. Als Medientechnologie basieren Sprachmodelle auf technischen Infrastrukturen und Produktionsbedingungen, die sich entlang von Fragen nach Umweltverträglichkeit, Arbeitsbedingungen und Ressourcenverteilung adressieren lassen. Für die zugrundeliegende technische Infrastruktur der Systeme werden kontinuierlich große Mengen unterschiedlicher Ressourcen in den Einsatz gebracht. Von Frischwasser zum Kühlen der Serverfarmen über seltene Edelmetalle für die technisch-materielle Grundlage hin zu einem enormen Strombedarf zur Generierung, Aufrechterhaltung sowie der konkreten Nutzung der Systeme. Die Endlichkeit der einzelnen Rohstoffe erfordert Entscheidungen, die nicht nur ökologische und ökonomische, sondern auch machtgeprägte Verteilungslogiken auf den Plan rufen. Das Wasser, welches dem Kühlen der heiß laufenden Server dient, wird insbesondere in trockenen Gebieten auch an anderer Stelle gebraucht, und die Produktion des benötigten Stroms kann nicht unbedingt (allein) durch umweltbewusste Quellen produziert werden. Im letzten Jahr ging durch die Medien, dass Google darüber nachdenkt, eigene Mini-Atomkraftwerke in Anwendung zu bringen, um den durch KI-Systeme gestiegenen eigenen Strombedarf zu decken. Edelmetalle wiederum sind selten und haben zugleich viele relevante Anwendungsbereiche. Neben der Entscheidung, wann und zu welchem Zwecke diese seltenen Stoffe weiterverarbeitet werden sollen, steht oftmals auch die teils umweltbelastende oder auf Ausbeutungslogiken basierende Gewinnung der Materialien in der Kritik (vgl. Crawford 2021; OECD 2022). Wenn Systeme wie ChatGPT für möglicherweise unpassende, unnötige, problematische oder sogar rechtlich fragwürdige sowie ethisch inakzeptable Prozesse in Anwendung gebracht werden, könnte man zurecht argumentieren, dass die umweltlichen Dimensionen der Systeme eine doppelte Ebene der Kritik eröffnen. *Doppelt* in dem Sinne, dass einerseits grundsätzlich zu reflektieren

7 Wenngleich sich dieser Text insbesondere auf die Trainingsdatenqualität als konkreten Anwendungsfall konzentriert, sollen zu Anfang auch einige allgemeinere Ebenen aufgerufen werden, die als Perspektive wiederum bei der Generierung sowie Nutzung von Trainingsdaten ebenfalls eine Rolle spielen (können).

ist, inwieweit die Nutzung der Systeme bzw. der dabei erhaltene Output die hierfür verbrauchten Ressourcen wert ist. Diese Ebene verstärkt sich andererseits noch, wenn die Ressourcen für einen Prozess verbraucht werden, der keinen substanziellen Mehrwert mit sich bringt oder dessen Ergebnisse vielleicht sogar für negative Zwecke (Deep Fakes oder Desinformationen) in Anwendung gebracht werden.[8] Eine weitere gesamtgesellschaftliche Ebene ließe sich in den Veränderungen adressieren, die Große Sprachmodelle in einzelnen gesellschaftlichen Diskursen und Lebensbereichen auslösen, wie bei den bereits erwähnten Auswirkungen in Arbeits- und Bildungskontexten. Damit die Systeme so funktionieren, wie sie es tun, bedarf es zudem explizit menschlicher Arbeit. Mit Begriffen wie Data-Work, Click-Work oder auch als Ghost Work[9] bezeichnete Arbeitsprozesse dienen der Zuarbeit von maschinellen Systemen, beispielsweise in der Klassifizierung von Datensätzen oder der Moderation von Inhalten, die den Richtlinien der jeweiligen Anbieter widersprechen. Diese Arbeit wird in der Regel in Länder mit sehr viel geringeren Löhnen ausgelagert und geschieht ohne langfristige Absicherung oder Arbeitsschutz (vgl. Gray/Suri 2019; Distelmeyer 2025).

Der zunehmenden Bedeutung sowie der negativen Wirkmacht der Systeme tragen nicht nur mediale Berichterstattungen und wissenschaftliche Untersuchungen Rechnung. In Europa hat die KI-Verordnung deutlich gemacht, dass auch politische und juristische Akteur:innen die Wirkmacht von KI-Anwendungen ernst nehmen. Technologische Systeme werden hierbei dem Risiko ihrer Nutzung folgend klassifiziert und reguliert. Wenngleich damit ein erster wichtiger Schritt getan wurde, um dem risikobehafteten Einfluss von KI-Systemen Grenzen zu setzen, muss dennoch festgehalten werden, dass die KI-Verordnung ein *regionaler* Vorstoß gegen ein international wirkmächtiges Phänomen ist. Dazu kommt, dass der Einsatz der Technologie sich oft schneller verbreitet und etabliert, als es Regularien und Forschungen einfangen können. Es scheint daher umso relevanter,

8 Ein reflektierter Umgang mit den Systemen sollte idealerweise bedeuten, mit Blick auf den großen Ressourcenbedarf oder die Fehleranfälligkeit der Systeme in Frage zu stellen, ob für die konkreten Handlung, die Nutzung von großen Sprachmodellen tatsächlich die beste oder einzige Lösung ist.

9 Jan Distelmeyer verweist in seinem Beitrag (Distelmeyer 2025) zurecht darauf, dass dieser Begriff die Unsichtbarmachung, die kritisiert werden soll, rekonstituiert und zugleich nicht für alle Arbeiter:innen zutreffend ist.

dass die entsprechenden Entwicklungen kontinuierlich kritisch perspektiviert werden.[10]

3. Systeminhärente Logiken

Neben den gesamtgesellschaftlichen Herausforderungen ist insbesondere die fehlerhafte Ausgabe von Informationen Fokus kritischer Untersuchungen. Die von Sprachmodellen generierten Inhalte können aus unterschiedlichen Gründen problematisch sein. Zu den größten Herausforderungen gehören die Halluzinationen der Systeme sowie die zum Teil in ihnen verankerten Diskriminierungen. Als Halluzination werden von KI-Systemen generierte Inhalte bezeichnet, die plausibel und wahrheitsgemäß erscheinen, aber de facto aus einer zufälligen statistischen Verteilung entstanden sind und daher weder wahr noch durch konkrete Quellen belegt sind. Einzelne aus den systeminhärenten Logiken der Systeme entstehende Halluzinationen werden regelmäßig durch viral gehende Fallbeispiele auch medial verhandelt.[11] Ist es vielleicht noch amüsant, wenn ChatGPT selbstsicher simple Additionen falsch beantwortet oder die Zugspitze zum höchsten Berg der Welt erklärt, dann sieht es schon anders aus, wenn das System souverän wissenschaftliche Quellen und Thesen für diskriminierende oder sexistische Inhalte ausgibt. Solche Halluzinationen der Systeme können in Form von glaubwürdig aussehenden Desinformationen weitreichende Folgen haben, wenn sie zum Beispiel intendiert generiert und von einzelnen Personengruppen anschließend zu manipulativen Zwecken weiterverbreitet werden.[12] Oder aber, wenn Personengruppen mit solch halluzinierten Inhalten konfrontiert werden, denen es entweder nicht möglich ist, den Wahrheitsgehalt zu überprüfen oder denen nicht bewusst ist, dass sie dies bei der Nutzung Großer Sprachmodelle tun sollten. Obgleich dies für

10 Ein Konzept hierfür sind Living Guidelines, die in regelmäßigen Abständen aktualisiert und den Entwicklungen folgend angepasst werden (vgl. ERA Forum/DG RTD 2024).

11 Viral ging so unter anderem das Video eines amerikanischen Anwalts, der sich von ChatGPT juristische Vergleichsfälle für seinen aktuellen Prozess ausgeben ließ, die sich schlussendlich als fabriziert herausstellten (vgl. Bohannon 2023).

12 Das Institut für strategischen Dialog hat im Jahr 2025 einen Bericht veröffentlicht, der darlegt, auf welche Weise rechtsextreme Akteur:sinnen in Deutschland generative KI gezielt einsetzen, um ihre Narrative online zu verbreiten (vgl. Hiller/Maristany de las Casas 2025).

alle Nutzer:innen gilt, scheint die Thematik noch relevanter, wenn man spezifische Bevölkerungsgruppen in den Blick nimmt. Vulnerable Gruppen wie Kinder, Senioren oder gesellschaftliche Minderheiten können aus unterschiedlichen Gründen besonders anfällig für die negativen Folgen von Sprachmodellen sein – insbesondere, wenn die Nutzung mit fehlender *literacy* verbunden ist, also einem mangelnden Verständnis der Funktionslogik der Systeme. Zu den Problemen, die Halluzination und Desinformation mit sich bringen, kommt hinzu, dass die Systeme möglicherweise sexualisierte, gewaltvolle, diskriminierende oder Copyright- und Persönlichkeitsrechte missachtende Inhalte ausgeben. Wenngleich dies in den Richtlinien der jeweiligen Unternehmen ausgeschlossen wird und ChatGPT bei konkreten Worten und Anfragen auch ausgibt „Dieser Inhalt verstößt möglicherweise gegen unsere Nutzungsrichtlinien", können entsprechenden Grenzen und Beschränkungen durch abweichendes Prompting gezielt (*jailbreaking*) sowie unabsichtlich umgangen werden (vgl. Liu et al. 2023). Obgleich also Filter und Grenzen für gewaltvolle, diskriminierende oder sexuelle Inhalte eingerichtet worden sind, können diese absichtlich sowie auch unabsichtlich umgangen werden und auf diese Weise entsprechende Inhalte auch an Kinder und Jugendliche ausgegeben werden, die davor geschützt werden sollten.[13] Auch andere Personengruppen können gefährdende oder sich negativ auswirkende Inhalte ausgegeben bekommen, beispielsweise wenn aufgrund traumatischer Erfahrungen konkrete ‚Triggerpunkte' existieren.

Die Herausforderungen, die sich aus der konkreten und individuellen Nutzung in Kombination mit den inhärenten Logiken der Systeme ergeben, können auf verschiedene Systemlogiken und Anwendungsprozesse zurückgeführt werden. Ein konkretes Scharnier der Funktionslogik sind die jeweils genutzten Daten. Sie bilden die Trainingsgrundlage der Systeme und damit das Fundament der Modellierung, sie sind Input (Prompt) sowie Output, aber auch elementarer Bestandteil der Anwendungslogik. Sie sind nicht nur notwendiges technisches Element der auf maschinellem Lernen basierenden Modellierungsverfahren, sie bilden zugleich die Grenzen und den Möglichkeitsraum der daraus entstehenden Sprachmodelle. Wenngleich nicht alle mit Großen Sprachmodellen verknüpften Probleme und Herausforderungen durch einen Blick auf die Trainingsdaten erklärt oder durch den Einsatz von entsprechenden ethischen Standards gelöst

13 Verstärkend kommt an dieser Stelle hinzu, dass es für ChatGPT zwar ein Mindestalter gibt, dieses bei der Nutzung allerdings nicht abgefragt wird.

werden können, lassen sich doch einige der angeführten Probleme auf die konkreten Datenmengen zurückführen.

4. Datenauswertung als entscheidendes Scharnier von Sprachmodellen

Derzeitige Entwicklungen des maschinellen Lernens hängen entscheidend an den jeweils vorhandenen und in den Einsatz gebrachten Trainingsdaten. Ohne zu tief in die technischen Funktionalitäten einzusteigen, sei darauf verwiesen, dass für die Entwicklung generativer KI sowie Großer Sprachmodelle sogenannte Deep Learning-Verfahren angewandt werden. Hierbei werden künstliche neuronale Netze mit einer großen Menge gelabelter oder auch ungelabelter Daten gespeist und trainiert. Im Laufe der Trainingsphase werden die Gewichtungen der Systeme kontinuierlich angepasst, sodass am Ende das Modell die ihm aufgetragene Aufgabe erfolgreich durchführen kann. Im Falle von Sprachmodellen ist dies das Identifizieren von konkreten Worten und ihren Relationen zueinander sowie das Antizipieren der darauf basierenden gewünschten Antwort mittels Mustererkennung. Daten bilden auf diese Weise einerseits die Grundlage für das Training der Modelle – ohne sie kann kein Sprachmodell entstehen –, setzen dadurch aber auch die Grenzen und Möglichkeitsbedingungen fest. Inhalte, die nicht Teil der Datenmenge sind, können nicht ausgegeben werden. Daneben schreiben sich auch implizit in den Daten vorhandene Relationen in die Modellierungen ein, dies kann im Extremfall dazu führen, dass ein wie auch immer gearteter Bias in den Datenmengen sich auch in den Ergebnissen der Systeme widerspiegelt. Dies allein zeigt bereits, welche Bedeutung der Implementierung der ‚richtigen' Daten für die Funktionalität der jeweiligen Systeme zukommt. So hat das Bundesamt für Sicherheit in der Informationstechnik (BSI) Anfang Juli 2025 einen methodischen Leitfaden zur Datenqualität in KI-Systemen vorgestellt und auch in der KI-Verordnung werden Daten und Daten-Governance als entscheidende Ebene adressiert und reguliert.

Neben der zum Zwecke des Datenschutzes im Jahr 2016 bereits in Kraft getretenen Datenschutz-Grundverordnung (DSGVO), deren Regulierungen auch für seitdem entstandene Technologien Gültigkeit besitzt, listet auch die KI-Verordnung explizite Kriterien und Regulierungen für den Einsatz von Daten im Diskursfeld rund um KI. Zu den spezifischen Regulierungen, die sich auf Große Sprachmodelle anwenden lassen, gehören neben der Datenqualität und der Daten-Governance auch Transparenzforde-

rungen, Risikobewertung und -management sowie die Pflicht zur Überwachung und Berichterstattung der Systeme. Mit der Transparenzforderung ist festgehalten, dass Anbieter von Sprachmodellen klar kommunizieren müssen, wenn die jeweiligen Inhalte von KI-Systemen generiert sind. Auf diese Weise soll verhindert werden, dass Nutzer:innen fälschlicherweise davon ausgehen, dass die Inhalte von menschlichen Akteur:innen erstellt wurden und entsprechende Zuschreibungen vornehmen. Unter dem Aspekt der Risikobewertung und dem Risikomanagement wird wiederum festgehalten, dass Sprachmodelle einer entsprechenden Bewertung unterzogen werden müssen, durch die Gefahren identifiziert und minimiert werden. Dieser Aspekt verweist auch bereits auf die Einteilung von Technologien basierend auf ihrem Risiko-Level. Große Sprachmodelle als Teilbereich generativer KI können je nach Anwendungskontext als Hoch-Risiko-System eingestuft werden. Aus diesem Grund sind Anbieter von Sprachmodellen wie OpenAI auch angehalten, die Leistung und Sicherheit ihrer Systeme kontinuierlich zu kontrollieren und entsprechende Berichte über die jeweilige Nutzung und die damit verbundenen Risiken zu erstellen. Nicht zuletzt hält die KI-Verordnung fest, dass es eine gewisse Form der Datenqualitätssicherung und Daten-Governance braucht. Dies bedeutet, dass Firmen Maßnahmen ergreifen sollten, um sicherzustellen, dass die Daten, die für das Training von Sprachmodellen eingesetzt werden, entsprechende Qualitätsmerkmale erfüllen und insbesondere frei von diskriminierenden oder schädlichen Inhalten sind. Damit gesetzliche Rahmungen oder ethische Standards für Trainingsdaten in Anwendung gebracht werden können, braucht es konkrete Trainingsdaten, die an diesen Standards und Richtlinien entlang bewertet werden können.

Eine erste wichtige Frage vor der Generierung Großer Sprachmodelle ist daher, ob es für den entsprechenden Anwendungsfall bereits entsprechende Datenmengen gibt, die die gewünschten Inhalte abbilden, oder ob erst Daten erhoben werden müssen. Für den Fall, dass es noch keine (adäquaten) Daten gibt und diese beschaffen oder erfasst werden müssen, ist es relevant zu überprüfen, ob für den konkreten Anwendungsfall eine datafizielle Erfassung überhaupt möglich ist, einerseits mit Blick auf technische Möglichkeitsbeschränkungen wie unzureichender Sensorik oder aber bei nur schwer (objektiv) zu dokumentierenden Phänomenen.

Darüber hinaus sollte in jedem Fall entschieden werden, ob eine Datenerfassung die möglichen Risiken des jeweils konkreten Anwendungsfall Wert ist. Fallen die entsprechenden Bereiche beispielsweise in den Schutz

der Privatsphäre und das Recht der informationellen Selbstbestimmung? Werden vulnerable Gruppen einem potentiell diskriminierenden Blick geöffnet? Ist bereits absehbar, dass die erfassten Daten durch die Einbindung in ein Sprachmodell möglicherweise problematischen Nutzungslogiken ausgeliefert sein werden? Wenngleich nicht all diese Fragen vorab beantwortet werden können, sollten sie für eine ethisch angeleitete Generierung von Modellierungen mitbedacht werden. Gesetzt den Fall, dass es sowohl möglich ist, Daten zu erheben oder auf vorhandene Datensätze zurückzugreifen und es auch keine ethischen oder gesetzlichen Einwände gegen die Erhebung oder Auswertung der jeweiligen Datensätze gibt, kann es dennoch sein, dass aus den scheinbar unproblematischen Daten, problematische Modellierungen entstehen.[14] Es sollte daher auch gefragt werden: Welche Qualität haben die genutzten Datensätze? Und auf welche Weise können Daten kritisch betrachtet und ethischen Standards folgend erhoben und eingesetzt werden?

5. Trainingsdatenqualität als Grundbedingung für ethische Sprachmodelle

Das Forschungsprojekt KITQAR, an dem die Universität Tübingen (Internationales Zentrum für Ethik in den Wissenschaften), das Hasso-Plattner-Institut (HPI), die Universität zu Köln und der Verband der Elektrotechnik, Elektronik und Informationstechnik (VDE) beteiligt sind, hat in seiner Laufzeit die Qualität von KI-Test- und Trainingsdaten in der digitalen Arbeitsgesellschaft erforscht. Ein zentrales Ergebnis dieses Projekts ist die Generierung und Veröffentlichung eines Glossars zur Datenqualität (vgl. Mohammed et al. 2023). Datenqualität wird dabei definiert als die „Eignung von Daten für einen bestimmten Anwendungszweck" (Mohammed et al. 2023: 1). Die Qualität selbst kann anhand ganz unterschiedlicher Kriterien bewertet werden.

Einige dieser Kriterien sind automatisch messbar oder überprüfbar, während andere nur von Expert:innen bewertet werden können. Manche der Kriterien sind nicht nur voneinander abhängig, sondern beeinflussen oder nivellieren sich. Wenngleich für den Einsatz von Trainingsdaten in der digitalen Arbeitsgesellschaft entwickelt, lassen sich die entsprechenden Kriterien auch für die Entwicklung von Großen Sprachmodellen in Anwendung

14 Beispielsweise dann, wenn die Zusammenführung zuvor unabhängiger Daten und Datensätze in einem Modell zu neuen Auswertungslogiken führt.

bringen. In dem vom Bundesamt für Sicherheit in der Informationstechnik (BSI) veröffentlichten methodischen Leitfaden zur Datenqualität in KI-Systemen werden wiederum zehn Kriterien aufgelistet, die sich direkt sowie indirekt auch im Glossar von KITQAR wiederfinden lassen. Für das BSI sind bei der Entwicklung generativer KI folgender Kriterien entscheidend: Repräsentativität, Vollständigkeit, Genauigkeit, Konsistenz, Korrektheit, Einheitlichkeit, Gültigkeit, Eindeutigkeit, sichere Quellen sowie die Überprüfung auf Personenbezug bei den jeweils genutzten Daten und damit auch die Einhaltung des Datenschutzes. Die zehn Kriterien werden in dem Leitfaden durch konkrete Bausteine wie Vielfalt, Ausgewogenheit, Konsistenzsicherung oder Expertenanalyse ergänzt.[15] KITQAR listet 29 Kriterien zur Bewertung der Datenqualität auf, die im Kontext für KI-Test- und Trainingsdatenqualität in der digitalen Arbeitsgesellschaft von Bedeutung sind. Von der *Aktualität* der Daten bis hin zu Fragen der *Zugänglichkeit* werden die verschiedenen Kriterien nicht nur erklärt und in Relation zueinander gesetzt, sondern auch mit der DSGVO und dem KI-Act untermauert.

Sowohl KITQAR als auch das BSI inkludieren in ihren Kriterien verschiedene Ebenen der Trainingsdatenqualität. Einige beziehen sich auf die Generierung oder die jeweilige Herkunft der Datenmenge, andere fokussieren den konkreten Inhalt der einzelnen Daten sowie ihre Relation zueinander und wieder andere adressieren den kontinuierlichen Umgang mit bereits eingebundenen Datenmengen. Für alle Kriterien gilt jedoch, dass in der Regel nicht alle von ihnen zugleich in Anwendung gebracht werden können. Oftmals braucht es Abwägungen, die zu einem Trade-Off von ethischen Werten und Risiken führen können. Um die Wirkmacht, aber auch die Grenzen von ethischen Kriterien für Trainingsdatenqualität aufzuzeigen, sollen nachfolgend exemplarisch einige der Kriterien am Beispiel Großer Sprachmodelle genauer ausgeführt werden. Trainingsdaten bilden einerseits die notwendige Grundbedingung für die Modellbildung von Großen Sprachmodellen, müssen aber selbst oftmals erst produziert, zusammengeführt und nach konkreten Kriterien aufgearbeitet werden, bevor sie nutzbar sind. Eine erste entscheidende Perspektive von Kriterien für Trainingsdatenqualität richtet sich daher auf die Herstellung sowie die Akquirierung der jeweiligen Trainingsdaten. KITQAR listet so *Ansehen* als eines ihrer Kriterien und referiert damit auf die Vertrauenswürdigkeit der Datenquelle. Dieses Kriterium anzuwenden, bedeutet einerseits, dass

15 Insgesamt listet der Leitfaden 15 Bausteine auf.

Datenquellen, zu denen bereits (gute) Erfahrungswerte vorliegen, jenen vorzuziehen sind, bei denen man nicht einschätzen kann, welche Qualität die jeweiligen Daten haben werden. Der Leitfaden des BSI wiederum verweist darauf, dass stets auf sichere Quellen zurückgegriffen werden muss, so zum Beispiel im wissenschaftlichen Kontext auf Veröffentlichungen, die mit Peer Review überprüft worden sind. Die Realität sieht jedoch oft anders aus. ChatGPT stand bereits mehrfach in der Kritik dafür, dass seine Trainingsdaten durch effizientes Scraping aus dem Netz gezogen wurden, ohne stark zu differenzieren, welche Quellen sie dabei anzapfen.[16] Die meisten der sowohl von KITQAR als auch in den Leitlinien aufgelisteter Kriterien widmen sich der inhaltlichen Qualität der Trainingsdaten. Von den 29 Kriterien für Trainingsdatenqualität lassen sich für die ethische Betrachtung Großer Sprachmodelle vor allem die Kriterien Ausgewogenheit, Datenschutz, Diversität, Fairness, Korrektheit, Privatsphäre und Repräsentativität benennen.

Datensätze sind, gemäß dem Glossar, dann *ausgewogen*, wenn die Datenpunkte innerhalb des repräsentierten Wertebereichs im Verhältnis zueinander gleich verteilt sind. Für den Fall, dass für die Erfassung von Kund:innendaten beispielsweise nach Altersgruppen sortiert werden soll, braucht es für eine ausgewogene Datenmenge, pro in das Training inkludierter Altersgruppe, eine entsprechend gleich große Datenmenge. Auch das BSI führt Ausgewogenheit als Kriterium an und ordnet diesem die Funktion zu, Verzerrungen in Form von Unter- oder Überrepräsentation entgegenwirken zu können. Auf Große Sprachmodelle übertragen, braucht es ein alternatives Beispiel. Exemplarisch davon ausgehend, dass ein Sprachmodell dazu generiert und trainiert werden soll, um in einem Bildungsprogramm für Migrant:innen, den Schüler:innen aus verschiedenen Ländern personalisierte Lerninhalte bereitzustellen, müsste für einen ausgewogenen Datensatz sichergestellt werden, dass das Modell eine gleich verteilte und damit ausgewogene Menge an Daten zu jedem relevanten Migrationshintergrund erhält. Zu beachten ist, dass Ausgewogenheit hier lediglich bedeutet, alle relevanten Gruppen des Trainingsdatensatzes im Verhältnis gleich einzubeziehen, um so die Über- oder Unterrepräsentation einzelner Gruppen zu vermeiden. Das bedeutet nicht zwingend gleichermaßen eine allumfassende bzw. die Gesellschaft vollständig abbildende Datenerfassung.

16 Als erstes Medienunternehmen hat die New York Times daher sowohl OpenAI als auch Microsoft auf Grund von Urheberrechtsverletzung verklagt (vgl. Freeman et al. 2024).

Entscheidend ist, welches Ziel die jeweilige Modellierung hat und welche weiteren Kriterien in den Einsatz gebracht werden. Um bei dem Beispiel des Glossars zu bleiben: Wenn ein Unternehmen ermitteln möchte, welche Produkte Kund:innen, die über 30 Jahre alt sind, präferieren, ist es wenig zielführend, die Altersgruppe unter 30 in die Modellierung einzubeziehen. Auf Große Sprachmodelle übertragen hieße das, wenn die Zielgruppe des Sprachmodells ausschließlich aus französischsprachigen Herkunftsländern stammt, ist es vermutlich weniger zielführend, die spanische Sprache mittels entsprechender Trainingsdaten in das System einzupflegen.

Damit ein Datensatz *divers* ist, muss, gemäß dem KITQAR-Glossar, „jede Entität der Domäne in der Datenmenge repräsentiert" sein, also jeder Entitätstyp der Gesamtmenge mindestens einmal vorkommen (Mohammed et al. 2023: 3). Es geht bei diesem Kriterium nicht um eine möglichst ausgewogene Verteilung, sondern darum, alle Einzelentitäten der vorhandenen Gesamtmenge zu inkludieren. Wenn also Mitarbeitende einer Firma anhand ihres Alters erfasst werden sollen, dann muss jedes Alter, zu dem es mindestens einen Mitarbeitenden gibt, Teil der Trainingsdatenmenge sein. Das BSI wiederum fasst diese Ebene unter dem Begriff der Vielfalt und zielt darauf ab, die Varianz der Datenmengen zu maximieren. Der Leitfaden verweist an dieser Stelle sogar explizit auf Sprachmodelle und argumentiert, dass die Vielfalt der menschlichen Sprache – explizit im Sinne von Dialekten, Jargons und Akzenten – in Sprachmodellen implementiert sein sollten. Die beiden Kriterien *Ausgewogenheit* und *Diversität* können deckungsgleich sein, müssen es aber nicht. Eines von beiden an die jeweiligen Trainingsdaten anzulegen, bedeutet daher nicht unbedingt, dass auch das andere erfüllt ist.

Eng verwoben mit den beiden Kriterien Diversität und Ausgewogenheit listen sowohl KITQAR als auch das BSI *Repräsentativität*. Dieses meint, dass jede Entität der zu repräsentierenden Gesamtmenge die gleiche Chance hat, in der jeweiligen Datenmenge repräsentiert zu sein (vgl. Mohammed et al. 2023). Ein repräsentativer Datensatz spiegelt also die statistischen Verteilungsverhältnisse der realen Gesamtmenge wider. Dabei sollte allerdings kritisch mitbedacht werden, dass es vorkommen kann, dass durch einen statistisch repräsentativen Datensatz das daraufhin trainierte Modell für einzelne Entitäten weniger geeignet ist, da nicht ausreichend repräsentative Daten vorhanden sind. Wenn beispielsweise ein Sprachmodell Studierenden Fragen zu typischen Herausforderungen im Universitätskontext beantworten können soll, kann es durchaus sein, dass ein auf repräsentativen Daten trainiertes Modell weder für Personen im Rollstuhl

noch für Personen unter 16 Jahren nützliche Antworten ausgeben kann, da diese nur einen geringen Anteil der Studierenden ausmachen und daher in Relation weniger Einfluss auf die Modellbildung genommen haben. Das BSI definiert den Begriff wiederum mit Rückbezug auf und in Abgrenzung zu Konzepten wie Ausgewogenheit, Vielfalt oder Gültigkeit. Der Leitfaden nutzt das Beispiel eines Systems, das dafür eingesetzt wird, die Kreditwürdigkeit einzelner Kund:innen zu errechnen. Mit Blick auf die Trainingsdatenqualität wird exemplifiziert: „Um die Repräsentativität zu gewährleisten, müssen die Trainingsdaten eine breite und faire Strichprobe der gesamten Bevölkerung enthalten" (BSI 2025: 7).

Fairness befasst sich im Diskursfeld rund um KI-Anwendungen in der Regel mit der Identifizierung, Analyse und Quantifizierung von Verzerrungen (Bias), die Individuen und Gruppen nach bestimmten Merkmalen wie Geschlecht, Ethnie oder Behinderung unzulässig diskriminieren (vgl. AIEIG 2020; Heesen et al. 2021; Mehrabi et al. 2021). Systeme können beispielsweise dann als fair bezeichnet werden, wenn ihre Trainingsdaten die Standards zur Freiheit von diskriminierenden Verzerrungen erfüllen oder die darauf basierenden Systeme keine diskriminierenden Inhalte ausgeben. Fairness ist als Begriff jedoch durchaus heterogen konzipiert. Von einem fairen Sprachmodell zu sprechen, kann auch meinen, dass keine ethischen Werte missachtet oder Personen in der Generierung oder Nutzung diskriminiert oder ausgebeutet wurden. Fairness kann also auf die Qualität der Daten, die Bedingungen ihrer Produktion oder die Bedingungen ihrer Anwendung oder aber auf die Fairness der schlussendlichen Modellierungen abzielen.

Ähnlich breit anwendbar ist die Ebene von Privatsphäre und Datenschutz. Als entscheidende Grundlage für die Generierung der Modellierungen existiert ein großes Interesse daran, möglichst viele Daten über möglichst viele Bereiche zu haben. Dabei gleichermaßen Datenschutz und Privatsphäre zu wahren – insbesondere, wenn globale Perspektiven und länderspezifisch unterschiedliche Regulierungen und Verständnisse zu Bedenken sind – birgt eigene Herausforderungen. Dies beginnt bei den eingangs aufgeworfenen Fragen danach, an welcher Stelle und in welchen Zusammenhängen welche Daten erhoben werden sollen oder dürfen. Und auch wenn die Erhebung datenschutzkonform geschieht, existieren weitere Herausforderungen. So kann unzureichender Schutz der Daten zu Missbrauch oder Verlust der Privatsphäre führen. Neben den Diskussionen darum, welche Inhalte von KI-Systemen aus dem Netz oder auch während

der konkreten Nutzung abgegriffen und für das Training genutzt werden dürfen, besteht auch die Gefahr, dass gespeicherte Inhalte beispielsweise aufgrund von Datenlecks durch externe Parteien zugegriffen und weiter genutzt werden können. Wenn Große Sprachmodelle für medizinische Diagnosen oder für Formen der Gesprächstherapie genutzt werden, fließen in die Systeme höchst private, persönliche und vor allem je nach Auswertungslogik vulnerable Informationen ein, die anschließend abgegriffen werden können. Dies führt wiederum zu neuen Formen der mangelnden Kontrolle über die eigenen Daten. Dieser Aspekt gewinnt an Relevanz, wenn man die Nutzung der Systeme durch vulnerable Gruppen in den Blick nimmt, denen aus diversen Gründen vielleicht nicht möglich ist, zu erfassen, welche Folgen die Preisgabe intimer Daten haben kann. OpenAI ist sich dieser Problematik durchaus bewusst. Fragt man ChatGPT selbst, welche Gefahren darin liegen können, intime Informationen mit dem System zu teilen, gibt dieses an, die Probleme reichten von unzureichendem Datenschutz oder nicht ausreichend geschützter Privatsphäre über das möglicherweise überhöhte Vertrauen in die Fähigkeiten des Systems bis hin zu emotionaler Abhängigkeit seitens der Nutzer:innen, gekoppelt mit fehlender Empathie seitens des Systems.[17] Obgleich das System seinen Nutzer:innen also auf Nachfrage durchaus erklärt, dass es problematisch sein kann, persönliche oder intime Daten einzugeben, hat OpenAI nur begrenzt Kontrolle darüber, welche Inhalte Nutzer:innen in das System einspeisen. ChatGPT reagiert, sobald die Eingabe gemacht wurde, die Daten von dem System also bereits verarbeitet und damit sowohl digital übermittelt als auch in das System eingespeist worden sind. Dazu kommt, dass auch hier ein Trade-Off zwischen einzelnen Kriterien adressiert werden kann. Denn der (Daten)Schutz vulnerabler Gruppen kann zur Folge haben, dass die Daten einzelner vulnerabler Gruppen nicht in das Training der jeweiligen Modelle mit einbezogen werden (können). Dies wiederum vermindert möglicherweise die Diversität des Modells und kann auch bedeuten, dass die durch das Systeme ermöglichten Unterstützungsangebote – wie das Übersetzen von Inhalten – nicht für alle Bevölkerungsgruppen gleichermaßen effizient ausgearbeitet werden können.

Mögliche Maßnahmen, um diesen Daten-Problemen entgegenzuwirken, sind die Implementierung von Sicherheitsmaßnahmen wie Verschlüsselungen und Anonymisierungen, DSGVO-konforme Datenverarbeitung sowie

17 Diese Informationen sind die Antwort von ChatGPT auf Basis des Prompts „Welche Konsequenzen kann es haben, dir zu intime Informationen zu teilen?“.

transparente Datenschutzerklärungen und Opt-Out Optionen, oder auch die Reduktion der Datensammlung auf ein Minimum. Ebenfalls dazu beitragen kann die Einhaltung und Überprüfung der bis hierhin exemplarisch aufgelisteten Kriterien für Trainingsdatenqualität. Der exemplarische Blick auf einzelne Kriterien hat bereits gezeigt, dass einzelne ethische Standards im Widerspruch zueinanderstehen, sowie – je nach Anwendungskontext – in ihrer Relevanz variieren können. Ein erstes Fazit kann daher sein, dass es nicht nur entscheidend ist, die Kriterien stets bedacht und reflektiert in den Einsatz zu bringen, sondern es durchaus auch interdisziplinäre und anwendungsfallbezogene Diskussionen darum geben muss, welche ethischen Standards jeweils in Einsatz gebracht werden können oder müssen. Das wiederum zeigt auch, dass es nur schwer eine Automatisierung solcher Standards geben kann.

6. Fazit

Die bis hierhin schlaglichtartige Betrachtung ethischer Perspektiven hat deutlich gemacht, dass ethische Standards für Trainingsdatenqualität ein wichtiger Ansatz sein können, um konkrete Risiken von Sprachmodellen abzuschwächen. Nichtsdestotrotz bleibt es dabei, dass Sprachmodelle in ihrer Nutzung auch bei Beachtung von Trainingsdatenqualität sowohl in ihrer systeminhärenten Logik, ihren konkreten Anwendungsmomenten als auch in ihrer allgemeinen gesamtgesellschaftlichen Bedeutung durchaus negative Wirkmacht entfalten können. Die Nutzung sowie die Generierung der Systeme sollte daher stets unter Berücksichtigung der mit dieser Medientechnologie verbundenen Konsequenzen durchgeführt werden. Dies bedeutet einerseits, dass weiterhin kritische Untersuchungen der Nachhaltigkeit der Systeme aus ethischer Sicht von Relevanz sein sollten, ebenso wie der Blick auf die hinter den Systemen liegenden menschlichen Arbeit oder mit Blick auf die gesellschaftlichen Veränderungen, die sie auslösen. Darüber hinaus scheint es ebenfalls notwendig, die konkreten Nutzungspraktiken zu adressieren, die Große Sprachmodelle ermöglichen. Insbesondere, wenn diese möglicherweise negative Folgen für die einzelnen Nutzenden haben können. Wichtig zu untersuchen wäre darüber hinaus auch, inwieweit sich mit Blick auf die zum Teil abweichende Realität des Trainings von Großen Sprachmodellen, die jeweiligen Kriterien der Trainingsdatenqualität überhaupt produktiv anwenden lassen können. Wenn Systeme wie ChatGPT ihre Daten über Scraping abgreifen und mit riesigen

Datenmengen arbeiten, lassen sich dann Kriterien wie Diversität, Fairness und Ausgewogenheit noch adäquat anbringen? An dieser Stelle ließe sich auch fragen, wo die Grenzen einer ethischen Betrachtung liegen oder liegen können und ob es vielleicht Bereiche gibt, in denen es sinnvoll ist, entsprechende Systeme gar nicht erst in Anwendung zu bringen.

Literatur

Afshar, Melissa Fleur (2024): People Are Using ChatGPT to Help Them Achieve Their 'Dream Life', in: Newsweek, 05. November 2024 (online unter: https://www.newsweek.-com/people-using-chatgpt-help-achieve-dream-life-1979801 – letzter Zugriff: 5.11.2025).

AI Ethics Impact Group (AIEIG) (2020): From Principles to Practice – An interdisciplinary framework to operationalise AI ethics, Gütersloh.

Ayre, Julie / Cvejic, Erin / McCaffery, Kirsten J. (2025): Use of ChatGPT to obtain health information in Australia, 2024: insights from a nationally representative survey, in: The Medical Journal of Australia 222 (4/2025), S. 210–212.

Bohannon, Molly (2023): Lawyer Used ChatGPT in Court and Cited Fake Cases. A Judge Is Considering Sanctions, in: Forbes, 08. Juni 2023 (online unter: https://www.forbes.com-/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/ – letzter Zugriff: 5.11.2025).

Bundesamt für Sicherheit in der Informationstechnik (2025): QUAI-DAL. Teildokument B: „02-Qualitätskriterien & Bausteine", in: Bundesamt für Sicherheit in der Informationstechnik, 01. Juli 2025 (online unter: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/-QUAI-DAL_B_Qualitaetskriterien.pdf?__blob=publicationFile&v=4 – letzter Zugriff: 5.11.2025).

Chun, Wendy Hui Kyong (2021): Discriminating Data. Correlation, Neighborhoods, and the New Politics of Recognition, Cambridge, MA.

Crawford, Kate (2021): The Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence, New Haven, London.

Distelmeyer, Jan (2025): Mit KI zu tun bekommen – Daten, Arbeit und Interfaces: Was können wir von Plattformen wie ChatGPT (und sie von uns) wissen?, in: cargo 65 (3/2025), S. 5.

European Research Area Forum & Directorate General for Research and Innovation (2024): Living guidelines on the responsible use of generative AI in research, in: Europäische Kommission, 15. April 2024 (online unter: https://european-research-area.ec.europa.eu/news-/living-guidelines-responsible-use-generative-ai-research-published – letzter Zugriff: 5.11.2025).

Europäische Union (2024): Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 über künstliche Intelligenz und zur Änderung bestimmter Rechtsakte der Union, in: Amtsblatt der Europäischen Union, L 206, S. 1–161 (online unter: https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A32024R1689 – letzter Zugriff: 5.11.2025).

Freeman, Joshua et al. (2024): Exploring Memorization and Copyright Violation in Frontier LLMs: A Study of the New York Times v. OpenAI 2023 Lawsuit, in: arXiv, 09. Dezember 2024 (online unter: https://doi.org/10.48550/arXiv.2412.06370 – letzter Zugriff: 4.8.2025).

Gallotta, Roberto et al. (2024): Large Language Models and Games: A Survey and Roadmap, in: arXiv, 09. Dezember 2024 (online unter: https://doi.org/10.48550/-arXiv.2402.18659.18659v1 – letzter Zugriff: 4.8.2025).

Gray, Mary L. / Suri, Siddharth (2019): Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass, Boston, New York.

Heesen, Jessica / Reinhardt, Karoline / Schelenz, Laura (2021): Diskriminierung durch Algorithmen vermeiden: Analysen und Instrumente für eine digitale demokratische Gesellschaft, in: Gero Bauer et al. (Hg.), Diskriminierung und Antidiskriminierung. Beiträge aus Wissenschaft und Praxis, Bielefeld, S. 129–148.

Hiller, Anna / Maristany de las Casas, Pablo (2025): Generative KI und die deutsche extreme Rechte. Narrative, Taktiken und digitale Strategien (online unter: https://isdgermany.org/-generative-ki-und-die-deutsche-extreme-rechte-narrative-taktiken-und-digitale-strategien/ – letzter Zugriff: 5.11.2025).

Lee, Ian (2024): 4 Theses on Boyfriend Dan GPT, in: The Last Organizer, 31. März 2024 (online unter: https://medium.com/thelastorganizer/4-theses-on-dan-gpt-98bb2a682b5b – letzter Zugriff: 5.11.2025).

Liu, Yi et al. (2023): Jailbreaking ChatGPT via Prompt Engineering: An empirical study, in: arXiv, 23. Mai 2023 (online unter: https://arxiv.org/abs/-2305.13860 – letzter Zugriff: 5.11.2025).

Loh, Wulf (2024): Generative KI, digitale Teilhabe und epistemische Ungerechtigkeit, in: RphZ – Zeitschrift für Religion, Politik und Gesellschaft 10 (2/2024), S. 215–233.

Mehrabi, Ninareh et al. (2021): A survey on Bias and Fairness in Machine Learning, in: ACM Computing Surveys 54 (6/2021), Article 115.

OECD (2022): Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint (= OECD Digital Economy Papers, No. 341), Paris.

OpenAI (2025a): AI boyfriend (online unter: https://chatgpt.com/g/g-RwZG2pDs2-ai-boyfriend – letzter Zugriff: 5.11.2025).

OpenAI (2025b): TherapyAI (online unter: https://chatgpt.com/g/g-8yHB0UD8j-therapyai – letzter Zugriff: 5.11.2025).

Paris, Martine (2025): ChatGPT Hits 1 Billion Users? 'Doubled In Just Weeks' Says OpenAI CEO, in: Forbes, 13. April 2025 (online unter: https://www.forbes.com/sites/martineparis/-2025/04/12/chatgpt-hits-1-billion-users-openai-ceo-says-doubled-in-weeks/– letzter Zugriff: 5.11.2025).

Heesen, Jessica et al. (2023): Künstliche Intelligenz im Journalismus. Potenziale und Herausforderungen für Medienschaffende. Whitepaper aus der Plattform Lernende Systeme, München. https://doi.org/10.48669/pls_2023-1

Raile, Paolo (2024): The usefulness of ChatGPT for psychotherapists and patients, in: Humanities and Social Sciences Communications 11, Article 47.

Reyes, Marta (2025): Why You Shouldn't Say "Thank You" and "Please" to ChatGPT, in: Medium, 24. April 2025 (online unter: https://medium.com/@martareyessuarez25/why-you-shouldnt-say-thank-you-and-please-to-chatgpt-2910da23b3f3 – letzter Zugriff: 5.11.2025).

Scheiter, Katharina et al. (2025): Künstliche Intelligenz in der Schule. Eine Handreichung zum Stand in Wissenschaft und Praxis, hrsg. im Rahmen des KI-Begleitprozesses im Rahmenprogramm empirische Bildungsforschung, Bonn (online unter: https://www.empirische-bildungsforschung-bmbfsfj.de/img/KI_Review.pdf – letzter Zugriff: 5.11.2025).

Mohammed, Sedir et al. (2023): Ein Glossar zur Datenqualität (1.2), in: Zenodo, 06. März 2023 (online unter: https://doi.org/10.5281/zenodo.7702426 – letzter Zugriff: 5.11.2025).

Shaker, Noor / Togelius, Julian / Nelson, Mark J. (2016): Procedural Content Generation in Games, Cham.

TED (2025): OpenAI's Sam Altman talks ChatGPT, AI agents and superintelligence – Live at TED2025, in: YouTube, 11. April 2025 (online unter: https://www.youtube.com/watch?v-=5MWT_doo68k – letzter Zugriff: 5.11.2025).

Thompson, Tommy (2024): The Changing Landscape of AI for Game Development, in: Paul Roberts (Hg.), Game AI Uncovered, Boca Raton, S. 1–11.

Yannakakis, Georgios N. / Togelius, Julian (2024): Artificial Intelligence and Games, 2. Aufl., Cham.