iit perspektive

Working Paper des Instituts für Innovation und Technik (iit) Nr. 75

# Ethik und Auswirkungen von KI mit minimalem Risiko

## Empfehlungen für die Forschungs- und Technologieförderung ausgehend von LLM

Rebecca Puchta, Arno Wilhelm-Weidner

iit perspektive

Working Paper des Instituts für Innovation und Technik (iit) Nr. 75

Rebecca Puchta, Arno Wilhelm-Weidner

# Ethik und Auswirkungen von KI mit minimalem Risiko – Empfehlungen für die Forschungs- und Technologieförderung ausgehend von LLM

## Einleitung

Im Fokus dieser iit-perspektive stehen ethische Herausforderungen der am 1. August 2024 in Kraft getretenen KI-Verordnung der Europäischen Union (KI-Verordnung, gemäß Fassung vom 13. Juni 2024). Die KI-Verordnung verfolgt einen risikobasierten Ansatz zur Bewertung der Einsatzfelder, Verwendungszwecke und Funktionsweisen von KI-Technologien. Die KI-Verordnung umfasst dabei vier Risikostufen. Zu ihnen zählen inakzeptable, hohe, begrenzte sowie minimale Risiken der Inverkehrbringung, Inbetriebnahme und Verwendung von KI-Technologien (KI-Verordnung, Erwägungsgrund Nr. 1).

Der Gesetzestext birgt zwei Auffälligkeiten. Einerseits zeigt er, dass KI-Praktiken mit minimalem Risiko in der KI-Verordnung weder explizit als solche benannt werden noch dezidiert ausgeführt wird, wie diese Risiken minimiert werden können. Ähnlich wie KI-Praktiken mit begrenztem Risiko bleiben solche mit minimalem Risiko wenig hinreichend definiert. Daraus ergeben sich, lautet sowohl das Lob als auch die Kritik an der KI-Verordnung, Handlungsspielräume beispielhaft für wirtschaftliche Akteure wie Start-ups, aber auch Rechtsunsicherheiten und umstrittene Auslegungen der Vorschriften wie beispielsweise zwischen zivilgesellschaftlichen, öffentlichen und wirtschaftlichen Akteuren. Dies ist der Ausgangspunkt für die Betrachtung ausgewählter ethischer Herausforderungen von KI mit minimalem Risiko, deren Ziel es ist, Empfehlungen für die Forschungs-, Bildungs- und Technologieförderung herzuleiten.

Andererseits wendet sich die KI-Verordnung zur Umgehung dieser Problematik „bestimmten KI-Systemen" zu. Unter diesem unkonkreten, weit gefassten Begriff fallen vor allem „KI-Modelle mit allgemeinem Verwendungszweck", die in englischsprachigen Kontexten als „general purpose artificial intelligence" diskutiert werden und spätestens seit dem Erfolg von ChatGPT im Jahr 2023 den öffentlichen Diskurs um KI prägen. Derartige KI-Modelle mit allgemeinem Verwendungszweck, so die KI-Verordnung, würden mitunter systemische Risiken bergen. Damit wird ein Begriff des risikobasierten Ansatzes aufgerufen, der sowohl hinsichtlich ethischen als auch technischen Gesichtspunkten eine genauere Betrachtung erfordert. Denn durch ihn wird die Bestimmung von begrenzten und minimalen Risiken zu einer komplexen Praxis, die die Entwicklung von KI-Technologien zukünftig prägen wird.

Im Folgenden werden zunächst die Grundannahmen der KI-Verordnung in Bezug auf KI-Systeme skizziert, um die Rolle von KI-Modellen mit allgemeinem Verwendungszweck in der Verordnung in einem nächsten Schritt zu fokussieren. Um die Spezifik und Komplexität dieser KI-Modelle zu eruieren und weiterführend Empfehlungen für die Förderung in Forschung, Wissenschaft und Bildung sowie der Technologie- und Innovationsentwicklung zu formulieren, stehen Large Language Modelle (LLM) im Mittelpunkt. Auf Basis zweier Fallbeispiele zeigen wir, welche ethischen Herausforderungen bei der Bestimmung systemischer Risiken von LLM bestehen. Davon ausgehend werden strategische Überlegungen und konkrete Empfehlungen für die Innovations- und Forschungsförderung abgeleitet, mit dem Ziel, einen Beitrag dazu zu leisten, KI-Modelle mit allgemeinem Verwendungszweck ethischer zu gestalten. Die iit-perspektive fokussiert dabei vorrangig Fragen, die konkret nutzende Individuen von KI-Technologien betreffen.

## Large Language Modelle in der KI-Verordnung: Ethische Implikationen und Herausforderungen

### Was ist die KI-Verordnung?

Die KI-Verordnung ist das europäische Instrument zur Regulierung der Entwicklung, des Einsatzes und des Gebrauchs von Systemen Künstlicher Intelligenz (KI). Seit ihrem Inkrafttreten zum 1. August 2024 bestimmt sie die Art und Weise, wie KI-

Technologien u. a. mit den in der Charta der Grundrechte der Europäischen Union festgeschriebenen Werten, Schutzpflichten und Leitbildern zukünftig in Einklang gebracht werden können. Ziel der Verordnung, die im August 2026 vollumfängliche Gültigkeit erlangen wird, ist es, die Implementierung und Anwendung von KI in allen Bereichen der EU menschenzentriert sowie vertrauenswürdig zu gestalten und auf diese Weise sicherzustellen, dass Demokratie, Rechtsstaatlichkeit und Umweltschutz nicht gefährdet werden (KI-Verordnung, Erwägungsgrund Nr.1).

Dieses Ziel wird durch einen „risikobasierten Ansatz" verfolgt, demzufolge es in Bezug auf KI-Praktiken vier Risikostufen gebe. Neben verbotener KI, die ein unannehmbares, inakzeptables Risiko birgt, gibt es als Hochrisiko-KI eingestufte KI-Praktiken sowie solche, die entweder ein begrenztes oder bloß minimales beziehungsweise geringes Risiko aufweisen. Ermessen wird der Grad des Risikos einer KI jeweils daran, ob und wenn ja, inwiefern sie Menschen in ihren Rechten, ihrer Sicherheit und Gesundheit negativ beeinflusst, einschränkt, manipuliert oder behindert. Neben der Reduktion von Menschen auf Zahlenwerte wie bei der verbotenen Praxis des Social Scoring oder der Missachtung von Rechten sowie der Diskriminierung von Menschen beispielsweise durch biometrische Hochrisiko-KI-Praktiken, wurden in der KI-Verordnung auch digitale Phänomene wie Des- und Misinformation, Fake News und Deepfakes sowie Wahlmanipulationen als potenzielle KI-Schäden angeführt, die es zum Schutz rechtstaatlicher und demokratischer Prozesse zu verhindern, kontrollieren und minimieren gelte.

**Bestimmte KI-Systeme und KI-Modelle mit allgemeinem Verwendungszweck**

Auffällig an der KI-Verordnung ist, dass KI-Praktiken mit minimalem Risiko nicht explizit definiert werden. Hingegen werden minimale Risiken negativ anhand der Definition „bestimmter KI-Systeme" in Kapitel VI und in Bezug auf „KI-Modelle mit allgemeinem Verwendungszweck" in Kapitel V charakterisiert. Mit Letzterem ist laut Verordnung ein KI-Modell gemeint, „das eine erhebliche allgemeine Verwendbarkeit aufweist und in der Lage ist, unabhängig von der Art und Weise seines Inverkehrbringens ein breites Spektrum an unterschiedlichen Aufgaben kompetent zu erfüllen und das [dabei] in eine Vielzahl nachgelagerter Systeme oder Anwendungen integriert werden kann" (KI-Verordnung, Artikel 3, Absatz 63).

Im Folgenden fokussieren wir große, generative KI-Modelle, die digitale Inhalte wie Text, Audio, Bild und Video flexibel handhaben, damit diverse Aufgaben erfüllen, in andere KI-Systeme integriert werden können und dabei dazu beitragen, dass die Modelle skaliert, erweitert, optimiert und modifiziert werden können. Bei diesem Prozess greifen die Modelle auf die Resultate riesiger Datenmengen zurück, mittels deren Analyse und Sortierung das Modell trainiert wurde (KI-Verordnung, Erwägungsgründe Nr. 79, 99 und 105). Große Large Language Modelle wie z. B. BERT (Google, 2018), ChatGPT (OpenAI, 2022) oder LLaMA (Meta AI, 2023) zählen zu jenen KI-Modellen, die gegenwärtig ganze Fachdiskurse und die öffentliche Medienberichterstattung über KI prägen (Bender et al., 2021, 613). Genauso zählen aber auch bestimmte Chatbots sowie große Suchmaschinen und Online-Plattformen (KI-Verordnung, Erwägungsgrund Nr. 120) zu KI-Systemen mit allgemeinem Verwendungszweck.

Ähnlich wie die Bestimmung minimaler Risiken ist die Risikobewertung der KI-Modelle mit allgemeinem Verwendungszweck diffizil, weil sie über den Umweg der Festlegung systemischer Risiken erfolgt, die diese Modelle von Beginn an bergen und während ihrer Anwendung zudem dynamisch entwickeln können (KI-Verordnung, Erwägungsgrund Nr. 110). Zu diesen systemischen Risiken zählen neben Unfällen und Störungen kritischer Infrastruktur negative Auswirkungen für die Gesundheit und Sicherheit der Bevölkerung genauso wie Formen der negativen Beeinflussung und Manipulation demokratischer und öffentlicher Prozesse. Neben der Gefährung der wirtschaftlichen Sicherheit eines Landes wird nicht zuletzt die Verbreitung illegaler, falscher oder diskriminierender Inhalte mittels KI als systemisches Risiko erachtet.

Für diesen Beitrag ist es vor dem Hintergrund dieses allgemeinen Problemaufrisses zentral, zunächst die Vielfalt der mit KI-Modellen mit allgemeinem Verwendungszweck verbundenen potenziellen systemischen Gefahren zu eruieren. Um dies im Umfang der vorliegenden Studie adäquat vornehmen zu können, konzentrieren sich die folgenden Ausführungen ausschließlich auf Large Language Modelle.

## Large Language Modelle: Funktionsweise, Risiken und Ethik

**Eingrenzung von Large Language Modellen**

Large Language Modelle (LLM) nehmen in der KI-Verordnung eine Sonderstellung ein, schließlich sind sie darin noch wenig berücksichtigt worden (Henning 2024). Der KI-Verordnung zufolge zählt ein LLM jedoch als „KI-Modell mit allgemeinem Verwendungszweck" (KI-Verordnung, Seite 50). Dies bedeutet, dass LLM in der Lage sind, einer Vielzahl von Zwecken sowohl für die direkte Verwendung als auch für die Integration in andere KI-Systeme zu dienen. Gut abgrenzen lassen sich LLM von der Funktionsweise einfacher Chatprogramme eines Smartphones, wie der Kulturwissenschaftler Michael Seemann konstatiert. Während beim Eintippen einer Nachricht auf dem Smartphone das nächste Wort als Wahrscheinliches in der Abfolge eines Satzes vorgeschlagen wird, würden LLM die Vorhersage ganzer Wortreihen, Bild- oder Tonfolgen errechnen (Seemann 2023, 9). Indem im Trainingsprozess riesige Mengen von

Texten, Bildern oder anderen Inhalten auf der Suche nach der Häufigkeit und Regelmäßigkeit ihres jeweiligen Vorkommens durchsucht werden, lernt die Maschine über diese Texte, Bilder oder anderen Inhalte, wie sie sich statistisch zueinander verhalten (Ebd. 11).

**Diskussion von LLM als KI mit minimalem Risiko**

Mehr als bei vielen anderen technischen Systemen hängt der Aspekt der Risikobewertung von LLM vom Verwendungszweck ab. Der individuelle Gebrauch von LLM zur Unterstützung für Formulierungen bei Briefen, der Suche nach Wortspielen oder zur Übersetzung kann in vielen Fällen harmlos sein. Der Gebrauch von LLM für Hausaufgaben, Abschlussarbeiten oder der Teilnahme an politischen Diskussionen kann hingegen die adäquate, heißt leistungserforderliche Herausbildung kognitiver Fähigkeiten in Schule und Hochschule verzögern und komplizieren als auch die Informiertheit und Richtigkeit der verwendeten Ergebnisse stark beeinträchtigen. Gleichzeitig kann der Gebrauch von LLM zur Recherche durch die Halluzinationen dieser Systeme massive Probleme hervorrufen. Als Halluzinationen werden faktisch falsche Aussagen eines LLM bezeichnet, die aufgrund der komplexen Wahrscheinlichkeitsrechnung der LLM nicht immer wissensbasiert, sondern auch frei erfunden sein können. Zudem wird, um ein Einsatzfeld zu nennen, das hier keine Berücksichtigung mehr findet, der Einsatz von LLM für Spam und Cyberangriffe bis hin zur einfacheren Herstellung von Biowaffen durch Amateure immer einfacher (Rubinic et al. 2024).

Aufgrund dessen kann in Bezug auf LLM allgemein und spezifisch nur schwer vom Bestehen minimaler Risiken ausgegangen werden. Dies wird im Folgenden ausgeführt. Hierzu werden zunächst zwei Fallbeispiele gegeben, die anschließend in theoretische Überlegungen zu dieser systemischen Problematik einfließen.

Fallbeispiel (1)

Ein Mann ist unsicher, welche Partei er wählen möchte. Deshalb nutzt er ein LLM wie zum Beispiel ChatGPT, um im Austausch mit ihm über die Agenden politischer Parteien und deren unterschiedliche Standpunkte informiert zu werden. Der Mann legt dem technischen System seine persönlichen Interessen, Sorgen und Probleme offen dar und gibt dabei viele Daten von sich preis. Die aggregierten Daten des Mannes werden von der Firma, die das LLM betreibt, als Werbeprofil verkauft, auf dessen Basis dem Mann in Zukunft gezielter Werbung angezeigt werden kann. Die entsprechende Einwilligung dazu hat der Mann beim Start des LLM durch das Anklicken der AGB beiläufig bereits erteilt. Die Informationen, die der Mann vom LLM erhält, setzen sich aus bekannten Fakten und Halluzinationen der dem KI-System zugrundeliegenden Modelle zusammen und spiegeln weder die aktuelle Parteienlandschaft noch aktuelle politische Standpunkte wider. Hingegen präsentiert das LLM auf Basis der errechneten Antworten des Mannes Inhalte von Verschwörungstheorien. Vom LLM werden diese völlig glaubhaft beschrieben und führen so weit, dass der Mann diese übertriebenen Geschichten als glaubwürdig interpretiert.

Fallbeispiel (2)

Eine Wissenschaftlerin eines großen Forschungsinstituts nutzt ein öffentlich verfügbares LLM, um die von ihr erhobenen statistischen Daten aus einem Experiment automatisiert auswerten zu lassen. Ohne die eigentlich notwendigen Tests zur Güte der Daten zu machen, werden die Daten durch das LLM verarbeitet. Das LLM-Ergebnis ist gut aufbereitet, die statistischen Tests haben klare Ergebnisse und lassen sich gut präsentieren, weshalb die Wissenschaftlerin sie für die Verwertung ihrer wissenschaftlichen Ergebnisse nutzen und weitergeben möchte. Da die Verteilung der Daten die verwendeten Tests eigentlich nicht zugelassen hätte und die Daten auch sonst nicht auf Unsauberkeiten hin überprüft wurden, werden die kondensierten Ergebnisse weitergereicht, ohne dass sie genutzt werden sollten. Gleichzeitig wurden die Daten bei der Nutzung durch das LLM an die amerikanischen Server des Systems gesendet und werden seither in Kombination mit dem Prompt der Wissenschaftlerin und der Antwort zum weiteren Training des LLM genutzt. Durch geschickte bösartige Prompts schaffen es andere Wissenschaftler:innen, das LLM so zu nutzen, dass es die vertraulichen Daten des Forschungsinstituts wieder ausgibt. Sie veröffentlichen die Daten und ziehen eilig eigene Schlüsse daraus, wodurch sie sich die Arbeit des Forschungsinstituts aneignen.

**Theoretische Einordnung und Reflektion systemischer Risiken von LLM**

Im Anschluss an diese Fallbeispiele werden nun im Folgenden wesentliche theoretische Aspekte ausgeführt, um die mit LLM einhergehenden systemischen Risiken einordnen zu können. Aus der KI-Verordnung geht hervor, dass systemische Risiken die Funktionsweise von KI-Modellen selbst betreffen (KI-Verordnung, Artikel 3, Nummer 56[1]). Funktionsweise meint dabei sowohl die Art und Weise, wie LLM technisch als Modelle funktionieren, als auch wie die LLM in verschiedenen Gebrauchskontexten Anwendung finden und genutzt werden. Systemische Risiken können ethische Konsequenzen haben oder sich aus ethischen Herausforderungen ergeben. Unter Zuhilfenahme des viel beachteten und für diese iit-perspektive titelbildgeben-

1 Systemisches Risiko definiert die KI-Verordnung als „Risiko, das für die Fähigkeiten mit hoher Wirkkraft von KI-Modellen mit allgemeinem Verwendungszweck spezifisch ist und aufgrund deren Reichweite oder aufgrund tatsächlicher oder vernünftigerweise vorhersehbarer negativer Folgen für die öffentliche Gesundheit, die Sicherheit, die öffentliche Sicherheit, die Grundrechte oder die Gesellschaft insgesamt erhebliche Auswirkungen auf den Unionsmarkt hat, die sich in großem Umfang über die gesamte Wertschöpfungskette hinweg verbreiten können."

den Texts „On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" der Computerlinguistikerinnen und Informatikerinnen Emily M. Bender, Timnit Gebru, Angelina McMillan-Major und Margaret Mitchell (Bender et al. 2021), sollen jene unerwünschten systemischen KI-Schäden mit ethischer Tragweite aufgeschlüsselt werden. Die Autorinnen sprechen dabei vier Hauptrisiken von LLM an. Zu diesen zählen die Umweltkosten der KI-Modelle, deren mangelnde Nachvollziehbarkeit, deren Potenzial, Falschinformationen in Umlauf zu bringen und des Weiteren die Gefahr, dass deren Ergebnisse über- und falschbewertet werden. Diese Aspekte werden im Folgenden ausgeführt und kurz in Bezug zu den Fallbeispielen gestellt.

*Umweltkosten*

Um die Energieintensität großer und, wie mit Fallbeispiel 2 zum Ausdruck kommt, weltweit vernetzter KI-Modelle mit allgemeinem Verwendungszweck besser zu verstehen, sieht die KI-Verordnung vor, Normungsaufträge zu lancieren, durch welche die Energiekosten und -effizienz eruiert und bestimmt werden (KI-Verordnung, Artikel 40; Samsi et al. 2023). LLM erfordern hohe Rechenkapazitäten und das Housing der dafür erforderlichen Rechner in Datenzentren ist sowohl energie-, d. h. strom- und wasser- als auch flächenintensiv (Pasquinelli 2024, 273). In den Vereinigten Staaten wird erwartet, dass KI-Praktiken im Jahr 2026 sechs Prozent des Strombedarfs der Nation ausmachen werden (Ren und Wierman 2024). In Bezug auf das Jahr 2030 wird global betrachtet bereits davon ausgegangen, dass voraussichtlich 30 Prozent des weltweiten Gesamtenergieverbrauchs im Bereich des maschinellen Lernens und damit der Künstlichen Intelligenz liegen (Bolón-Canedo et al. 2024). Die Kosten für Energie liegen dabei nicht ausschließlich bei Betreibern von KI-Modellen, sondern sind auf alle Infrastrukturbetreiber, Netzwerke sowie die User:innen in unterschiedlichem Maße verteilt.

*Unergründlichkeit und Nicht-Nachvollziehbarkeit*

LLM sind unergründlich und hochkomplex und vielleicht auch deshalb so erfolgreich. Ihre technischen und rechnerischen Operationen überschreiten die menschliche Wahrnehmungsfähigkeit bei Weitem, wie der Medienwissenschaftler Shane Denson konstatiert hat (2016). Deshalb steht die Entwicklung und Testung von Modellen mittels Trainingsdaten im Zentrum, wie der Soziologe und Systemtheoretiker Armin Nassehi in seinem Buch „Muster. Theorie der digitalen Gesellschaft" (2019) beschrieben hat. Nassehi folgend ließe sich KI allgemein als selbstlernende Technik definieren, die sich in den letzten Jahrzehnten stark ausdifferenziert hat (z. B. Natural Language Processing, Deep Learning, Cognitive Computing und so weiter). Dabei gleichen sich die unterschiedlichen Verfahren grundlegend darin, dass sie Muster anhand der Festlegung von Häufigkeiten, dem Erkennen von Ähnlichkeiten und dem Finden von Regelmäßigkeiten extrahieren (Ebd. 229). In Plattformökonomien werden diese Muster als Datenbündel im Hinblick auf ihre Netzwerkeffekte und ihr Skalierungspotenzial ausgeschöpft. Aus diesem grundlegenden Ordnungsprozess heraus ergibt sich das Risiko, dass die Vielfalt der Daten stark und mitunter zu stark reduziert wird, sodass die Unterschiedlichkeit und Besonderheit digitaler Objekte und Inhalte verkannt wird und das erkannte Muster nicht für die einzelnen Daten, sondern eine spezifische Menge an Daten repräsentativ ist.

Diese Charakteristik bezieht sich auf den Umfang, die Spezifik, die Qualität und Repräsentativität der Trainingsdaten, mit welchen KI-Modelle trainiert werden. Die Datensätze von LLM wie ChatGPT werden mitunter durch Crawling-Techniken erweitert, wodurch ausgewählte Daten aus dem World Wide Web herangezogen werden. Die dem Crawling zur Verfügung stehenden Datensätze sind dabei nicht repräsentativ (Tao et al. 2024, 2). Unklar bleibt z. B., mittels welcher Klassifikatoren Trainingsdaten erhoben und strukturiert wurden. Uneinheitlich sind zudem die Zugangsmöglichkeiten zum sowie die konkreten Nutzungsweisen des Internets, zumal sie sich weltweit und auch entlang von individuellen Präferenzen und Handlungsweisen sehr stark unterscheiden. Nicht zuletzt variieren sowohl Praktiken der Kuration als auch der Dokumentation von Daten immens (Bender et al. 613). Unergründlich bleibt also kurzum, welche Daten das Modell tragen und auf welchem Datenschema die Algorithmen trainiert wurden. Der Effekt: Die Aussagekraft der Modelle ist stark eingeschränkt, weil die dem Training zugrundeliegenden Daten beispielsweise die Eigenheiten der Sprachgewohnheiten, Alltagspraktiken, Gewohnheiten oder Konzepte (meist minorisierter) Bevölkerungsgruppen nicht abbilden und dabei gesellschaftliche Hierarchien fortschreiben (Ebd. 614).

Dabei gelten gerade die Reichweite und die Integrierbarkeit von KI-Modellen mit allgemeinem Verwendungszweck in andere KI-Systeme laut KI-Verordnung als deren zwei zentrale Qualitäten. Zum einen rekurrieren KI-Modelle auf riesigen Big-Data-Datensätzen. Durch die fortwährende Integration von immer neuen Daten werden diese sowohl umfassender als auch intransparenter. Zum anderen fungieren die KI-Modelle aufgrund ihrer Integrierbarkeit selbst als so bezeichnete „foundational models" beziehungsweise Basismodelle, deren leichte Einführung in andere Systeme und Modelle zugleich die Wahrscheinlichkeit für die bloße Übernahme von Fehlern erhöhe (Mock et al. 2024). Wie Fallbeispiel 2 zu veranschaulichen vermag, muss die Güte der Daten als Quelle systemischer Risiken bewertet werden, zumal in Beispiel 2 die Integrierbarkeit und Reichweite zu urheberrechtlichen Herausforderungen und Fragen zur Abgrenzung von Forschungsspionage führt.

*Falschinformationen, Fehleranfälligkeit und Halluzinationen*

Im Anschluss an die obigen Erläuterungen müssen LLM Michael Seemann zufolge des Weiteren „gravierende Wissenslü-

cken, Interpretationsfehler, Logikfehler und Halluzinationen" bescheinigt werden (2023, 27). Ungeachtet der Tatsachen, dass LLM Anweisungen von Menschen durch zufriedenstellende Antworten bearbeiten oder grammatikalisch korrekte und flüssig lesbare Texte sowie Programme erzeugen können und darüber hinaus diverse Sprachen kennen, Übersetzungsleistungen beschleunigen und als kreativ zu bewerten sind, ist deren hohe und noch immer nicht nachvollziehbare Fehleranfälligkeit äußerst kritisch zu bewerten (Ebd., 28).

Wie Pegah Maham und Sabrina Küspert in ihrem Policy Brief „Governing General Purpose AI. A Comprehensive Map of Unreliability, Misuse and Systemic Risks" für den Think Tank „interface" (ehemals „Stiftung Neue Verantwortung") im Juli 2023 erklären, sei es noch nicht zu gewährleisten, dass KI-Modelle sich wie beabsichtigt verhalten. Unklar sei noch immer, wie genau die internen Mechanismen, die mitunter selbst für ihre Entwickler:innen nicht in allen Auswirkungen verständlich seien, zu kontrollieren sind. Dies verringere die Vertrauenswürdigkeit der Modelle, führe zur Verbreitung von falschen, nicht vollständigen und wie in Bezug auf Fallbeispiel 1 exemplifizierten manipulativen oder irreführenden Informationen. Dies erhöhe nicht nur das Unfallrisiko, sondern führe zu Verletzungen der Privatsphäre (Maham und Küspert 2023, 21–25).

*Bewertung der Bedeutung LLM-generierter Informationen*
Zuletzt sind die ethischen Implikationen und Folgen der Indienstnahme und Inverkehrbringung von LLM im Hinblick auf deren Stellenwert zu bewerten, die sie im öffentlichen Diskurs und im individuellen Gebrauch durch Nutzer:innen genießen und gewinnen. Implizit, so Bender und ihre Co-Autorinnen, nehmen Menschen eine bestimmte statistische Beziehung zu dem Input und dem Output sowie der Form von LLM ein (Bender 2021, 615). Dieser Vorgang betrifft die Art und Weise dessen, wie Menschen LLM-generierte Informationen als sinnvoll bewerten und als bedeutungstragend verbreiten (Ebd., 611). Auf den ersten Blick können die mithilfe von Technik rasch generierten und einfach modellierten Informationen objektiv oder neutral wahrgenommen werden. Doch Technik operiert nie unbeeinflusst. Die derartige Bewertung technisch generierter Informationen birgt unzählige Probleme, nicht zuletzt, weil erst beziehungsweise gerade die Verwendung und Distribution von (potenziell) verzerrten Informationen eben jene Biases festigen und fertigen.

Diese Gefahr thematisiert Fallbeispiel 1, in welchem ein LLM den Meinungsbildungsprozess eines Bürgers für eine anstehende demokratische Wahl beeinflusst. Es ist davon ausgehend nicht unbedenklich oder im Hinblick auf öffentliche und demokratische Prozesse nicht ungefährlich, KI-Modelle mit allgemeinem Verwendungszweck als „Allzweck-KI" zu definieren oder gar zu bagatellisieren, wie es während der Entwurfsphase und Ausarbeitung der KI-Verordnung durch multinationale Unternehmen wie Microsoft und Google unternommen wurde (Kühl 2023). Zwar kann dieser Begriff aus kulturwissenschaftlicher Sicht als der kühne Wunsch eingeordnet werden, mittels des Einsatzes von KI ein vielseitig einsetzbares und im Sinne wirtschaftlicher Interessen effizientes sowie effektives Schlagwort für den Einsatz kostspieliger KI zu entwickeln. Gleichzeitig wirkt dieser Begriff im schlimmsten Fall jedoch auch auf ein Missverständnis hin, dass eben jenes KI-Modell allgemeingültiges Wissen generieren und diesem Wissen Authentizität zugesprochen werden könne. Diese Dynamik befördert die Verbreitung nicht nachvollziehbarer oder sachlich unkorrekter Informationen.

## Handlungsempfehlungen

Vor dem Hintergrund der skizzierten technischen Funktionsweisen und gesellschaftlichen Dynamiken von LLM, die als Beispiel eines KI-Modells mit allgemeinem Verwendungszweck erachtet wurden, reflektiert diese iit-perspektive abschließend ausgewählte Empfehlungen für die Forschungs-, Wissenschafts- und Technologieförderung – mit dem Ziel, die Diskriminierungsarmut und Informationsqualität von KI-Modellen allgemeiner Verwendungszwecke zu erhöhen und den souveränen Umgang mit bei der Bewertung minimaler Risiken von KI-Modellen zu sichern. Dabei werden allgemeine strategische Überlegungen mit Ideen für die technische Weiterentwicklung und die Beforschung von KI-Modellen verknüpft.

### Allgemeine und strategische Überlegungen

*Förderung von Informationsbewertungskompetenz*
In Anbetracht der dargelegten ethischen Herausforderungen im Umgang mit KI-Modellen von minimalem Risiko gilt es aus unserer Sicht Informationsbewertungskompetenz aller Bevölkerungsteile über Altersstufen und Ausbildungsgrade hinweg zu fördern. Der Begriff „Informationsbewertungskompetenz" verbindet dabei den für den wissenschaftlichen Kontext relevanten Begriff Informationskompetenz (Sauerwein 2024) mit dem der Bewertungskompetenz, wie er bislang vor allem im MINT-Bereich fachdidaktisch Berücksichtigung erfuhr (Bögeholz et al. 2018). Zugleich schließt der Begriff Formen der KI-, Digital- sowie Daten- und Medienkompetenz nicht aus, rahmt oder grundiert sie vielmehr. Informationsbewertungskompetenz adressiert genauer die individuellen und gesellschaftlichen Formen des souveränen Umgangs mit Informationen, die durch digitale Geräte und Infrastrukturen erzeugt und vernetzte Umgebungen verteilt werden. Deshalb klammert der Begriff das komplexe Zusammenspiel zwischen Daten und Metadaten, Distributions- sowie Vernetzungslogiken digitaler Infrastrukturen und Plattformen, Datei- und Medienformate usf. nicht aus.

**Empfehlungen für die Forschungs- und Innovationsförderung**

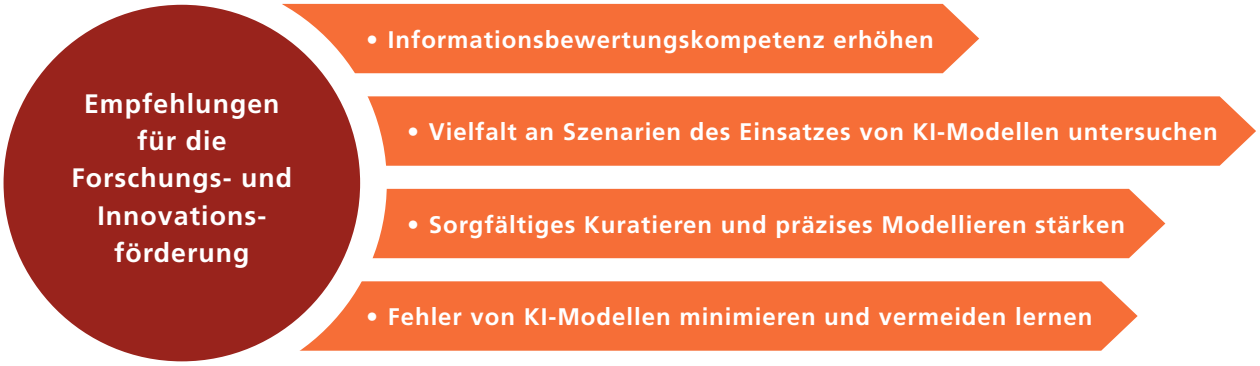


**Abbildung 1:** Empfehlungen für die Forschungs- und Innovationsförderung

Informationsbewertungskompetenz richtet sich auf die Kompetenz und Fähigkeit, KI-generierte, datengetriebene und durch Medien prozessierte und verbreitete Informationen in ihren jeweiligen Entstehungskontexten und Gebrauchsweisen reflektieren, einschätzen, untersuchen und prüfen zu können. Der Begriff definiert damit zugleich die Kompetenz, die Gewinnung und Vermittlung, die Präsentation und allgemeine Nutzung von Informationen als Prozesse der Aufmerksamkeitslenkung zu erkennen und sie als Bedingung individueller und gemeinschaftlicher Wissensproduktion anzuerkennen. Die Förderung von Informationsbewertungskompetenz kann eine von vielen für User:innen relevanten Lösungen sein, die gesellschaftlichen Schäden von KI zu minimieren. Denn laut den Forscherinnen des Wissenschaftszentrums für Sozialforschung Berlin, Jelena Cupać und Mitja Sienknecht, habe die KI-Verordnung bereits während der Entwurfsphase darin gezögert, den „Schaden für die Gesellschaft" durch KI-Modelle mit allgemeinem Verwendungszweck zu benennen (Cupać und Sienknecht 2022). Und selbst nach Fertigstellung der KI-Verordnung monieren Expert:innen wie zum Beispiel von netzpolitik.org weiterhin, dass die Regeln für LLM erst noch geschrieben werden würden (Henning 2024; Kurz und Ullrich 2025).

Unserem Verständnis nach unterscheidet sich Informationsbewertungskompetenz von der in der KI-Verordnung eingeforderten und definierten KI-Kompetenz, weil KI-Kompetenz sich fast ausschließlich auf die „KI-Wertschöpfungskette" im ökonomischen Sinn bezieht (KI-VO Artikel 3, Absatz 56 sowie Erwägungsgrund Nr. 20) und dabei die ethischen Herausforderungen der Kommerzialisierung von Daten mittels KI zu wenig konkret berücksichtigt werden. Gegenüber von KI-Kompetenz, der zufolge es darum ginge, sich „der Chancen und Risiken von KI und möglicher Schäden, die sie verursachen kann, bewusst zu werden" (KI-VO Artikel 3, Absatz 56), richtet sich Informationsbewertungskompetenz weniger nur auf die Bewusstwerdung. Vielmehr geht es darum, den sozialen Wert von Informationen sowie deren gesellschaftliche Bedeutung im Informationszeitalter selbst ins Zentrum zu rücken und sie als umkämpftes Gut (nicht-)demokratischer Gesellschaften neu zu betrachten. Gerade die Komplexität und Unergründlichkeit von KI-Modellen, aber auch die Offenheit des politischen Prozesses bis hin zur vollumfänglichen Anwendung der KI-Verordnung machen es wesentlich, KI-generierte Informationen adäquat bewerten zu können.

Obgleich die Kompetenz, Informationen auf Urheberschaft und Nachvollziehbarkeit, ihre Aktualität und Qualität sowie institutionelle Rahmung hin adäquat bewerten zu können, neben der Recherchekompetenz zum Beispiel als Teil der wissenschaftlichen Ausbildung erachtet werden kann, hat Informationsbewertungskompetenz durch die Digitalisierung, die Prozesse der Verdatung und die zunehmende Vernetzung in nicht-hochschulischen Kontexten längst den Status einer alltäglichen Anforderung erlangt.

ChatGPT gibt auf die gestellte Frage nach der Bewertung KI-basierter Informationen an, dass dafür mehrere Komponenten relevant seien. Es gelte ein Verständnis dafür zu entwickeln, wie KI funktioniere; den Kontext, die verwendeten Quellen, die Annahmen und Logik der KI kritisch einzubeziehen; Fachwissen einzubinden, Vergleiche vorzunehmen, designierte Schwachstellen von KI ernst zu nehmen und Unsicherheiten dieser zu erkennen. Man könnte argumentieren, dass ChatGPT selbst vorschlägt, Informationsbewertungskompetenz im Umgang mit LLM zur Anwendung zu bringen. Unter welchen Entstehungs- und Vermittlungsbedingungen sich Informationsbewertungs-

kompetenz jedoch entwickelt und floriert, welche Bedeutung dieser Kompetenz beigemessen wird und welche Bedarfe sich in schulischen und hochschulischen wie beruflichen und alltäglichen Kontexten bei ihrer Herausbildung und Sicherung konkret zeigen, gilt es zu erforschen.

### *„Liebes ChatGPT, kannst Du mir sagen, wie ich lerne, die Informationen eines KI-Modells zu bewerten?“*

*Szenarienbasierte Förderung*
Um dem Fehlen von Regulierungserfahrung in Bezug auf „KI-Modelle mit allgemeinem Verwendungszweck“ zu begegnen, schlägt die KI-Verordnung die Ausarbeitung von Praxisleitfäden (KI-Verordnung Artikel 56) und die Förderung von KI-Reallaboren (KI-Verordnung Artikel 57) vor. Für deren ordnungsgemäße Konzeption, Ausarbeitung und Indienstnahme ist es empfehlenswert, interdisziplinäre Projekte zu fördern, die sich mit verschiedensten Variationen möglicher Regulierungsverfahren von LLM auseinandersetzen. Diese interdisziplinären Konstellationen können die Praxisleitfäden sowohl begleiten und unterstützen als auch erproben und testen. Insbesondere können die Leitfäden und Reallabore als Methodik und Instrument der Forschungsförderung selbst betrachtet werden, zumal sie sich dazu eignen, wissenschaftsbasierte Empfehlungen zum Umgang mit LLM nach dem Wortlaut der KI-Verordnung zu generieren und zu konkretisieren.

**Technische Überlegungen**
*Förderung sorgfältiger Kuration und präziser Modellierung*
Die institutionalisierte Förderung der Erforschung sowie die Forschungsförderung von KI-Modellen selbst verstärken ethische Herausforderungen im Umgang mit KI. Zu beobachten, so Bender und Co, sei beispielhaft die Diskrepanz in der Erforschung von LLM, die sich stark auf die bloße Korrektur in der Erfassung von natürlicher Sprache konzentriere. Dabei werde vernachlässigt, dass das Sprachsystem erst im Gebrauch, d. h. in konkreten Situationen, spezifischen Kontexten, individuellen Sprachgewohnheiten Bedeutung erlangt (Bender und Koller 2020). Da KI-Modelle lediglich statistische Form extrahieren, gelte es technische Zugriffe zu entwickeln, die es erlauben, die Fähigkeiten der Modelle sorgfältiger zu deklinieren und akkurater zu charakterisieren (Bender et al. 615), statt sie bloß nach statistischen Logiken im Hinblick auf ihre Größe, Reichweite und Skalierung zu betreiben.

Der Datenwissenschaftler Arash Hajikhani und die Betriebswirtschaftlerin Carolyn Cole haben deshalb längst die Präzisierung von KI-Modellen mittels spezialisierter KI-Modelle gefordert. Spezialisierte KI-Modelle meint Modelle, die auf die Erkennung, Analyse und Erzeugung eines ganz spezifischen Themenbereichs oder eines spezifischen Datei-, Medien- oder Textformats ausgerichtet sind (Hajikhani et al. 2024, 737). Mittels der Gegenüberstellung eines allgemeinen KI-Modells mit einem spezialisierten KI-Modell wäre eine Feinabstimmung von KI-Modellen möglich, so die Idee der Forschenden. Die Förderung derartiger Ansätze stünde dem noch immer zu verzeichnenden Trend der Skalierung entgegen, ohne dabei auf Größeneffekte verzichten zu müssen.

*Förderung der Minimierung und Vermeidung von Fehlern*
Angesichts der Tatsache, dass die Funktionsweise, die Fehleranfälligkeit und internen Logiken von LLM noch zu wenig verstanden werden, gilt es, Forschungs- und Entwicklungsvorhaben zu fördern, die eben diese Desiderate beheben. Wir empfehlen hiermit Forschungs- und Entwicklungsvorhaben, die sich der Frage annehmen, wie KI mit allgemeinem Verwendungszweck ethisch besser werden kann. Denkbare Varianten sind die Beforschung der Zusammenstellung von LLM-Trainingsdatensätzen und in diesem Zuge die Bereitstellung von sowohl ethisch und datenschutzrechtlich einwandfreien als auch nützlichen und qualitativ hochwertigen Datensätzen. Eine andere Option sind Projekte, die bestehenden LLM-Output in ethischer Hinsicht verbessern. Ein entsprechendes Modell beziehungsweise System könnte im Sinne eines Abgasfilters zwischen (austauschbaren) LLM und Nutzenden eingesetzt werden. Der Output des LLM würde auf seine Ethik hin analysiert werden und Nutzer:innen könnten zusätzliche Information zuteilwerden, die sie in die Ergebnisverwendung, deren Kontextualisierung und Reflektion einbeziehen. Nicht zuletzt ist denkbar, dass die Datensätze von LLM, die keine ethischen Herausforderungen erzeugen, weiter zum Training genutzt werden, und andere konsequenterweise nicht.

## Ausblick

In diesem Artikel haben wir ethische Problemstellungen der aktuellen KI-Regulierung und der zu regulierenden Systeme theoretisch und anhand von Fallbeispielen aus dem Bereich von KI mit allgemeinem Verwendungszweck vorgestellt und Handlungsempfehlungen für die zukünftige Förderung entwickelt. Der Bereich der LLM ist im Augenblick sehr dynamisch. Gleichzeitig wird immer deutlicher, dass es sich empfiehlt, den Trend um LLM und ihren Neuheitswert nicht zu Ungunsten begründeter sozialer und rechtlicher Annahmen, ethischer Grundsätze sowie elementarer Normen und gesellschaftlicher Regeln zu vernachlässigen. Die Förderung von Forschungsaktivitäten, -arbeiten und -kooperationen zur wissensbasierten Klärung der oben dargelegten Herausforderungen ist ein wesentliches Instrument der erfolgreichen und verträglichen Regulierung von KI.

## Literatur


Bender, Emily M., Timnit Gebru, Angelina McMillan-Major und Margaret Mitchell. 2021. „On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Conference on Fairness, Accountability, and Transparency* (*FAccT '21*), 3-10.03.2021, Canada.

Bender, Emily M. und Alexander Koller. 2020. „Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Bögeholz, Susanne, Corinna Hößle, Dietmar Höttecke und Jürgen Menthe. 2018. „Bewertungskompetenz". In *Theorien in der naturwissenschaftsdidaktischen Forschung*. Hrsg. v. Dirk Krüger, Ilka Parchmann und Horst Schecker. Springer: Berlin, Heidelberg.

Bolón-Canedo, Verónica, Laura Morán-Fernández, Brais Cancela und Amparo Alonso-Betanzos. 2024. „A review of green artificial intelligence: Towards a more sustainable future". *Neurocomputing* 599.

Denson, Shane. 2016. „Speculation, Transition, and the Passing of Post-Cinema." *Cinéma & Cie: International Film Studies Journal,* 26-27. In „Post What? Post When? Thinking Moving Images beyond the Post-Medium/Post-Cinema Condition." Hrsg. v. Vinzenz Hediger und Miriam De Rosa.

Hajikhani, Arash und Caroyln Cole. 2024. „A critical review of large language models: Sensitivity, bias, and the path toward specialized AI." *Quantitative Science Studies* 5 (3).

Henning, Maximilian. 2024. „Die eigentlichen Regeln für ChatGPT kommen noch." 03.11.2024 *netzpolitik.org*. URL: https://netzpolitik.org/2024/europaeische-ki-verordnung-die-eigentlichen-regeln-fuer-chatgpt-kommen-noch/.

KI-Verordnung. 2024. Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008, (EU) Nr. 167/2013, (EU) Nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 und (EU) 2019/2144 sowie der Richtlinien 2014/90/EU, (EU) 2016/797 und (EU) 2020/1828. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Kühl, Eike. 2023. „KI-Regulierung: Wie Google und Microsoft Stimmung gegen den AI Act machen". 01.03.2023. *heise.de*.

Kurz, Constanze und Stefan Ullrich. 2025. „Dürfen wir Ihre Informationen durch unserer KI jagen?" 17.02.2025 netzpolitik.org. URL: https://netzpolitik.org/2025/privatsphaere-duerfen-wir-ihre-informationen-durch-unsere-ki-jagen/.

Maham, Pegah und Sabrina Küspert. 2023. „Governing General Purpose AI. A Comprehensive Map of Unreliability, Misuse and Systemic Risks." *Stiftung Neue Verantwortung*.

Nassehi, Armin. 2019. *Muster. Theorie der digitalen Gesellschaft*. C. H. Beck: München.

Pasquinelli, Matteo. 2024. *Das Auge des Meisters. Eine Sozialgeschichte Künstlicher Intelligenz*. Unrast: Münster.

Ren, Shaolei und Adam Wierman. 2024. „The Uneven Distribution of AI's Environmental Costs. Harvard Business Review." 15.07.2024. URL: https://hbr.org/2024/07/the-uneven-distribution-of-ais-environmental-impacts.

Rubinic, Igor, Marija Kurtov, Ivan Rubinic, Likic Robert, Paul I. Dargan und David M. Wood. 2024. „Artificial intelligence in clinical pharmacology: A case study and scoping review of large language models and bioweapon potential." *Br J Clin Pharmacol.* 2024. 90 (3): 620-628.

Samsi, Siddharth, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michael, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari und Vijay Gadepally. 2023. „From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference." *IEEE High Performance Extreme Computing Conference (HPEC).* Conference Paper. arXiv.

Sauerwein, Tessa. 2024 „#networkframework – Next Generation Informationskompetenz für Forschende". *Bibliotheksdienst* 58 (12).

Seemann, Michael. 2023. „Künstliche Intelligenz, Large Language Models, ChatGPT und die Arbeitswelt der Zukunft". Working Paper Forschungsförderung 304, Hans-Böckler-Stiftung: Düsseldorf.

Tao, Yan, Olga Viberg, Ryan S. Baker und René F. Kizilcec. 2024. „Cultural bias and cultural alignment of large language models." *PNAS Nexus* 3 (9).

**Herausgeber**

Prof. Dr. Volker Wittpahl
Institut für Innovation und Technik (iit)
in der VDI/VDE Innovation + Technik GmbH
Steinplatz 1, 10623 Berlin

**Zitation**

Puchta, Rebecca und Wilhelm-Weidner, Arno (2025). Ethik und Auswirkungen von KI mit minimalem Risiko – Empfehlungen für die Förderung ausgehend von LLM. iit-perspektive Nr. 75. Institut für Innovation und Technik (iit), Berlin.

**Autor:innen**

Rebecca Puchta
Tel.: +49 (0) 30 310078 3266
E-Mail: rebecca.puchta@vdivde-it.de

Dr. Arno Wilhelm-Weidner
Tel.: +49 (0) 30 310078 5866
E-Mail: wilhelm-weidner@iit-berlin.de

**iit perspektive Nr. 75**

Februar 2025
Layout: Poli Quintana
DOI: 10.23776/2025_01
Bildnachweise: Fina – stock.adobe.com (Papagei), Rawpixel.com – stock.adobe.com (Netz)

DOI: 10.23776/2025_01