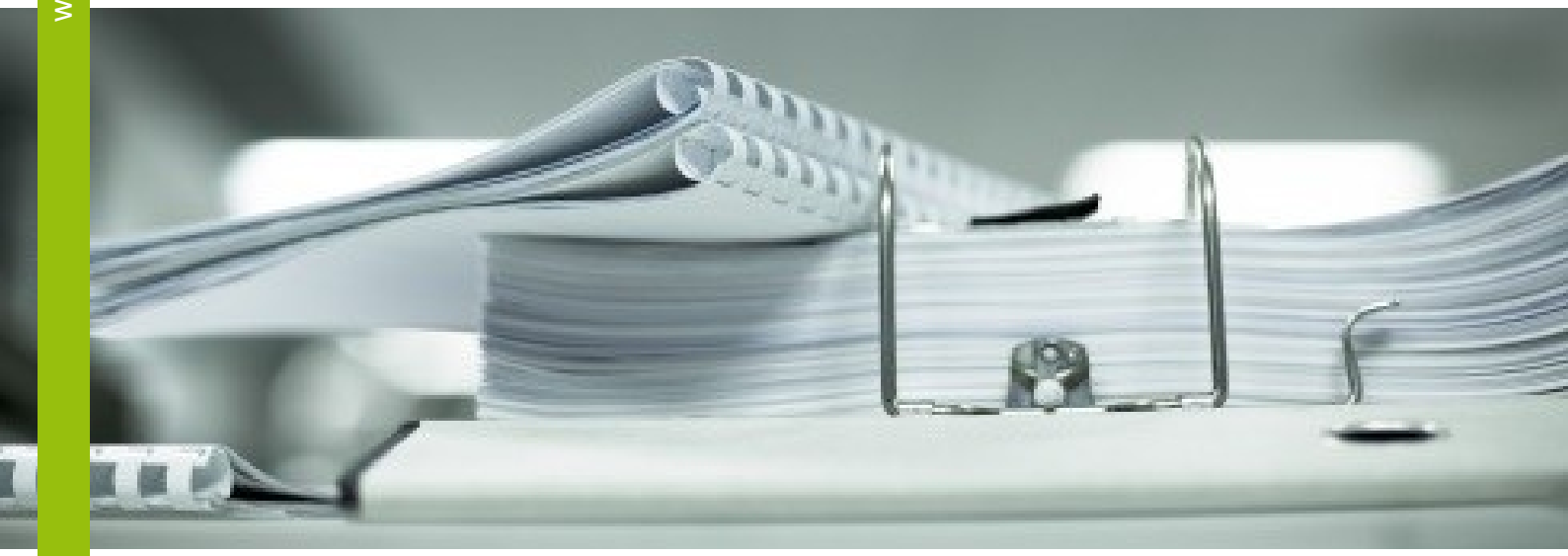
Policy Brief | 17.09.2025



Bessere Spielregeln für digitale Berater

Chatbots als universelle Experten oder als „epistemisches Risiko“ für die Nachhaltigkeitstransformation?

// Dr. Peter Gailhofer[1]

LLM-basierte Chatbots dringen in viele Gesellschaftsbereiche vor – auch in umweltpolitisch relevanten Kontexten. Sie versprechen, komplexe Fragen evidenzbasiert zu beantworten und damit nachhaltigere Entscheidungen zu ermöglichen. Gleichzeitig bergen sie Risiken: Die zugrundeliegenden Sprachmodelle sind **überzeugend, aber unzuverlässig**. Sie vermitteln falsche oder verzerrte Informationen ebenso eloquent wie gesicherte Evidenz. Gerade im Umweltbereich, wo Entscheidungen existenzielle Folgen haben können, ist dies problematisch.

Das Policy Paper fopkussiert zwei zentrale Problemkomplexe:

- **Simuliertes Wissen:** Chatbots wirken wie sachkundige Experten, verfügen aber nicht über echte „epistemische“ Fähigkeiten. Sie reproduzieren verbreitete, aber falsche Annahmen und präsentieren sie mit hoher sprachlicher Überzeugungskraft. So entsteht „Bullshit“ im Sinne von Harry Frankfurt: Inhalte, die auf Überzeugung statt Wahrheitsgehalt zielen.

- **Ausblendung normativer Konflikte und Prioritäten:** LLMs verschieben normative Spielräume, indem sie Wertfragen als scheinbar neutrale Fakten präsentieren. Es entsteht eine **„epistemische Fassade“**, die Konflikte unsichtbar macht und Verfahren zur transparenten Abwägung und Deliberation zu unterlaufen droht.

Beide Problemkomplexe können – insbesondere, wenn die Systeme in sensiblen Bereichen, etwa in der Verwaltung eingesetzt werden – zu schwerwiegenden sozialen und ökologischen Folgen führen. Um sie zu lösen, ist die **epistemische Vertrauenswürdigkeit** der Systeme sicherzustellen. Das Papier definiert diese als die Fähigkeit eines Systems, Informationen

[1] Bei der Erstellung dieses Policy Papers wurden KI-basierte Chatbots unterstützend eingesetzt – sowohl für Rechercheaufgaben (einschlägige Literatur und Methodenbeschreibungen) als auch für redaktionelle Zwecke (Zusammenfassungen, Formulierungsvorschläge) sowie für die Erstellung einzelner Inhalte (z. B. Tabellen).

bereitzustellen, die sachlich belastbar, überprüfbar und für den jeweiligen Kontext angemessen sind – und dabei sowohl faktische Unsicherheiten als auch zugrunde liegende Wertannahmen offenlegen, anstatt sie zu verdecken.

Die Ursachen fehlender Vertrauenswürdigkeit sind **sozio-technischer Natur**: verzerrte Trainingsdaten, Designentscheidungen der Anbieter, ökonomische Anreize, algorithmische Verstärkungen und die Interaktion mit Nutzer*innen. Daraus lässt sich folgern, dass denkbare Lösungsansätze ebenfalls sozio-technisch sein sollten – also technische, organisatorische und rechtliche Elemente kombinieren.

Das Papier identifiziert vier übergeordnete Hebel zur Stärkung der epistemischen Vertrauenswürdigkeit. Diese erfordern die kontinuierliche Einbindung wissenschaftlicher und fachlicher Expertise als integralen Bestandteil von Entwicklung, Anwendung und Governance der Systeme:

- **Transparenz & Erklärbarkeit:** Die Anbieter der Systeme sollten ihre Quellen, Unsicherheiten und Entscheidungswege nachvollziehbar offenlegen – zumindest für die wissenschaftliche Analyse.
- **Datenqualität & Evidenzbindung:** Die Zuverlässigkeit von Antworten hängt maßgeblich von der Auswahl, Aufbereitung und Validierung der zugrunde liegenden Daten ab. Verfahren wie Retrieval-Augmented Generation (RAG) und die Kuration domänenspezifischer Datenräume verdeutlichen die zentrale Rolle wissenschaftlicher Expertise für die Sicherung belastbarer Evidenz.
- **Menschliche Expertise im Prozess:** Verfahren wie Human-in-the-Loop, Reinforcement Learning from Human Feedback (RLHF) oder Supervised Fine-Tuning binden Fachwissen gezielt in die Entwicklung und Anwendung von Modellen ein. Damit wird Expertise nicht nur ergänzend, sondern strukturell in die Systeme integriert.
- **Evaluation & Governance:** Benchmarks, Leaderboards, Peer-Review und partizipative Bewertungsverfahren schaffen Rahmenbedingungen für eine kontinuierliche Qualitätskontrolle. Ihre Wirksamkeit hängt wesentlich von institutionalisierter Expertise und fachlicher Mitgestaltung ab.

Der geltende Rechtsrahmen (etwa die KI-Verordnung, die Verordnung über digitale Dienste, das Lauterkeits- und Äußerungsrecht) bietet erste Ansätze, weist aber erhebliche Lücken auf. Der rechtliche Rahmen stellt daher noch nicht sicher, dass die Betreiber von Chatbots die Maßnahmen treffen, mit denen die erheblichen Risiken fehlender epistemischer Vertrauenswürdigkeit gemindert – und umgekehrt Potenziale evidenzgestützter „universeller Berater“ gehoben werden könnten.

Zentrale Handlungsempfehlungen

- Die Untersuchung empfiehlt vor diesem Hintergrund die Prüfung **rechtlicher Nachschärfungen**, etwa durch sektorale Präzisierungen des AI Act.
- Dabei könnte – analog zur Umweltverträglichkeitsprüfung (UVP) – eine **„Epistemische Verträglichkeitsprüfung" (EVP)** insbesondere für Anwendungen vorgeschrieben werden, die spezialisiert in besonders sensiblen Bereichen eingesetzt werden sollen.
- Sektorale **Standards für Datenqualität und Unsicherheitskommunikation** in umweltrelevanten Anwendungen könnten gesetzliche Regelungen im Umwelt- und Planungsrecht ergänzen,

Daneben könnte der Rechtsrahmen ergänzt werden durch

- den Aufbau **institutionalisierter Mechnismen und Verfahren zur Evaluation** (Benchmarks, Audits, Peer-Review),
- sowie durch die Schaffung von **Organisationspflichten** für menschliche Entscheidungsträger.

Das Ziel solcher Instrumente ist nicht die Etablierung einer zentralen „Wahrheitsinstanz", sondern **(insbesondere) durch Verfahrensvorgaben zu belastbaren Outputs beizutragen**.

1 Einleitung

KI-Anwendungen, insbesondere solche, die auf großen Sprachmodellen basieren, durchdringen alle Gesellschaftsbereiche und versprechen dabei revolutionäre Veränderungen. Vor allem KI-basierte Chatbots[2] werden zunehmend zu Ratgebern für alle denkbaren Entscheidungen, sollen uns das gesamte Weltwissen erschließen und dabei auch noch unsere individuellen Präferenzen berücksichtigen[3]. Aus ökologischer Sicht bietet diese Entwicklung Chancen wie Risiken. Das Versprechen der neuen Tools ist tatsächlich verlockend: Gerade in Zeiten, in denen wissenschaftliche Evidenz unter Beschuss steht und Desinformation alltäglich geworden ist, wirkt die Vorstellung technisch vermittelten, evidenzbasierten und objektiven Handlungswissens überaus attraktiv. Aus ökologischer Perspektive scheinen sich dafür enorme Möglichkeiten zu eröffnen: Ob als automatisierte und hyperbeschleunigte Berater für die klimaangepasste Stadtplanung, für die sozial-ökologische Optimierung von Lieferketten oder als freundliche Ratgeber für nachhaltigen Konsum – KI-basierte Systeme könnten uns dabei helfen, komplexe Umweltfragen besser zu verstehen, evidenzbasiert nachhaltigere Entscheidungen zu treffen und damit die ökologische Transformation zu beschleunigen.

Doch wie verlässlich sind diese digitalen Berater tatsächlich? Im letzten Spendenprojekt des Öko-Instituts[4] wurde die Faktentreue von KI-Sprachmodellen zu vier wichtigen Nachhaltigkeitsthemen geprüft. Die zwiespältigen Ergebnisse dieser Prüfung werfen grundlegende Fragen auf: Selbst bei eindeutig formulierten, eher naturwissenschaftlich ausgerichteten Fragen wurden erhebliche Ungenauigkeiten und fragwürdige Bewertungen festgestellt, die Informationen der Bots waren häufig einseitig oder lückenhaft, wobei dies teilweise nicht leicht erkennbar war. Dahinter stecken grundsätzliche Probleme: KI-Systeme sind „persuasive“– also hochgradig überzeugende[5] – aber unzuverlässige und vielfach voreingenommene „Experten“. Die Folgen von sehr überzeugenden, aber nur vermeintlich richtigen Ratschlägen könnten gerade im Umweltsektor, in dem evidenzbasierte Entscheidungen eine existenzielle Bedeutung haben, gravierend sein.

Solche Risiken können gemindert werden. Wirksame Maßnahmen hierfür brauchen allerdings ein klares Bild vom zu lösenden Problem: Aufbauend auf den Diagnosen im Spendenprojekt sollen deshalb im Folgenden zunächst die „epistemischen“[6] Risiken – solche Risiken also, die aus der disruptiven Wirkung der neuen Technologien *auf unsere Erkenntnis und unser Handlungswissen folgen* – erläutert werden, um dann über entsprechende Lösungsansätze aus Umweltsicht nachzudenken. Ein Ergebnis dieser Überlegungen sei vorweggenommen: bei der Lösung dieser Probleme wird menschliche Expertise absehbar eine wesentliche Rolle spielen.

---

[2] Chatbots sind computerbasierte Systeme, mit dem die Nutzenden über eine Onlineschnittstelle in natürlicher Sprache kommunizieren. Chatbots können Anfragen dank sprachverarbeitender Technologie beantworten, dafür nutzen sie vermehrt KI-Systeme, Definition in Anlehnung an Albrecht (2023), ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen, TAB Hintergrundpapier Nr. 26, online verfügbar unter: https://publikationen.bibliothek.kit.edu/1000158070/150614893

[3] S. https://inflection.ai/blog/why-create-personal-ai

[4] Siehe https://www.oeko.de/blog/fakt-ist-sprachliche-unsicherheiten-der-kuenstlichen-intelligenz/..

[5] Z. Begriff s. https://de.wikipedia.org/wiki/Persuasive_Kommunikation

[6] Z. Begriff s. https://de.wiktionary.org/wiki/epistemisch.

1.1 Simuliertes Wissen: Wenn KI überzeugend spricht, aber nichts versteht

Die wissenschaftliche Auseinandersetzung mit den Fähigkeiten und Grenzen großer Sprachmodelle ist noch am Anfang, hat sich in den vergangenen Jahren aber deutlich vertieft. Die Vielfalt der Risiken der neuen „Wahrheitsmaschinen“[7] für Informationsvermittlung, Kommunikation und politische Diskurse, die sich dabei zeigt,[8] kann hier nicht abgebildet werden. Die Darstellung konzentriert sich stattdessen auf zwei übergeordnete Problemkomplexe, die mit der hohen Überzeugungskraft der sprachmodellbasierten ChatBots zu tun haben und aus Umweltsicht besonders relevant erscheinen.

Der erste Problemkomplex betrifft die systematische Unzuverlässigkeit bei der Wiedergabe von Tatsachen – also die Faktentreue der Systeme. Wie Sandra Wachter und Kollegen in einem jüngeren Paper formulieren,[9] sind große Sprachmodelle bestenfalls "incidental truthtellers" – vermeintlich wahre Aussagen entstehen ganz vorrangig als gleichsam zufällige Folge von Wahrscheinlichkeitsverteilungen in den Daten, die zum Training der Systeme genutzt werden. Die Systeme verfügen zugleich nicht über eine irgendwie geartete Expertise und die „epistemischen“ Fähigkeiten – zur Bewertung von Plausibilität, Schlüssigkeit, Konsistenz einer Information –, die nötig wären, um wahre von falschen Aussagen unterscheiden zu können. Menschliches Feedback beim Training der Modelle kann hier grobe Fehler vermeiden, neue Methoden[10] die Fehleranfälligkeit – wenn auch in umstrittenem Ausmaß –[11] verringern. In der Kombination mit typisch menschlichen “Biases” ihrer Nutzer*innen – wir neigen dazu, die Systeme zu vermenschlichen und sprachlichen Äußerungen Bedeutung und Intentionalität zu unterstellen – führen solche Probleme aber, wie Wachter et. al.[12] es fassen dazu, dass LLMs vielfach „homogenisiertes, stark vereinfachtes und nicht repräsentatives Wissen“ in großem Maßstab verbreiten.

Die umweltpolitische Relevanz solcher Probleme kann beispielhaft mit dem typischen Problem des, sogenannten "common token bias"[13] verdeutlicht werden: Weitverbreitete Irrtümer in den Quelldaten werden mit derselben Überzeugungskraft reproduziert wie wissenschaftlich gesicherte Erkenntnisse. Ein anschauliches Beispiel aus dem Bereich der Umweltpolitik: Während korrekte Angaben zum Energieverbrauch von Wärmepumpen statistisch häufiger in technischen Dokumentationen oder wissenschaftlichen Analysen zu finden sind, können sich gleichzeitig populäre Mythen über deren Ineffizienz hartnäckig in allgemeineren Quellen halten – ohne entsprechende Feinabstimmung könnten beide Informationstypen undifferenziert in die Modelle

[7] Munn, Magee & Arora (2023), Truth machines: synthesizing veracity in AI language models, *AI & Soc* **39**, 2759–2773 (2024). https://doi.org/10.1007/s00146-023-01756-4

[8] Für eine umfassende Darstellung s. etwa Huang et. al. (2024), TrustLLM: Trustworthiness in Large Language Models, arXiv:2401.05561v6 [cs.CL] 30 Sep 2024.

[9] Wachter, Mittelstadt & Russell (2024) Do large language models have a legal duty to tell the truth? R. Soc. Open Sci.11240197, http://doi.org/10.1098/rsos.240197

[10] Vgl. etwa Lightman et.al. (2023), Let`s verify step by step, arXiv:2305.20050v1 [cs.LG] 31 May 2023. Für GPT-5 verzeichnet OpenAI deutlich geringere Halluzinationsraten, z. B. in Benchmark-Vergleichen. Dennoch bleibt das Halluzinationsphänomen strukturell bestehen und gerade in fachlich komplexen oder datenarmen Kontexten durchaus prominent. Die Reduktion der Fehlerquote mildert also Symptome, beseitigt aber nicht die epistemischen Risiken, denen unser Policy Paper nachgeht, s. https://www.nature.com/articles/d41586-025-02853-8, s.a. https://opentools.ai/news/open-ais-new-findings-cracking-the-code-on-ai-hallucinations-with-gpt-5.

[11] Vgl. Shojaee et.al. (2025), The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity, online verfügbar unter https://machinelearning.apple.com/research/illusion-of-thinking

[12] Wachter, Mittelstadt & Russell (2024) Do large language models have a legal duty to tell the truth? R. Soc. Open Sci.11240197, http://doi.org/10.1098/rsos.240197

[13] S. hierzu etwa Mina et.al. (2025), Cognitive Biases, Task Complexity, and Result Interpretability in Large Language Models, online verfügbar unter: https://aclanthology.org/2025.coling-main.120.pdf.

einfließen. Im schlechtesten Fall befinden sich fundierte Informationen, z.B. in Fachartikeln, hinter einer Bezahlschranke und sind für Training und Betrieb der Modelle nicht zugänglich – die Systeme könnten ihre Antworten stattdessen aus tendenziösen Medieninhalten oder sozialen Netzwerken herleiten.

Besonders problematisch wird dies wegen einer Eigenschaft, die den Systemen den wenig schmeichelhaften Titel hervorragender "Bullshitter" eingebracht hat – in Anlehnung an den Philosophen Harry Frankfurt,[14] der mit dem Begriff „Bullshit“ eine Form der Kommunikation bezeichnete, die sich ausschließlich um die persuasive[15] Wirkung und nicht um ihre inhaltliche Richtigkeit kümmert. KI-basierte Sprachassistenten wirken häufig eloquent und überzeugend, präsentieren „jede Antwort mit unerschütterlicher Zuversicht, ähnlich wie ein Experte, der aus dem Stegreif eine Erklärung abgibt“,[16] selbst wenn ihre Aussagen wenig oder gar nichts mit wissenschaftlicher Evidenz zu tun haben. Eine solche von der inhaltlichen Richtigkeit losgelöste Überzeugungskraft ist hochriskant, vor allem eben, wenn sie gesellschaftlich relevante Entscheidungsprozesse durch Chatbots millionenfach skaliert, personalisiert und hochgradig beschleunigt auf falschen Tatsachengrundlagen informiert.

1.2 „Bullshit“ statt verantwortlicher Entscheidung: Verschiebt KI normative Spielräume?

KI-Systeme wie große Sprachmodelle erscheinen auf den ersten Blick als neutrale Werkzeuge: Sie analysieren Informationen, fassen Daten zusammen, formulieren Empfehlungen. Doch bei näherer Betrachtung zeigt sich, dass die Systeme unausweichlich Werturteile[17] treffen – etwa darüber, welche Perspektiven sie berücksichtigen, welche Unsicherheiten sie betonen oder ausblenden, und wie sie konkurrierende Interessen gewichten. Fakten und Wertentscheidungen sind oft eng verflochten – auch in der Umweltpolitik: Klimaprojektionen, hydrologische Modelle und toxikologische Bewertungen sind nicht deterministisch, sondern mit struktureller Unsicherheit behaftet – sei es durch Szenarien, Eingangsdaten oder Extrapolationen. Diese Unsicherheiten sind nicht eliminierbar, sondern müssen sichtbar gemacht und politisch verhandelt werden. Werturteile sind in vielen Anwendungsbereichen nicht nur unvermeidbar, sondern hochgradig folgenreich – weil solche impliziten Wertannahmen darüber entscheiden, welche Interessen Priorität erhalten**,** welche Risiken und Kosten wir in Kauf nehmen**,** wessen Perspektiven Gehör finden. Entscheidungen beruhen eben nicht nur auf Fakten, sondern auch auf oft unausgesprochenen Überzeugungen darüber, was als relevant, fair oder zumutbar gilt.

Solche Wertungen und Priorisierungen sind in gesellschaftlich und politisch relevanten Bereichen institutionell eingehegt – etwa in der Umweltverwaltung. Hier gibt es bewährte Verfahren zur Verarbeitung von Unsicherheit und Wertkonflikten: Abwägungsregeln, Begründungspflichten, transparente Entscheidungsprozesse, Rechtsschutzmöglichkeiten. Wenn ein LLM in einer Umweltbehörde etwa hilft zu beurteilen, ob ein Industrieprojekt *zumutbare* Umweltbelastungen verursacht oder ob ausreichende wissenschaftliche Evidenz für die Zulassung einer Chemikalie vorliegt, dann

[14] Vgl. https://de.wikipedia.org/wiki/On_Bullshit.
[15] Zum Begriff der persuasiven Kommunikation s. https://de.wikipedia.org/wiki/Persuasive_Kommunikation.
[16] Munn, Magee & Arora (2023), Truth machines: synthesizing veracity in AI language models, *AI & Soc* **39**, 2759–2773 (2024). https://doi.org/10.1007/s00146-023-01756-4
[17] Zu den normativen Priorisierungen von LLMs vgl. Liu, Liu & Yu (2025), What’s the most important value? INVP: INvestigating the Value Priorities of LLMs through Decision-making in Social Scenarios, Proceedings of the 31st International Conference on Computational Linguistics, pages 4725–4752 January 19–24, 2025

bewegt es sich **nicht nur im Bereich von technischen und tatsachenbezogenen Aspekten,** sondern **im Zentrum solcher Wertungsfragen**. Für deren Beantwortung übernehmen die Entscheidenden, in „analogen“ Verfahren Verantwortung. Sie können gegebenenfalls in Gerichtsverfahren, jedenfalls aber in gesellschaftlichen und professionellen Debatten diskutiert und revidiert werden.

Genau hier liegt ein Risiko beim Einsatz von „persuasiven“ Chatbots: Denn diese können, etwa bei ihrem Einsatz in der Umweltverwaltung, den Eindruck erwecken, es gäbe auf komplexe Fragen klare, neutrale Antworten – selbst dann, wenn unterschiedliche ethische Bewertungen naheliegend und notwendig wären. So entsteht, wie Silvie Delacroix es in einem jüngeren rechtswissenschaftlichen Papier formuliert,[18] eine Art **"epistemische Fassade"**: Ein Anschein von Sicherheit, der möglicherweise umstrittene Wertfragen unsichtbar macht. Das kann die offenen Verfahren und Diskurse untergraben, die für Entscheidungen im Umweltbereich – und viele andere gesellschaftlich bedeutsame Bereiche – unverzichtbar sind. Es reicht deshalb nicht, LLMs einfach nur technisch genauer oder sicherer zu machen. Sie müssen so gestaltet sein, dass sie Unklarheiten, kollidierende Interessen und Wertekonflikte nicht verdecken, sondern sichtbar machen und dadurch einer Diskussion zugänglich stellen. Delacroix betont, dass es dafür neue Schnittstellen zwischen Mensch und Maschine braucht, die nicht Eindeutigkeit erzwingen, sondern Raum für ethische Mehrdeutigkeit, Diskussion und Interpretation lassen – und damit eine Grundlage für verantwortliche Entscheidungen bilden.

Die Analyse des Problems greift also zu kurz, wenn sie mögliche Auswirkungen der Systeme auf Wertentscheidungen ignoriert. Um beide Aspekte zu berücksichtigen, kann man das übergeordnete Problem als eines der „epistemischen“ Vertrauenswürdigkeit der Systeme bezeichnen[19]: Darunter ist dann die Fähigkeit eines Systems zu verstehen, belastbare, überprüfbare und für den jeweiligen Kontext angemessene Beiträge zu liefern – und dabei weder faktische Unsicherheiten, noch Wertannahmen zu kaschieren, sondern sichtbar zu halten und für Diskurse zu öffnen.

2 Sozio-technische Lösungen für ein sozio-technisches Problem

Um solche Risiken zu mindern und die transformativen Potenziale vertrauenswürdiger KI-Systeme zu heben, ist ein **Verständnis der Ursachen** systemischen „bullshittings“[20] (im philosophischen Sinne) notwendig. Um zu einem solchen Problemverständnis zu kommen, muss man sich aber erneut mit einer gewissen Komplexität herumschlagen. Denn die Herausforderungen für vertrauenswürdige Systeme sind nicht nur technischer Natur, sondern das Ergebnis vielschichtiger soziotechnischer Wechselwirkungen.

Dazu gehören bekanntermaßen die Auswahl und Struktur der Daten, mit denen die Modelle trainiert werden. Wie erwähnt, schöpfen LLMs aus gigantischen Korpora, die womöglich unkuratiert aus dem Internet extrahiert wurden – mit allen darin enthaltenen Verzerrungen: Einseitigen Narrativen, stereotypen Rollenmustern,

[18] Delacroix (2025), Moral Perception and Uncertainty Expression in LLM-Augmented Judicial Practice (March 14, 2025), Minds and Machines, Forthcoming, online zugänglich unter: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4787044.

[19] Vgl. Heersmink, de Rooij, Clavel Vázquez & Colombo (2024), A phenomenology and epistemology of large language models: transparency, trust, and trustworthiness. Ethics Inf Technol 26, 41 (2024). https://doi.org/10.1007/s10676-024-09777-3

[20] Zum Begriff im Zusammenhang mit KI Chatbots s. Townsen Hicks, Humphries & Slater (2024), ChatGPT is bullshit. Ethics Inf Technol 26, 38 (2024). https://doi.org/10.1007/s10676-024-09775-5

wissenschaftlich anmutenden Halbwahrheiten oder ideologischen Voreingenommenheiten. Ein aktuelles Papier beobachtet gar, dass sich in LLMs auf der Basis solcher Daten eigenständige Wertesysteme herausbilden, die sich über die Systeme hinweg immer mehr ähneln und durchaus ethisch problematisch, jedenfalls nicht ganz leicht erkennbar sein können – und, wie gerade beschrieben, auch vermeintlich sachliche Antworten beeinflussen könnten.[21]

Doch das ist nur ein Teil des Problems: Verzerrte, oder falsche Aussagen können auf subjektive Designentscheidungen von Entwickler*innen oder Anbietern der Chatbots zurückgehen. Der Kampf um Marktanteile zwischen unterschiedlichen Anbietern von LLMs führt dazu, dass den Systemen letztlich wirtschaftlich motivierte Zielgrößen vorgegeben und die Systemoutputs z.B. auf Nutzerbindung statt auf Wahrheitsgehalt optimiert werden. Hinzu kommen algorithmische Verzerrungen, die bestimmte inhaltliche Tendenzen verstärken oder abschwächen können – und nicht zuletzt: das Verhalten der Nutzer*innen selbst. Denn auch die Nutzer*innen prägen die Modelle mit – ob über Feedbackmechanismen, die bloße Wiederholung bestimmter Fragen, oder die Art und Weise, wie sie ihre prompts formulieren. Modelle haben häufig Probleme, Fehler in den Eingaben ihrer Nutzer zu erkennen und lernen, das zu sagen, was „gut ankommt“, statt das, was gut belegt ist – ein Phänomen, das in der Forschung als „sycophancy“ bezeichnet wird.[22] Durch solche strukturellen Probleme könnten subtile Fehleinschätzungen der Nutzer*innen verstärkt und verstetigt werden.

Die gute Nachricht: Das zunehmende Wissen um erforderliche „sozio-technische Leitplanken“ der Systeme ermöglicht auch gezielte Interventionen. Es identifiziert mit diesen „Leitplanken“ eben auch Möglichkeiten, deren epistemische Vertrauenswürdigkeit zu verbessern. Dabei sollte nicht auf einzelne – etwa auf Prompts oder Interfaces,[23] fokussierte – sondern auf einen *Mix* unterschiedlicher Instrumente gesetzt werden. Die nähere Betrachtung dieser Instrumente lässt wiederum darauf schließen, dass jüngere Warnungen vor einem „Ende der Wissensarbeit“[24] womöglich übertreiben: Wo es um Vertrauen geht, bleibt menschliche Expertise auf absehbare Zeit eine relevante Größe.

2.1 Hebel zur Stärkung epistemischer Vertrauenswürdigkeit

Eine zentrale Vorbedingung für wirksame Interventionen ist **Transparenz**. Nur wenn wir verstehen, wie KI-Systeme zu ihren Ergebnissen kommen, können wir sie auch gezielt verbessern. Eine solche Erklärbarkeit der Modelle und Anwendungen ist auch die Voraussetzung dafür, dass menschliche Entscheider – ob in Behörden oder anderswo – wirklich noch als verantwortliche Akteure begriffen werden können: Mark Coeckelbergh spricht aus philosophischer Perspektive von "epistemischer Handlungsfähigkeit" – der Möglichkeit, die eigenen Überzeugungen auf Basis nachvollziehbarer Gründe zu formen. Bleiben die inneren Prozesse von LLMs undurchsichtig, verlieren Nutzer*innen die Kontrolle über ihre Erkenntnisgrundlagen.[25] Die Folge wäre eine Art epistemischer Fremdbestimmung. Einen technischen Lösungsansatz bietet

[21] Mazeika et. al. (2025), Utility Engineering: Analyzing and Controlling Emergent Value Systems in Ais, online verfügbar unter https://arxiv.org/abs/2502.08640.
[22] S. hierzu etwa Huang et. al. (2024), TrustLLM: Trustworthiness in Large Language Models, arXiv:2401.05561v6 [cs.CL] 30 Sep 2024, S. 30.
[23] Im Kontext von KI-Chatbots bezeichnet ein Interface die Schnittstelle, über die Nutzer*innen mit dem KI-Modell interagieren und durch die das Modell in technische, soziale und organisatorische Umgebungen eingebettet wird.
[24] Holtel (2024), Droht das Ende der Experten?, Verlag Franz Vahlen 2024.
[25] Coeckelbergh (2025), LLMs, Truth, and Democracy: An Overview of Risks. Sci Eng Ethics. 2025 Jan 23;31(1):4. doi: 10.1007/s11948-025-00529-0. PMID: 39849172; PMCID: PMC11759458.

hier etwa die mechanistische Interpretierbarkeitsforschung:[26] Sie macht es möglich, interne Repräsentationen in Sprachmodellen zu identifizieren und damit Begründungszusammenhänge offenzulegen, statt nur Outputs zu bewerten. Entsprechende Erklärungsansätze könnten vor diesem Hintergrund etwa helfen, die Nutzer*innen für Risiken falscher Informationen und ihren eigenen Einfluss auf die Systemoutputs zu sensibilisieren.

Durch Transparenz allein werden mögliche Fehlerquellen der Systeme aber noch nicht beseitigt. Hierfür lassen sich drei weitere Hebel identifizieren: die Qualität der verwendeten Daten, die methodischen Verfahren zur Entwicklung und Weiterentwicklung der Systeme und Metriken, mit denen deren Leistung bewertet, gemessen und gesteuert wird. In allen drei Bereichen kann fachliche Expertise eine zentrale Rolle spielen: nicht nur in der Form einer punktuellen Kontrolle, sondern als integraler Bestandteil epistemisch verantwortlicher KI-Systeme.

2.1.1 Daten

Angesichts der Tatsache, dass KI-Modelle die unvermeidbaren Verzerrungen und Fehlinformationen ihrer Trainingsdaten reproduzieren, wird deutlich, wie zentral die Auswahl, Aufbereitung und Gewichtung der Daten für deren Vertrauenswürdigkeit ist. Die Lösung liegt nicht in der bloßen Menge der Daten, sondern in deren intelligenter und strukturierter Auswahl: Quellen müssen validiert, Unsicherheiten gekennzeichnet und unterschiedliche Perspektiven systematisch berücksichtigt werden. Diese Qualitätssicherung erfordert die Einschätzung durch Fachleute, die etwa bewerten können, welche Informationen als belastbare Evidenz gelten dürfen. In dieser Weise validierte Daten können nicht nur für das Training, sondern auch für den Betrieb relevant werden. Mit Verfahren wie *Retrieval-Augmented Generation* (RAG) lassen sich Modelle so ausstatten, dass sie bei der Entwicklung ihrer Antworten auf externe, verifizierbare Wissensquellen zurückgreifen.

Das „ClimSight“-Tool[27] illustriert, dass und wie solche Ansätze funktionieren können: Es soll präzise, nachvollziehbare und reproduzierbare Klimaanalysen „für jedermann“ bereitstellen und die Probleme mangelnder Verlässlichkeit und die fehlender Evidenzbasierung durch die Integration einer RAG-Architektur lösen, indem es validierte, domänenspezifische Datenbestände (wie Klimaprojektionen, lokale Bedingungen und wissenschaftliche Literatur) in die Ableitung seiner Antworten einbindet. Die im Projekt TRUSTLLM durchgeführte Untersuchung[28] zeigt unter Verweis auf wissenschaftlich kuratierte Datensätze wie *Climate FEVER*[29] oder *SciFact*[30] welche entscheidende Rolle solche von Expert*innen validierten Daten für die Verbesserung der Antwortqualität spielen können.

2.1.2 Methoden zur Entwicklung und Optimierung von LLM-Systemen

Neben der Datengrundlage sind auch die Methoden, die zum Training der Systeme eingesetzt werden, ausschlaggebend die Vertrauenswürdigkeit der Sprachmodelle.

[26] Lindsey et.al., On the Biology of a Large Language Model, Anthropic March 27, 2025, online verfügbar unter https://transformer-circuits.pub/2025/attribution-graphs/biology.html.
[27] S. Koldunov/Jung (2024), Local climate services for all, courtesy of large language models, *Commun Earth Environ* **5**, 13 (2024). https://doi.org/10.1038/s43247-023-01199-1.
[28] Huang et. al. (2024), TrustLLM: Trustworthiness in Large Language Models, arXiv:2401.05561v6 [cs.CL] 30 Sep 2024, S. 30.
[29] Vgl. Diggelmann et.al. (2020), CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims, online verfügbar unter: https://arxiv.org/abs/2012.00614.
[30] Vgl. Wadden t al., Fact or Fiction: Verifying Scientific Claims, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Auch hier kann menschliche Expertise eine wichtige Rolle spielen. Unter dem Überbegriff von *Human-in-the-Loop*-Ansätzen werden verschiedene Methoden beschrieben,[31] bei denen Menschen – insbesondere Fachexpert*innen – aktiv in Entwicklungs- und Optimierungsschritte eingebunden werden. Im Zusammenhang mit wichtigen Verfahren wie *Reinforcement Learning from Human Feedback* (RLHF) wird diese Rolle deutlich: Hier erhalten Modelle Rückmeldungen auf ihre Antworten und lernen, ihre Strategien auf dieser Basis zu verbessern. Dabei wird zunehmend auf die Notwendigkeit fachlich fundierter Rückmeldungen hingewiesen.[32] Insbesondere bei spezialisierten Anwendungen – etwa im Bereich Umwelt, Energie oder Planung – könnte die Integration von domänenspezifischem Expertenwissen über solche Verfahren zu deutlichen Qualitätsgewinnen führen.

Neben RLHF werden auch eine Vielzahl weiterer Ansätze vorgeschlagen, um die Modellausgaben an Standards auszurichten, etwa „Supervised Fine-Tuning*"*, bei dem Modelle gezielt anhand kuratierter Beispieldialoge weitertrainiert werden,[33] „Deliberatives Alignment",[34] das die Anwendung expliziter Domänenstandards durch regelgeleitetes Denken fördern, oder „Instruction Tuning",[35] das die Anpassung der Systeme an fachspezifische Aufgabenstellungen erleichtern soll.

Schließlich erfährt auch das sogenannte **Prompt Engineering**[36] erhebliche Beachtung. Dabei werden gezielt strukturierte, inhaltlich präzise und kontextbewusste Eingaben („Prompts") entwickelt, um LLM-Antworten systematisch zu verbessern. Besonders relevant kann dies zur Reduktion von Nutzer-vermittelten Fehlerquellen wie *Sycophancy* – also der unkritischen Bestätigung falscher Nutzerannahmen sein. Durch Prompts, die das Modell explizit auffordern, Unsicherheiten anzuzeigen oder multiple Perspektiven zu erwägen, kann die Genauigkeit, Relevanz und Verständlichkeit Chatbot-Antworten zumindest situativ verbessert werden.

2.1.3 Benchmarks

Ein dritter Ansatzpunkt liegt in der Ausgestaltung von sogenannten Benchmarks. Benchmarks bestehen aus Daten, Fragen oder Aufgaben, die bestimmte Fähigkeiten LLMs testen und Metriken zur Bewertung dieser Fähigkeiten liefern sollen. Anhand von Benchmarks lässt sich also die Qualität von Modellantworten systematisch bewerten – etwa durch Tests auf Konsistenz, Evidenz oder logische Stimmigkeit. Zwar deuten aktuelle Studien darauf hin, dass Benchmarks anfällig für systematische Mängel sind,[37] darunter Verzerrungen in der Datenerstellung, oder (kommerzielle) Fehlanreize. Dennoch können sie – wenn sorgfältig entwickelt – durchaus Einfluss auf die Entwicklungsrichtung der Technologie nehmen: Sie beeinflussen, welches Modell für

[31] Mosqueira-Rey, Hernández-Pereira, Alonso-Ríos *et al. (2023),* Human-in-the-loop machine learning: a state of the art. *Artif Intell Rev* **56**, 3005–3054 (2023). https://doi.org/10.1007/s10462-022-10246-w.

[32] Daniels-Koch/ Freedman (2022), The Expertise Problem: Learning from Specialized Feedback, arXiv:2211.06519v1 [cs.LG] 12 Nov 2022.

[33] Wang et.al. (2024), Simulated Task Oriented Dialogues for Developing Versatile Conversational Agents, Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part I Pages 157 – 172 https://doi.org/10.1007/978-3-031-56027-9_10

[34] Guan et.al. ((2025), Deliberative Alignment: Reasoning Enables Safer Language Models, rXiv:2412.16339v2 [cs.CL] 8 Jan 2025.

[35] Bergmann (2024), https://www.ibm.com/think/topics/instruction-tuning, online verfügbar unter: https://www.ibm.com/think/topics/instruction-tuning

[36] Barkley/ van der Merwe (2024), Investigating the Role of Prompting and External Tools in Hallucination Rates of Large Language Models, arXiv:2410.19385v1 [cs.CL] 25 Oct 2024.

[37] S. Eriksson et.al. (2025), Can we trust AI Benchmarks? An interdisciplinary review of current issues in AI evaluation, arXiv:2502.06559v2 [cs.AI] 25 May 2025

die betreffende Anwendung als „führend" gilt, welche Varianten für die weitere Entwicklung herangezogen werden und welche Eigenschaften überhaupt als erstrebenswert gelten. Noch deutlicher wird die Steuerungsfunktion von Metriken bei zunehmend autonomen Optimierungsverfahren wie *AlphaEvolve* von Google DeepMind: Dort konkurrieren verschiedene Modellvarianten direkt miteinander; die jeweils „erfolgreicheren" Varianten – gemessen an vordefinierten Benchmarks – werden weiterentwickelt, die anderen verworfen. Gerade wenn diese Optimierung automatisiert abläuft, bleibt die Gestaltung der Bewertungsmaßstäbe ein kritischer Hebel – denn diese bestimmen, was vom System unter „Erfolg" verstanden und optimiert wird. Das macht deutlich: Selbst in einem zunehmend automatisierten Entwicklungsprozess bleiben Menschen – insbesondere Fachleute – unverzichtbar, um vertrauenswürdige Maßstäbe zu setzen, die nicht nur die Übereinstimmung mit dem Stand der Wissenschaft, sondern auch ethische Ausrichtung und gesellschaftliche Relevanz berücksichtigen können.

2.1.4 Faktenbewertung im Diskurs: Partizipative Verfahren für vertrauenswürdige KI

Während heutige Benchmarks wie TruthfulQA[38] weiterhin auf vordefinierten, einmal zusammengestellten Datensätzen basieren und im Wesentlichen als statische Referenz dienen, enthalten einige bereits Elemente einer dynamischen Infrastruktur – so etwa „online Leaderboards“ mit kontinuierlichen Updates.[39] Ihr volles Potenzial als Qualitätssicherungsinstrumente erhielte diese Art Evaluierung jedoch erst durch eine breitere **institutionelle Verankerung**: etwa durch Gremien oder Communities, die regelmäßig die Faktentreue von Inhalten überprüfen, anpassen und kontextabhängige Domänenanforderungen einbringen. In dieser Perspektive könnten bewährte, partizipative Verfahren wie Peer Review oder eine kollaborative Governance durch die Fach-Community dazu dienen, solche Benchmarks in sozial eingebettete evidenzbasierte Qualitätskontrollen umzuwandeln.

Kollaborative Ansätze könnten jedoch auch noch einen Schritt weiter gehen. Wie Delacroix mit Blick auf KI-Systeme im Justizwesen darlegt, könnten erweiterte partizipative Ansätze das Problem der „epistemischen Vertrauenswürdigkeit", wie wir es oben beschrieben haben, umfassender adressieren. Diese könnten insbesondere in solchen praktischen Kontexten helfen, in denen eindeutige Wahrheiten schwer bestimmbar sind, bei komplexen gesellschaftlichen Fragen, bei denen Werturteile unvermeidbar sind – etwa darüber, welche Interessen Priorität erhalten, welche Risiken wir in Kauf nehmen und wessen Perspektiven Gehör finden. Partizipative Ansätze könnten darauf abzielen, sowohl faktische Unsicherheiten als auch Wertannahmen sichtbar zu halten und für demokratische Diskurse zu öffnen. Delacroix identifiziert vier Kernprinzipien solcher Verfahren: **Die Deliberation** über Unsicherheitsfragen innerhalb von Nutzer*innengemeinschaften, eine **iterative Verfeinerung** der Kommunikation über Unsicherheiten durch kollektives Feedback, einen **Multi-Stakeholder-Dialog** zur Einbeziehung verschiedener Perspektiven und die Entwicklung eines **institutionalisierten Gedächtnisses** für den kollektiven Umgang mit Unsicherheit. [40]

[38] Lin, Hilton & Adams (2023), TruthfulQA: Measuring How Models Mimic Human Falsehoods, online verfügbar unter https://arxiv.org/abs/2109.07958.

[39] So etwa “FACTS Grounding”, https://deepmind.google/discover/blog/facts-grounding-a-new-benchmark-for-evaluating-the-factuality-of-large-language-models/

[40] Delacroix (2025), Moral Perception and Uncertainty Expression in LLM-Augmented Judicial Practice (March 14, 2025), Minds and Machines, Forthcoming, online zugänglich unter: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4787044.

Ob sich solche Verfahren erfolgreich umsetzen lassen, bleibt freilich eine offene Frage.

Dass sich ein institutionalisiertes, partizipatives Feedback auch in umwelt- und klimapolitischen Kontexten sinnvoll denken lässt, illustriert folgendes Beispiel für den Einsatz von Chatbots im Bereich der Klimaanpassung[41]: Wenn ein Chatbot auf einen prompt eines Planers in einer Kommune hin beispielsweise Empfehlungen für Hochwasserschutzmaßnahmen zusammenstellt, wird er sich einerseits auf Fakten, Wechselwirkungen und Wahrscheinlichkeiten beziehen – zu erwartende Niederschlagsmengen, die Fließgeschwindigkeit von Gewässern, topographische Einflüsse –, andererseits aber auch Wertentscheidungen treffen – darüber, welche Stadtteile oder Gebäude prioritär geschützt werden sollen, wie stark ökologische Belange gegenüber Baukosten gewichtet werden oder welche Risiken als zumutbar gelten. Die Vertrauenswürdigkeit solcher Systemoutputs könnte durch ein wiederholtes, mehrstufiges Feedbackverfahren („dynamisches Review“) verbessert werden: Das System könnte zunächst einen Empfehlungsentwurf generieren. Problematische Aspekte – eine unsichere Datenbasis, konkurrierende Bewertungen, notwendige Priorisierungen oder Wertentscheidungen – müssten dann identifiziert und markiert werden, möglicherweise durch geschulte Anwender*innen oder hybride Ansätze, die etwa „LLM-as-a-Judge"-Verfahren[42] mit menschlicher Überprüfung kombinieren. Anschließend könnten Ansätze entwickelt werden, um die Fachcommunity in die Bewertung und Optimierung der Outputs einzubeziehen – etwa bei der Einschätzung der wissenschaftlichen Evidenzlage. Parallel könnten Verfahren etabliert werden, die juristische Expertise in die Analyse rechtlicher Rahmenbedingungen und Entscheidungsspielräume einbinden. Das System würde diese Rückmeldungen integrieren und eine überarbeitete Version erstellen Schließlich können auch lokale Akteure beteiligt werden, indem sie unterschiedliche Handlungsoptionen anhand ihrer Präferenzen und Prioritäten bewerten. Im Zuge einer solchen mehrstufigen Validierung entstünden nicht einzelne „optimale" Empfehlungen, sondern transparent bewertete Handlungsoptionen mit expliziten Hinweisen auf ihre fachlichen Grundlagen, rechtlichen Grenzen und normativen Implikationen.

Erste Anwendungen[43] zeigen, wie solche Validierungsverfahren technisch umgesetzt werden könnten. Andere Konzepte[44] illustrieren, dass KI-Systeme auch Diskurs und Konsens innerhalb schwieriger Fragen erleichtern könnten, etwa durch die Identifikation von Interessenkonflikten, die Zusammenfassung verschiedener fachlicher Perspektiven und Expertisen oder die Moderation strukturierter Diskussionen. Bislang lässt sich nicht seriös bewerten, mit welchem Ergebnis, welchen neuen Risiken und mit welchem Aufwand solche Anwendungsperspektiven verbunden wären. Angesichts der ohnehin zunehmenden Nutzung der Systeme ist es aber umso wichtiger, solche Ansätze kritisch zu beforschen und zu erproben.

[41] Zu Einsatzmöglichkeiten und -fähigkeiten von Chatbots im Kontext von Klimawandel und Klimaanpassung s. Vaghefi, S.A., Stammbach, D., Muccione, V. *et al.* (2023), ChatClimate: Grounding conversational AI in climate science. *Commun Earth Environ* **4**, 480 (2023). https://doi.org/10.1038/s43247-023-01084-x

[42] Einen Überblick zu solchen Verfahren bieten Gu et.al. (2025), A Survey on LLM-as-a-Judge, arXiv:2411.15594v5 [cs.CL] 9 Mar 2025.

[43] S. Anonymous Authors, Designing an open-source LLM interface and social platforms for collectively driven LLM evaluation and auditing, https://openwebui.com/assets/files/whitepaper.pdf.

[44] Tessler et.al. (2024), AI can help humans find common ground in democratic deliberation, Science Vol 386, Issue 6719, DOI: 10.1126/science.adq2852.

2.2 Regulierungsansätze: Rechtliche Vorgaben für epistemisch vertrauenswürdige Chatbots?

Es bestehen also eine ganze Reihe technischer Ansätze und Verfahren, die eine Verbesserung der epistemischen Vertrauenswürdigkeit versprechen. Dabei bleibt aber die Frage noch offen, wie Anbieter und Nutzer*innen der Systeme dazu gebracht werden können, diese Ansätze tatsächlich umzusetzen. Marktdynamiken allein werden kaum ausreichen, um die notwendigen Standards zu etablieren – ohne rechtliche Pflichten oder andere regulatorische Anreize werden sich die Defizite in Sachen Faktentreue und Vertrauenswürdigkeit der Modelle nicht beheben lassen. Das liegt vor allem daran, dass sich die Tech-Unternehmen ein Wettrennen um das populärste Modell liefern – unter der Prämisse, dass sich am Ende bestenfalls einige wenige Systeme und Anwendungen durchsetzen. Nach dieser Logik erscheinen vielleicht zeit- und kostenintensive Maßnahmen zur Verbesserung der Vertrauenswürdigkeit womöglich als kurzfristig schwerwiegender Wettbewerbsnachteil.

Es wird deshalb regulatorische Anreize brauchen, damit die technischen Möglichkeiten, die epistemische Vertrauenswürdigkeit zu verbessern, auch wirklich im nötigen Umfang genutzt werden. Bislang, das zeigen auch Wachter und Kollegen,[45] gibt es aber nur wenige rechtliche Vorgaben, die Anbieter oder Nutzer*innen dazu verpflichten würden, sich im beschriebenen Sinne um die epistemische Vertrauenswürdigkeit von Chatbots zu kümmern. Zwar bestehen eine ganze Reihe von Regeln und Sorgfaltsstandards, die zu einer wahrheitsgemäßen Kommunikation, beziehungsweise zur Vermeidung falscher oder missverständlicher Aussagen oder Informationshandlungen verpflichten können. Diese „passen“ aber entweder gar nicht auf Chatbots und LLMs, sind nur in bestimmten, engen Anwendungsfeldern einschlägig, oder nur auf bestimmte, unmittelbar schädliche Anwendungsszenarien.

Die Instrumente im bestehende Rechtsrahmen und ihre Lücken – das zeigt das Öko-Institut in einem großen, gerade fertig gestellten Projektbericht für das Umweltbundesamt –[46] bieten aber reichhaltiges Anschauungsmaterial dazu, wie man es besser machen kann.

So enthält die Europäische Verordnung über Künstliche Intelligenz[47] in ihrer verabschiedeten Fassung zwar nur wenige umweltrelevante Regelungen, verdeutlicht aber immerhin, wie mit rechtlichen Mitteln auf KI-Systeme und ihre Outputs Einfluss genommen werden kann. Diese Vorgaben greifen präzise einige der zuvor skizzierten sozio-technischen Stellschrauben auf: So regelt die Verordnung in ihrem Artikel 10 die Einrichtung einer Daten-Governance, nach denen die Anbieter bestimmter Systeme z.B. dazu verpflichtet sind, die zu deren Einsatzzweck passenden Daten in entsprechender Qualität zu verwenden und schädliche „biases“ zu vermeiden – eine Antwort etwa auf oben beschriebene Probleme, wie den „common token bias" und unkuratierter Trainingsdaten. Art. 15 regelt Pflichten, die helfen sollen, die erforderliche Genauigkeit, Fehlerfreiheit und Freiheit von „Unstimmigkeiten“ der Systeme am

[45] Wachter, Mittelstadt & Russell (2024) Do large language models have a legal duty to tell the truth? R. Soc. Open Sci.11240197, http://doi.org/10.1098/rsos.240197

[46] S. Gailhofer et. al. (2025), Umweltrechtliches Regulierungskonzept für algorithmenbasierte Entscheidungssysteme. Potenziale des Umweltrechts für die ökologische Ausrichtung Künstlicher Intelligenz und autonomer Systeme. Zum Zeitpunkt des Verfassens dieses Papiers ist dieser Bericht finalisiert, aber noch nicht veröffentlicht. Weitere Informationen unter: https://www.oeko.de/projekte/detail/umweltrechtliches-regulierungskonzept-algorithmenbasierter-entscheidungssysteme/.

[47] Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz, https://eur-lex.europa.eu/eli/reg/2024/1689/oj?locale=de.

Maßstab einschlägiger Leistungsmetriken sicherzustellen, insbesondere auch solche Probleme, die aus der Interaktion der Systeme mit ihren Nutzern herrühren können – womit etwa das erläuterte Problem der „sycophancy“ adressiert werden müsste. Daneben sieht die Verordnung vor, dass die Nutzer solcher Systeme in „Betriebsanleitungen" z.B. eben über deren erwartbare „Genauigkeit" informiert werden müssen – wodurch die „epistemische Fassade" der Systeme entschärft werden könnte.

Diese Vorgaben gelten aber nur für sogenannte „Hochrisikosysteme“ – und sind für die Chatbots, um die es hier geht, nur in Ausnahmefällen relevant. Daneben enthält die Verordnung auch Vorgaben für LLMs, die dort als „Systeme mit allgemeinem Verwendungszweck“ bezeichnet werden. Danach müssen gerade für große Modelle Trainingsverfahren und Trainingsmethoden und Informationen über Art und Qualität der für das Training verwendeten Daten dokumentiert und den Anbietern von Anwendungen und Systemen, die auf Large Language Modelle aufbauen, zur Verfügung gestellt werden. Schließlich gilt für besonders große Modelle die Pflicht „systemische Risiken“ zu analysieren, also solche, die gerade aus gesellschaftlichen Wechselwirkungen von „soziotechnischen“ Systemen mit Nutzer*innen, Unternehmen und anderen Organisationen folgen. Gerade eine „systemische“ Betrachtung müsste auch subtilere, aber schwerwiegende Gefahren in den Blick nehmen, die aus verfälschten, missverständlichen oder verzerrten Informationen folgen können. Es scheint aber, auch nach Lektüre des Praxisleitfadens,[48] der zur Konkretisierung dieser Pflichten vorgelegt wurde, nicht so, als müssten Anbieter der großen Systeme zukünftig die hier besprochenen Probleme der Generierung von „persuasivem Bullshit“ als systemische Risiken vertieft analysieren.

Zusammengefasst enthält die KI-Verordnung also zwar einige – „sozio-technisch“ plausible – Vorgaben für die Anbieter und Betreiber der Systeme, durch die deren Anfälligkeit für Ungenauigkeiten, Fehler und Verzerrungen und ihre Ursachen angegangen werden soll. Diese Vorgaben werden aber wegen des engen Risikobegriffs der Verordnung meistens gerade nicht solche Systeme betreffen, die – basierend auf Claude, GPT-4 oder anderen gebräuchlichen Chatbots – Umweltinformationen vermitteln, Prognosen oder Empfehlungen ausgeben. Wo das – wie bei den Pflichten der Verordnung, die gerade die großen LLMs betreffen – anders ist, bleibt abzuwarten, ob auch die „epistemische Vertrauenswürdigkeit“ und ihre „systemischen Risiken“ in den Fokus der Anbieter und Betreiber der Systeme genommen werden müssen. Eine entscheidende Lücke liegt also darin, dass kleinere Modelle und alltägliche Informationsvermittlung oft durch das Raster fallen. Eine Pflicht für LLMs „die Wahrheit zu sagen“,[49] die den Problemen gerecht würde, lässt sich der KI-VO jedenfalls nicht entnehmen.

Das gilt auch für andere Regelungen: Das Verbraucherschutzrecht (insbes. die Richtlinie über unlautere Geschäftspraktiken,[50] umgesetzt im deutschen Gesetz gegen den unlauteren Wettbewerb) greift nur bei kommerziellen Aktivitäten mit Einfluss auf Verbraucherentscheidungen. Wer seinen Chatbot nutzt, um Produkte zu bewerben, muss also auf korrekte Angaben achten – die reine Informationsvermittlung bleibt aber unberührt, egal ob der Chatbot falsche Informationen über die Umweltbelastung

[48] EU Kommission am 10.7.2025, The General-Purpose AI Code of Practice, online verfügbar unter: https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai?utm_source=chatgpt.com

[49] Die sich in der hier diskutierten Form freilich an ihre Betreiber und Nutzer richten würde – Debatten um „eigene“, deren Rechtspersönlichkeit voraussetzende Pflichten werden hier nicht diskutiert, s.a. Gailhofer (2025),a.a.O., Abschnitt 2.5.

[50] Richtlinie 2005/29/EG vom 11. Mai 2005 über unlautere Geschäftspraktiken im binnenmarktinternen Geschäftsverkehr zwischen Unternehmen und Verbrauchern, online verfügbar unter: https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=celex:32005L0029.

von Produkten ausgibt oder historische Fakten verfälscht. Auch die Regelungen der (geplanten, zuletzt zumindest vorerst von der Kommission gestoppten) Green Claims-Richtlinie enthalten zwar konkrete Vorgaben, um die Validität von Umwelt- und Nachhaltigkeitsaussagen, zu verbessern, gelten aber eben nur für ausdrückliche Umweltaussagen, die Gewerbetreibende über Produkte, oder über Geschäftspraktiken von Unternehmen gegenüber Verbrauchern treffen. In ähnlicher Weise enthält das deutsche Äußerungsrecht zwar interessante Ansatzpunkte für die rechtliche Sanktionierung von falschen Tatsachenbehauptungen, weist aber große Lücken auf: Mit Blick auf die „Auto-Vervollständigung" Funktion der google-Suche hat der deutsche Bundesgerichtshof entschieden,[51] dass google unter bestimmten Bedingungen haften kann, wenn dabei rufschädigende Falschaussagen generiert werden. Die beanstandete Funktion weist einige Parallelen zu den Ausgaben von Large Language Models auf: Beide erzeugen Inhalte nicht auf Grundlage überprüfter Fakten oder inhaltlicher Kompetenz, sondern auf Basis statistischer Wahrscheinlichkeiten – bei der Autovervollständigung anhand von Mustern in früheren Sucheingaben anderer Nutzer*innen, bei LLMs anhand von Wahrscheinlichkeiten in großen Mengen sprachlicher Daten. Daher könnten Anbieter oder Betreiber von Chatbots zukünftig mit rechtlichen Folgen zu rechnen haben, wenn sie keine hinreichenden Vorkehrungen treffen, um Falschinformationen von LLMs zu identifizieren und möglichst auch zu vermeiden. Das gilt aber eben nur, wenn solche Falschinformationen konkrete Personen oder Unternehmen betreffen und daneben auch noch rufschädigend sind – die meisten der Probleme, die durch epistemisch unzuverlässige Systeme entstehen, bleiben deshalb außen vor.

Handlungsempfehlungen zur Verbesserung der epistemischen Vertrauenswürdigkeit

Die Anforderungen an die „epistemische Vertrauenswürdigkeit" der Systeme im bestehenden Recht weisen also signifikante Lücken auf: Zwar bestehen einige Regelungen, die die Einhaltung bestimmter Sorgfaltspflichten bei der Bereitstellung und beim Betrieb von KI-Systemen vorgeben. Diese Vorgaben betreffen aber nur ausnahmsweise die vorliegend gegenständlichen Chatbots und die Risiken, die aus „persuasivem Bullshit" (s.o.) folgen können. Zugleich bestehen einige rechtliche Anspruchsgrundlagen, nach denen sich Personen wehren können, wenn sie durch KI-generierten „Bullshit" in ihren persönlichen Rechten verletzt werden – allerdings unter so engen Voraussetzungen, dass das vorliegende Problem unzureichend adressiert wird. Auch wenn es also eine ganze Reihe von fachlichen Ansätzen gibt, die erstens zeigen, dass menschliche Fachexpertise für vertrauenswürdige Systeme erforderlich ist und zweitens beschreiben, durch welche Verfahren und Methoden dieses Erfordernis umgesetzt werden kann, bestehen bislang kaum rechtliche Anreize, solche Verfahren und Methoden tatsächlich einzusetzen.

Das muss nicht so bleiben: Wachter und Kollegen machen beispielsweise einen breit angelegten Vorschlag, um die Lücken in der Rechtslage hinsichtlich der epistemischen Vertrauenswürdigkeit von LLMs zu füllen. Sie argumentieren, dass Anbieter großer Sprachmodelle – gerade angesichts ihrer potenziell massiven gesellschaftlichen Wirkung – rechtlich verpflichtet werden sollten, die epistemische Qualität ihrer Systeme systematisch zu verbessern. Im Zentrum steht dabei eine „duty to tell the truth", also eine Pflicht, Modell-Outputs möglichst eng an überprüfbare Fakten und eine Vielfalt wissenschaftlicher Quellen zu binden – und eben die Produktion von

[51] BGH, Urteil vom 14.05.2013 - VI ZR 269/12, online verfügbar unter: https://open-jur.de/u/627117.html.

„persuasivem Bullshit“ (im philosophischen Sinne) zu vermeiden.[52] Ziel ist dabei nicht die Etablierung einer zentralen „Wahrheitsinstanz“, sondern die Beschreibung von Sorgfaltspflichten, solche Verfahren umzusetzen, die die Qualität der System-Outputs verbessern können: Vor allem Anbieter sollten ihre Modelle transparenter ausgestalten, öffentliche und zivilgesellschaftliche Beteiligung sicherstellen sowie Vertrauenswürdigkeit als zentrales Ziel ihrer Systeme verankern. Solche Vorschläge rücken also die Frage in den Vordergrund, wie verlässliches Wissen erzeugt und begründet wird, anstatt inhaltliche Vorgaben zu machen, oder allein auf technische Aspekte zu setzen.

Vorwürfe, eine solche Sorgfaltspflicht sei paternalistisch und widerspräche demokratischen Werten,[53] können dementsprechend kaum überzeugen. Kritikern ist allerdings zuzugestehen, dass rechtliche Vorgaben konkreter ausgestaltet, stärker sanktioniert und damit potenziell effektiver sein können, wenn sie für bestimmte Kontexte geschaffen werden, in denen die Schäden mangelhafter Vertrauenswürdigkeit schwerwiegend und naheliegend sind. In dem Projektbericht des Öko-Instituts für das Umweltbundesamt zur umweltrechtlichen Regulierung algorithmenbasierter Entscheidungssysteme[54] argumentieren wir – mit einem breiteren Fokus auf KI-spezifische Umweltrisiken –, dass eine KI-Regulierung risikobasiert und sektorspezifisch konkreter ausbuchstabiert werden sollte. Für entsprechende Regelungen bietet einerseits das Digital- und KI-Recht Vorlagen, das (wie gerade beschrieben) trotz seiner Lücken technologiespezifisch plausible Instrumente enthält. Andererseits verfügt auch das Umweltrecht über effektive Mittel, um wissenschaftliche Evidenz und fachliche Expertise zur Lösung sozio-technischer Probleme einzusetzen. Unter anderem aus diesem Grund sollte nicht nur untersucht werden, wie das Umweltrecht vom Technikrecht lernen kann, sondern sind auch umweltrechtliche Instrumente als mögliche Lösungsansätze für Herausforderungen der KI-Regulierung zu betrachten.

Mit Blick auf den Fokus dieses Papiers sei auf folgende Ideen verwiesen: So ist es naheliegend, die Instrumente der KI-Verordnung, die auf eine qualitativ hochwertige Datengrundlage abzielen, auf solche spezialisierten KI-Chatbots auszudehnen, die in umweltrechtlich regulierten Anwendungsfeldern eingesetzt werden, z.B. in der Landwirtschaft, in der Fachplanung, in der industriellen Anlagensteuerung. Ein naheliegender Ansatz besteht darin, dass Expert*innen – mindestens für sensible Spezialanwendungen –, aufbauend auf der Regelung des Artikel 10 KI-VO, Kriterien für die Daten-Governance und Datenqualität entwickeln, um die Basis für vertrauenswürdige Systemausgaben zu verbessern. Für die Umsetzung solcher institutionalisierter Verfahren, unter Einbindung von Fachexpert*innen oder spezialisierten Beiräten,[55] können umweltrechtliche Instrumente eine Vorlage liefern. Solche Beiräte oder Gremien könnten auch Leistungsmetriken für spezialisierte Anwendungen entwickeln und methodische Ansätze wie Benchmarking, oder RAG-Verfahren in konkrete, sektorspezifische Vorgaben oder Kriterien übersetzen. Auch die Pflicht, eine Betriebsanleitung zur Verfügung zu stellen (Artikel 13 KI-VO) sollte in solchen Anwendungsfeldern konkretere Hinweise zur Promptgestaltung, Quellenwahl und Interpretation enthalten, die unter Beteiligung wissenschaftlicher Communities entwickelt werden könnten.

[52] Z. Ganzen s. Wachter, Mittelstadt & Russell (2024) Do large language models have a legal duty to tell the truth? R. Soc. Open Sci.11240197, http://doi.org/10.1098/rsos.240197

[53] Paseri/ Durante (2025), Examining epistemological challenges of large language models in law. *Cambridge Forum on AI: Law and Governance*. 2025;1:e7. doi:10.1017/cfl.2024.7

[54] S. https://www.oeko.de/projekte/detail/umweltrechtliches-regulierungskonzept-algorithmenbasierter-entscheidungssysteme/

[55] Die Möglichkeit der Einberufung und organisationsbezogenen Regelung eines wissenschaftlichen Beirats sieht etwa § 10 des Düngegesetzes vor.

Ähnlich wie bei Umweltmanagementsystemen könnten KI-Systeme durch Zertifizierung und Audits in kritischen Anwendungsbereichen einer unabhängigen Bewertung unterzogen werden. Das Produktrecht zeigt ebenfalls, wie externe Audits systematisch zur Qualitätssicherung eingesetzt werden können – solche Ansätze könnten womöglich auch fruchtbar gemacht werden, um die Einhaltung von Verfahren zur Gewährleistung der epistemischen Vertrauenswürdigkeit nachzuweisen.

Rechtliche Regelungen könnten – um den schwer zu bewertenden Risiken „persuasiven Bullshits" systematisch zu begegnen – auch Anleihen am etablierten Instrument der Umweltverträglichkeitsprüfung (UVP) nehmen. Analog zur UVP, die komplexe Umweltwirkungen vor Projektrealisierung systematisch bewertet, könnte eine Art **„Epistemische Verträglichkeitsprüfung" (EVP)** für KI-Systeme in kritischen Anwendungsbereichen entwickelt werden, insbesondere dort, wo diese wissensproduzierend in Verwaltungsverfahren eingebunden sind. Diese würde die potenziellen Auswirkungen auf Entscheidungs- und Erkenntnisprozesse systematisch erfassen und bewerten. Dadurch könnten Ansätze zur Integration menschlicher Expertise und Mitbestimmung verfahrensrechtlich verankert werden. Die Realisierung solcher Verfahren wird durch neue, deliberative, moderierende KI-Tools auf der Basis von großen Sprachmodellen[56] in Zukunft zweifellos einfacher werden.

Damit ist den Risiken der alltäglichen und weit verbreiteten Nutzung großer Modelle noch nicht Rechnung getragen. Auch diesbezüglich könnten und sollten rechtliche Instrumente zunächst einmal helfen, unser Wissen über „epistemische“ Risiken zu erweitern. Eine naheliegende (und innovationsfreundliche) Möglichkeit hierzu zeigt wiederum das Technikrecht auf: Der noch junge Digital Services Act (DSA) verpflichtet „Very Large Online Platforms“ („VLOPS“) nach Art. 34 zur jährlichen Bewertung systemischer Risiken, einschließlich Desinformation und ggf. zur Ergreifung von risikomindernden Maßnahmen. Daneben enthält der DSA einen Anspruch auf Forschungsdatenzugang – damit qualifizierte und unabhängige Forschende diese komplexen Risiken erforschen und Gegenmaßnahmen entwickeln können. Diese Regelungen erfassen nur Plattformen mit über 45 Millionen Nutzern in der EU – reine LLM-Anbieter ohne Plattformfunktionen bleiben außen vor. Auch das muss aber nicht so bleiben. Daher sollten Fachexpert*innen Zugang zu den Modellen haben, die Chatbots zugrunde liegen, um diese gezielt auf epistemische Risiken zu untersuchen und zu bewerten.

[56] S. die Überlegungen oben zur “Faktenbewertung im Diskurs” und insbesondere Tessler et.al. (2024), AI can help humans find common ground in democratic deliberation, Science Vol 386, Issue 6719, DOI: 10.1126/science.adq2852.

Tabelle: Rechtsrahmen vs. epistemische Vertrauenswürdigkeit [57]

Regelung / Regelungsbereich	Relevante Instrumente/Mechanismen	Lücken / Begrenzungen	Möglicher Lösungsansatz
KI-Verordnung (VO (EU) 2024/1689)	• Daten-Governance (Art. 10) • Genauigkeits-/Konsistenzpflichten (Art. 15) • Dokumentationspflichten für GPAI/LLMs (Art. 53 KI-VO) • Risikomanagement für Hochrisikosysteme • besondere Pflichten für besonders große Basismodelle (Art. 55 KI-VO)	• Keine ausdrückliche Pflicht zur Faktentreue oder Offenlegung von Unsicherheiten/Wertannahmen • „Hochrisiko"-Definition zu eng, systemische Risiken vorauss. ebenso --> keine „Passung" auf „epistemische Risiken"	• Einführung einer horizontalen „epistemischen Verträglichkeitsprüfung (EVP)" nach Vorbild der UVP?
Digital Services Act (VO (EU) 2022/2065)	• Systemische Risikoanalysen (Art. 34) • Forschungsdatenzugang gegenüber VLOPs	• Beschränkt auf Plattformen > 45 Mio. Nutzer:innen • LLM-Anbieter ohne Plattformfunktion nicht erfasst	• Schaffung von Datenzugangs-Pflichten gegenüber LLM-Anbietern
Verbraucherschutzrecht (UCPD 2005/29/EG, UWG)	• Verbot irreführender Geschäftspraktiken • Pflicht zu zutreffenden Angaben bei Werbung/Verkauf	• Nur bei kommerzieller Kommunikation anwendbar • Reine Wissensvermittlung ohne Verkaufszweck bleibt unreguliert	
Green Claims-Richtlinienvorschlag (2023)	• Nachweispflichten für Nachhaltigkeitsbehauptungen • Validierung von Umweltaussagen	• Gilt nur für Unternehmenskommunikation ggü. Verbraucher:innen • gilt nicht für allgemeine Wissensvermittlung	• Übertragung der Nachweispflichten auf KI-gestützte Umwelt-Chatbots?
Äußerungs- und Deliktsrecht (BGH Autocomplete-Urteil VI ZR 269/12)	• Haftung für rufschädigende Falschaussagen	• Schutz nur für betroffene Personen/Unternehmen • Strukturelle epistemische Risiken bleiben unadressiert	
Umweltrechtliche Instrumente (analog)	• Umweltverträglichkeitsprüfung (UVP) • Beiräte (§ 10 DüngeG) • Zertifizierung/Audits • sektorale Datenstandards	• Bisher keine vergleichbaren Verfahren für KI-Systeme	• Ergänzung sektoraler Datenqualitätsstandards • Entwicklung einer „Epistemischen Verträglichkeitsprüfung" (EVP) für KI-Anwendungen in Verwaltung/Planung • Einrichtung von wiss. Gremien/Verfahren für Datenqualität und Benchmarks

Quelle: eigene Darstellung

[57] Die Tabelle ‚Rechtsrahmen vs. epistemische Vertrauenswürdigkeit' wurde mithilfe von ChatGPT (OpenAI, Version GPT-5) erstellt. Prompt (Zusammenfassung): ‚Bitte erstelle aufbauend auf dem Papier eine Tabelle, in der für zentrale Rechtsinstrumente (z. B. KI-VO, DSA, Verbraucherschutzrecht, Green Claims RL, Äußerungsrecht, Umweltrecht) jeweils die relevanten „epistemisch wirksamen" Instrumente/Mechanismen, bestehende Lücken/Begrenzungen sowie mögliche Lösungsansätze systematisch gegenübergestellt werden.'"

Fazit

Mit der fehlenden „epistemischen Vertrauenswürdigkeit“ von KI-basierten Chatbots gehen subtile Risiken einher, die bei alltäglichen Entscheidungen, insbesondere aber in administrativen, legislativen und politischen Kontexten gravierende Effekte haben könnten. Eine Reihe von technischen und methodischen Ansätzen könnte diese Risiken mindern – vor allem durch die gezielte Integration wissenschaftlicher Expertise. Die hier diskutierten Regulierungsansätze decken bei weitem noch nicht alle denkbaren Handlungsmöglichkeiten ab, zeigen jedoch auf, wie entsprechende Methoden und Verfahren durch rechtliche Instrumente aufgenommen werden könnten. Sie eröffnen hierfür ein mögliches Spektrum, das technische, verfahrensrechtliche und institutionelle Elemente verbindet: von Anforderungen an die Systemarchitektur über sektorale Datenqualitätsstandards bis hin zu dialogischen Verfahren mit Fachgemeinschaften.

Die Ausgestaltung solcher Instrumente kann sich an bewährten Modellen aus dem Umweltrecht ebenso orientieren, wie am entstehenden „KI-Recht“ und damit jeweils spezifische Aspekte der analysierten „sozio-technischen“ Probleme adressieren. Die Sicherung der epistemischen Vertrauenswürdigkeit von Large Language Models und Chatbots bleibt eine gesamtgesellschaftliche und politisch hochrelevante Herausforderung. Angesichts ihrer enormen Potenziale als verlässliche „universelle Experten“ einerseits und den gleichzeitigen Risiken einer massenhaften Verbreitung von „persuasivem Bullshit“ andererseits führt an einer verantwortungsvollen Auseinandersetzung mit dieser Problematik kein Weg vorbei.

Literatur

Albrecht, S. (2023), *ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen*. TAB-Hintergrundpapier Nr. 26. Karlsruher Institut für Technologie. https://publikationen.bibliothek.kit.edu/1000158070/150614893

Barkley, L., van der Merwe, B. (2024), Investigating the role of prompting and external tools in hallucination rates of large language models. *arXiv preprint*. https://arxiv.org/abs/2410.19385

Bergmann, D. (2024), What is *instruction Tuning*? IBM Think Blog. https://www.ibm.com/think/topics/instruction-tuning

Coeckelbergh, M. (2025), LLMs, truth, and democracy: An overview of risks. *Science and Engineering Ethics, 31*(1), 4. https://doi.org/10.1007/s11948-025-00529-0

Daniels-Koch, O., Freedman, R. (2022), The expertise problem: Learning from specialized feedback. *arXiv preprint*. https://arxiv.org/abs/2211.06519

Delacroix, S. (2025), Moral perception and uncertainty expression in LLM-augmented judicial practice. *Minds and Machines* (forthcoming). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4787044

Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., Leippold, M. (2020), CLIMATE-FEVER: A dataset for verification of real-world climate claims. *arXiv preprint*. https://arxiv.org/abs/2012.00614

Eriksson, H., …, Stojanovic, N. (2025), Can we trust AI benchmarks? An interdisciplinary review of current issues in AI evaluation. *arXiv preprint*. https://arxiv.org/abs/2502.06559

Frankfurt, H. G. (2005), *On Bullshit*. Princeton University Press.

Gailhofer, P., Hermann, A.; Franke, J.; Führ, M; Grünberger, T. (2025), *Umweltrechtliches Regulierungskonzept für algorithmenbasierte Entscheidungssysteme. Potenziale des Umweltrechts für die ökologische Ausrichtung Künstlicher Intelligenz und autonomer Systeme*. Öko-Institut e.V., Bericht für das Umweltbundesamt (im Erscheinen). https://www.oeko.de/projekte/detail/umweltrechtliches-regulierungskonzept-algorithmenbasierter-entscheidungssysteme/

Guan, J., …, Wang, W. Y. (2025), Deliberative alignment: Reasoning enables safer language models. *arXiv preprint*. https://arxiv.org/abs/2412.16339

Heersmink, R., de Rooij, B., Clavel Vázquez, M. J., Colombo, M. (2024), A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness. *Ethics and Information Technology, 26*(41). https://doi.org/10.1007/s10676-024-09777-3

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

Holtel, M. (2024), *Droht das Ende der Experten?* Verlag Franz Vahlen.

Huang, Y., Zhang, X., Zhang, C., …, Wang, W. Y. (2024), TrustLLM: Trustworthiness in large language models. *arXiv preprint*. https://arxiv.org/abs/2401.05561

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., Cobbe, K. (2023), Let's verify step by step. *arXiv preprint*. https://arxiv.org/abs/2305.20050

Lin, S., Hilton, J., Evans, O. (2021), TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint*. https://arxiv.org/abs/2109.07958

Lindsey, J., Thiel, G., Hume, T., …, Olah, C. (2025), On the biology of a large language model (Attribution Graphs). Anthropic, Transformer Circuits. https://transformer-circuits.pub/2025/attribution-graphs/biology.html

Liu, Y., Liu, Z., Yu, K. (2025), What's the most important value? INVP: Investigating the value priorities of LLMs through decision-making in social scenarios. *Proceedings of the 31st International Conference on Computational Linguistics*, 4725–4752.

Mazeika, J., …, Arora, S. (2025), Utility engineering: Analyzing and controlling emergent value systems in AIs. *arXiv preprint*. https://arxiv.org/abs/2502.08640

Mina, M., Ruiz-Fernández, V., Falcão, J., Vasquez-Reina, L., Gonzalez-Agirre, A. (2025), Cognitive biases, task complexity, and result interpretability in large language models. *Proceedings of COLING 2025.* https://aclanthology.org/2025.coling-main.120.pdf

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., …, Moret-Bonillo, V. (2023), Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review, 56*, 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Munn, L., Magee, L., Arora, V. (2024), Truth machines: Synthesizing veracity in AI language models. *AI & Society, 39*, 2759–2773. https://doi.org/10.1007/s00146-023-01756-4

Paseri, L., Durante, M. (2025), Examining epistemological challenges of large language models in law. *Cambridge Forum on AI: Law and Governance, 1*(e7). https://doi.org/10.1017/cfl.2024.7

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., Farajtabar, M. (2025), The illusion of thinking: Understanding the strengths and limitations of reasoning models via the

lens of problem complexity. Apple ML Research. https://machinelearning.apple.com/research/illusion-of-thinking

Tessler, M. H., Hawkins, R. D., Schulz, L. E., Goodman, N. D. (2024), AI can help humans find common ground in democratic deliberation. *Science, 386*(6719). https://doi.org/10.1126/science.adq2852

Vaghefi, S. A., Stammbach, D., Muccione, V., …, Bürger, G. (2023), ChatClimate: Grounding conversational AI in climate science. *Communications Earth & Environment, 4*, 480. https://doi.org/10.1038/s43247-023-01084-x

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., Hajishirzi, H. (2020), Fact or fiction: Verifying scientific claims. *Proceedings of EMNLP 2020*. https://doi.org/10.18653/v1/2020.emnlp-main.609

Wachter, S., Mittelstadt, B., Russell, C. (2024), Do large language models have a legal duty to tell the truth? *Royal Society Open Science, 11*, 240197. https://doi.org/10.1098/rsos.240197

Wang, X., Sen, P., Li, R., Yilmaz, E. (2024), Simulated task oriented dialogues for developing versatile conversational agents. In: *Advances in Information Retrieval: ECIR 2024 Proceedings, Part I* (pp. 157–172). Springer. https://doi.org/10.1007/978-3-031-56027-9_10

Öko-Institut | Freiburg | Darmstadt | Berlin

Das Öko-Institut ist eines der europaweit führenden, unabhängigen Forschungs- und Beratungsinstitute für eine nachhaltige Zukunft. Seit der Gründung im Jahr 1977 erarbeitet das Institut Grundlagen und Strategien, wie die Vision einer nachhaltigen Entwicklung global, national und lokal umgesetzt werden kann. Das Institut ist an den Standorten Freiburg, Darmstadt und Berlin vertreten.
oeko.de | info@oeko.de

Kontakt

Dr. Peter Gailhofer | Öko-Institut e.V. | +49 30 405085-352| p.gailhofer@oeko.de

Das Policy Paper wurde im Rahmen des Projekts „Schreiben mit künstlicher Intelligenz – Fakten oder Fiktion? Chancen und Risiken von KI-Sprachmodellen: Wie einfach ist es für Nutzer*innen, verlässliche Informationen zu Klima- und Umweltschutzthemen zu erhalten?“ erstellt.