Fachbereich WD 5

Bias und Halluzinationen von KI-Systemen

Kategorien, Ursachen und Minimierungsstrategien

Bias und Halluzinationen von KI-Systemen
Kategorien, Ursachen und Minimierungsstrategien

Aktenzeichen:	WD 5 - 3000 - 082/25
Abschluss der Arbeit:	19.11.2025
Fachbereich:	WD 5: Wirtschaft, Energie und Klima

Inhaltsverzeichnis

1. Einleitung

Künstliche Intelligenz (KI) ist mehr als ein technisches Werkzeug – sie gehört mittlerweile bewusst oder unbewusst in den Alltag vieler Menschen beginnend bei der Nutzung von Smartphones, beim Einsatz am Arbeitsplatz bis hin zu Anwendungen in Sicherheitsinfrastrukturen. Im Privatbereich berechnen beispielsweise Wearables Kalorien, während im Unternehmenseinsatz Algorithmen Lagerstände prognostizieren oder Cyber-Bedrohungen in Netzwerken erkennen (BSI, 2024).[1] In Deutschland nutzen laut unterschiedlichen Befragungen zwischen 10 % und 37 % der Unternehmen KI-Lösungen.[2] Bei Studierenden in Deutschland ist die Nutzungsquote von KI im Studium innerhalb von zwei Jahren von 63 auf 91 % gestiegen (2023-2025).[3] 62 % der Jugendlichen in Deutschland zwischen 12 und 19 Jahren nutzten 2024 KI-Anwendungen. Die häufigste Anwendung war mit 65 % der Einsatz als Hilfsmittel für die Schule und für Hausaufgaben – noch vor „Spaß haben“ und „sich informieren“.[4]

Auch der Arbeitsmarkt steht vor deutlichen strukturellen Veränderungen. Eine Studie des Internationalen Währungsfonds geht davon aus, dass in naher Zukunft global fast 40 % – in hochentwickelten Volkswirtschaften bis zu 60 % – der Arbeitsplätze durch KI beeinflusst werden.[5]

Die duale Wirkung von erhofftem Nutzen (z. B. Effizienzgewinn, neue Lern- und Arbeitsformen, gesellschaftliche Chancengleichheit) und schwer abschätzbaren Risiken (Substitution von Arbeitsplätzen, Diskriminierung, verstärkte Desinformation) prägt die Diskussion über KI.[6] Zugleich verläuft die Entwicklung im KI-Bereich rasant: Das erste „Key-Update“ des International AI Safety Report – nur einige Monate nach dem Hauptbericht – verweist auf messbare Verbesserungen bei den bestehenden Risiken in den KI-Systemen.[7]

---

1 BSI (2024), Transparenz von KI-Systemen, https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/White-paper-Transparenz-KI-Systeme.pdf?__blob=publicationFile&v=3.

2 Wissenschaftliche Dienste (2025), Der Einfluss von Künstlicher Intelligenz auf den Arbeitsmarkt, WD 5 - 3000 - 066/25, S. 11 ff.

3 von Garrel & Mayer (2025), Künstliche Intelligenz im Studium - Eine quantitative Längsschnittstudie zur Nutzung KI-basierter Tools durch Studierende (2023 & 2025). Online unter: https://doi.org/10.48444/h_docs-pub-533.

4 IWD (2025), Wie Jugendliche künstliche Intelligenz nutzen, https://www.iwd.de/artikel/wie-jugendliche-kuenstliche-intelligenz-nutzen-645347/.

5 Cazzaniga, Jaumotte, Li, et al. (2024), Gen-AI: Artificial Intelligence and the Future of Work, IMF Staff Discussion Notes 2024, 001 (2024), https://doi.org/10.5089/9798400262548.006.

6 OECD.AI Policy Observatory (2022), OECD Input to the United Nations High-Level Advisory Body on Artificial Intelligence, https://wp.oecd.ai/app/uploads/2024/03/OECD-Input-to-the-UN-High-Level-Advisory-Board-on-AI-final.pdf, S. 12.

7 International AI Safety Report: First Key Update (2025), https://internationalaisafetyreport.org/sites/default/files/2025-10/first-key-update_0.pdf.

Vor dem Hintergrund dieser Perspektiven untersucht diese Ausarbeitung die zwei Hauptrisiken von KI-Systemen genauer, die vor allem im Bereich der generativen KI eine Rolle spielen: **Bias** und **Halluzinationen**. Ausgangspunkt der Arbeit waren die folgenden Leitfragen:

- Welche Formen von Bias und Halluzinationen gibt es? Wie verbreitet sind Halluzinationen?
- Wie können Bias und daraus resultierende unerwünschte Diskriminierungen verhindert werden?
- Wie können Halluzinationen und daraus resultierende unerwünschte Falschinformationen verhindert werden?

Zur Klärung dieser Fragen geht Kapitel 2 grundlegend auf den Kontext von Vertrauen und Transparenz bei KI-Systemen ein. Kapitel 3 beschäftigt sich mit Bias-Kategorien, Ursachen und Maßnahmen zur Verzerrungsminimierung. Kapitel 4 widmet sich dem Thema Halluzinationen, konkret deren Definitionen, Kategorien, Entstehungsursachen und Lösungsansätze zur Reduktion.

2. „KI Black Box“, Vertrauen und Transparenz

Mit der schnellen Verbreitung digitaler Anwendungen mit Künstlicher Intelligenz (KI) steigt deren technische Komplexität. Viele zugrunde liegende KI-Basismodelle (sog. Foundation Models)[8] wirken als sog. „Black Boxes“, weil ihre internen Abläufe selbst für Experten kaum nachvollziehbar sind. Nutzerinnen und Nutzer sehen nur Eingaben und Ausgaben. Wie ein Modell zu einer konkreten Entscheidung gelangt, bleibt verborgen. Diese Intransparenz erschwert die Beurteilung des Wahrheitsgehalts der Ausgaben und macht es schwierig, angemessene Entscheidungen auf dieser Basis zu treffen. In Bereichen wie Medizin, Finanzwesen, öffentliche Verwaltung oder Justiz gewinnt diese Problematik zusätzlich an Bedeutung, da dort Fehlentscheidungen unmittelbare Folgen auf Lebensverhältnisse haben.[9]

Transparenz ist eine zentrale Voraussetzung, um dieses Vertrauen zu ermöglichen. Sie soll die Nutzenden befähigen, Entscheidungen eines KI-Systems einzuordnen, seine Grenzen zu erkennen und Risiken abzuschätzen. Transparenz umfasst daher nicht nur die Darstellung der Fähigkeiten eines Systems, sondern auch die Limitierungen, Fehlermuster und Rahmenbedingungen.

8 Siehe Bertelsmann Stiftung (o.D.), Das Ökosystem der KI-Basismodelle: Wie Daten, Energie und menschliche Arbeit KI-Basismodelle formen, reframe[Tech], https://www.reframetech.de/wissensseite-basismodelle/; IBM (o.D.), What are foundation models?, https://www.ibm.com/think/topics/foundation-models.

9 BSI (2024), Transparenz von KI-Systemen, https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/White-paper-Transparenz-KI-Systeme.pdf?__blob=publicationFile&v=3, S. 5.; OEDC (2025), What explainable AI is, why it matters and how we can achieve it, https://oecd.ai/en/wonk/what-explainable-ai-is-why-it-matters-and-how-we-can-achieve-it.

Fehlt diese Einordnung, bleiben Überprüfbarkeit und Verantwortlichkeit unklar, und Nutzerinnen und Nutzer können weder Fairness noch Zuverlässigkeit verlässlich beurteilen.[10]

Die europäische KI-Verordnung[11] greift diese Anforderungen auf und definiert verbindliche Transparenzpflichten. Artikel 50 verpflichtet Anbieter bestimmter Systeme – insbesondere solcher mit allgemeinem Verwendungszweck oder Anwendungen wie Emotionserkennung – dazu, klare Informationen über Funktionsweise, Zweckbestimmung, Leistungsgrenzen und mögliche Risiken bereitzustellen.[12] Diese Vorgaben bilden die Grundlage dafür, dass Systeme in der EU zugelassen und eingesetzt werden dürfen. Das Prinzip „transparency by design“ fordert zudem, Transparenz von Beginn an in die Entwicklung zu integrieren, statt sie nachträglich herzustellen.[13] Um Transparenz und damit Vertrauen in KI-Systeme zu gewährleisten, müssen Risiken in Form eines Risikomanagements bewusst benannt, identifiziert, bewertet, migriert und überwacht werden.[14]

Während Transparenz und grundlegende Risikobewertungen die Rahmenbedingungen definieren, rücken nun jene Risiken in den Fokus, die im praktischen Einsatz besonders häufig auftreten und erhebliche Auswirkungen haben können. Zwei Risikokategorien sind dabei zentral: **Bias**, also systematische Verzerrungen in Daten oder Modellverhalten, und **Halluzinationen**, also die Erzeugung faktisch falscher oder erfundener Inhalte durch generative Modelle. Beide Phänomene entstehen nicht zufällig, sondern sind strukturell mit der Funktionsweise moderner KI verbunden. Sie betreffen unterschiedliche Ebenen – von der Datengrundlage über das Modellverhalten bis zur Interaktion mit Nutzerinnen und Nutzern – und bilden damit die zentralen Risiken, die in den folgenden Abschnitten vertieft analysiert werden.

3. Bias in KI-Systemen

3.1. Menschlicher Bias, Bias in KI-Systemen und algorithmische Diskriminierung

Bias bzw. Verzerrungen beschreibt zunächst eine grundlegende Eigenschaft menschlichen Denkens. Kognitive Verzerrung entsteht aus der menschlichen, begrenzten Perspektive auf die Welt und dem Bedürfnis, Informationen zu vereinfachen und zu generalisieren.[15] „Bias“ bzw. „die

10 BSI (2024), Transparenz von KI-Systemen, https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/White-paper-Transparenz-KI-Systeme.pdf?__blob=publicationFile&v=3, S. 12.

11 https://artificialintelligenceact.eu/.

12 https://artificialintelligenceact.eu/article/50/.

13 BSI (2024), Transparenz von KI-Systemen, https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/White-paper-Transparenz-KI-Systeme.pdf?__blob=publicationFile&v=3, S. 16.

14 AI Action Summit (2025), International AI Safety Report - The International Scientific Report on the Safety of Advanced AI, https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf, S. 157 ff.

15 SAP (2024), Was ist KI-Bias? Ursachen, Auswirkungen und Strategien zur Bekämpfung, https://www.sap.com/germany/resources/what-is-ai-bias.

kognitive Verzerrung“ ist ein kognitionspsychologischer Sammelbegriff. Er beschreibt, wie Menschen zu verarbeitende Informationen systematisch falsch verstehen oder mental repräsentieren.[16] Es gibt eine Vielzahl an kognitiven Verzerrungen, die das menschliche Handel beeinflussen und die aus fünf grundlegenden Ursachen hergeleitet werden können: [17]

- Verkürzungen bei der Informationsverarbeitung
- Begrenzte Verarbeitungsfähigkeit des Gehirns
- Emotionale und moralische Motivationen
- Verzerrungen beim Speichern und Abrufen von Erinnerungen
- Sozialer Einfluss

Diese mentale Abkürzung erleichtert Lernprozesse, kann jedoch zu systematisch verzerrten Urteilen führen.[18] Ethisch relevant wird ein Bias dort, wo solche Verzerrungen nicht nur die Wahrnehmung strukturieren, sondern konkrete Nachteile für andere erzeugen. Die Übergänge zwischen individueller Voreingenommenheit und diskriminierendem Verhalten sind fließend. Beides basiert auf selektiven Bewertungen, die bestimmte Personen oder Gruppen benachteiligen.

Auf dieser Grundlage lassen sich sogenannte **algorithmische Verzerrungen** (im engl. Fachjargon „algorithmic Bias“ oder nur „Bias“) verstehen, die in KI-Systemen enthalten sein können. KI-Systeme sind keine neutralen Entitäten, sondern übernehmen Muster aus ihren Entwicklungs- und Einsatzkontexten.[19] „**Bias**“ in KI bezeichnet systematische Fehler, die bestimmten Gruppen oder Weltanschauungen Vorteile verschaffen und für andere nachteilige oder ungerechte Ergebnisse erzeugen. Diese Verzerrungen können aus verschiedenen Quellen stammen: fehlerhafte oder unausgewogene Trainingsdaten, unzureichende Modellierungsentscheidungen oder aus gesellschaftlichen Ungleichheiten, die in den Daten bereits eingeschrieben sind. Damit werden bestehende menschliche und strukturelle Voreingenommenheit nicht nur abgebildet, sondern potenziell verstärkt.

KI-Systeme spiegeln beispielsweise historische und aktuelle Ungleichheiten wider und können sie durch automatisierte Entscheidungen reproduzieren.[20] Bias kann dabei in allen Phasen des maschinellen Lernens auftreten – in der Datenerfassung, im Algorithmusdesign und in den vom

16 https://de.wikipedia.org/wiki/Kognitive_Verzerrung.

17 Für eine Übersicht siehe: https://www.visualcapitalist.com/every-single-cognitive-bias/.

18 Siehe im Folgenden SAP (2024), Was ist KI-Bias? Ursachen, Auswirkungen und Strategien zur Bekämpfung, https://www.sap.com/germany/resources/what-is-ai-bias.

19 Vgl. im Folgenden AI Action Summit (2025), International AI Safety Report - The International Scientific Report on the Safety of Advanced AI, https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf, S. 92.

20 IBM (o.D.), KI-Verzerrungen anhand von Beispielen aus der realen Welt beleuchten, https://www.ibm.com/de-de/think/topics/shedding-light-on-ai-bias-with-real-world-examples.

Modell generierten Vorhersagen. Allerdings betonen Experten die Beziehung zwischen menschlichem und algorithmischem Bias: KI erweitert die Reichweite dieser menschlichen Verzerrungen, indem sie sie großflächig skalieren und automatisieren kann., indem beispielsweise automatisierte Verwaltungsentscheidungen für ganze Bevölkerungsgruppen getroffen werden.[21]

Bias ist der zentrale Mechanismus, durch den algorithmische Diskriminierung entsteht. Verzerrte oder unausgewogene Daten, fehlerhafte Kennzeichnungen und algorithmische Gewichtungen übertragen bestehende gesellschaftliche Ungleichheiten in KI-Systeme und formen daraus systematische Benachteiligungen bestimmter Gruppen.[22] Da diese Verzerrungen häufig aus scheinbar neutralen Kriterien resultieren, treten sie überwiegend als indirekte oder intersektionale Diskriminierung auf.[23] Algorithmische Diskriminierung ist damit nicht nur das Ergebnis isolierter offensichtlicher Fehler, sondern auch die konsequente Fortsetzung menschlicher, institutioneller und historischer – teilweise verborgener – Bias-Strukturen im technischen System.

3.2. Auftreten von KI-Bias

Fälle von KI-Bias zeigen, dass dieser in zahlreichen gesellschaftlichen Bereichen auftreten kann und häufig aus unausgewogenen Trainingsdaten, strukturellen Ungleichheiten oder fehlerhaften Modellarchitekturen entsteht. Die illustrativen Beispiele in der folgenden Tabelle umfassen sowohl reale nachgewiesene Verzerrungen als auch hypothetische oder modellhafte Szenarien, die in Studien und Tests identifiziert wurden. Gemeinsam verdeutlichen sie, dass solche Verzerrungen nicht auf einen einzelnen Sektor beschränkt sind, sondern in sämtlichen Lebensbereichen mit dem Einsatz von KI-Systemen auftreten können.

Tabelle 1: Beispiele von Bias in KI-Systemen

Anwendungsbereich / Sektor	**Beschreibung des Bias-Falles**	**Quelle**
Personalwesen / Recruiting	Algorithmus sortiert Bewerbungen von Frauen aus, weil Trainingsdaten überwiegend männliche Lebensläufe enthielten	IBM (2024)
Personalwesen / Recruiting	„Natural Language Processing“-Probleme führen dazu, dass ein Algorithmus männlich konnotierte Begriffe bevorzugt	IBM (o.D.)

21 Ein Negativbeispiel ist die Toeslagenaffaire (zu dt. Kindergeldaffäre) in den Niederlanden Ende 2020, in der hauptsächlich migrantischen Familie zu Unrecht Sozialbetrug vorgeworfen wurde. (https://netzpolitik.org/2021/kindergeldaffaere-niederlande-zahlen-millionenstrafe-wegen-datendiskriminierung/).

22 AI Action Summit (2025), International AI Safety Report - The International Scientific Report on the Safety of Advanced AI, https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf, S. 92; IBM (o.D.), KI-Verzerrungen anhand von Beispielen aus der realen Welt beleuchten, https://www.ibm.com/de-de/think/topics/shedding-light-on-ai-bias-with-real-world-examples.; SAP (2024), Was ist KI-Bias? Ursachen, Auswirkungen und Strategien zur Bekämpfung, https://www.sap.com/germany/resources/what-is-ai-bias.

23 European Union Agency for Fundamental Rights (2022), Bias in Algorithms – Artificial Intelligence and Discrimination, https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf, S. 24.

Anwendungsbereich / Sektor	**Beschreibung des Bias-Falles**	**Quelle**
Personalwesen / Recruiting	KI-basierte Einstellungsmodelle ersetzen personalorientierte Prozesse und können bestehende Verzerrungen reproduzieren	Schwartz et al. (2022)
Gesundheitswesen	Diagnosen weniger genau bei schwarzen Patienten aufgrund unterrepräsentierter Daten	IBM (o.D.)
Gesundheitswesen	Verzerrte Diagnosen oder Empfehlungen, wenn Trainingsdaten nur aus einer ethnischen Gruppe stammen	SAP (2024)
Kreditvergabe / Finanzsystem	Strengere Kreditbewertung für einkommensschwache Regionen; Benachteiligung bestimmter Gruppen	SAP (2024)
Versicherungen	Höhere Versicherungsprämien für Minderheiten durch Nutzung von Postleitzahlen	SAP (2024)
Strafverfolgung / Predictive Policing	Modelle verstärken „Racial Profiling“ aufgrund historischer Verhaftungsdaten	European Union Agency for Fundamental Rights (2022)
Strafverfolgung / Predictive Policing	EU-Test eines automatisierten Systems zeigt potenzielle Verzerrungen und Feedback-Loops	European Union Agency for Fundamental Rights (2022)
Online-Werbung	Hochbezahlte Stellenanzeigen werden Männern häufiger angezeigt als Frauen	IBM (o.D.)
Inhaltsmoderation / Soziale Medien	Minderheitenbeiträge werden häufiger fälschlich als unangemessen eingestuft	SAP (2024)
Inhaltsmoderation / Soziale Medien	Algorithmen verstärken politische Echokammern	SAP (2024)
Inhaltsmoderation / Soziale Medien	Algorithmus zeigt ethnische und geschlechtsspezifische Verzerrungen	European Union Agency for Fundamental Rights (2022)
Bildung	Prognosemodelle bevorzugen Schüler und Schülerinnen aus besser ausgestatteten Schulen	SAP (2024)
Bildgenerierung	Midjourney zeigt ältere Personen fast ausschließlich als Männer	IBM (o.D.)
Bildgenerierung	Unterrepräsentation oder stereotype Darstellung ethnischer Gruppen	SAP (2024)
Spracherkennung	Schlechtere Ergebnisse für Akzente, Dialekte oder Nichtmuttersprachler	SAP (2024)

Anwendungsbereich / Sektor	Beschreibung des Bias-Falles	Quelle
Bevorzugung von Texten	Wenn KI die Wahl zwischen Texten eines Menschen und einer anderen KI hat, bevorzugt sie die KI-geschriebenen Texte (KI-für-KI-Bias)	Laurito, Davis, Grietzer et al. (2025)[24]

3.3. Kategorien von Bias und Diskriminierung in KI-gestützten Systemen

Bias in KI-gestützten Systemen entsteht durch technische, datenbezogene und soziale Faktoren, die entlang des gesamten KI-Lebenszyklus wirken. Verzerrungen können in unterschiedlichsten Formen auftreten und zu systematischen Benachteiligungen einzelner Gruppen führen. Für eine differenzierte Analyse sind insbesondere die Klassifikationen des Bundesamtes für Sicherheit in der Informationstechnik (BSI) und des AI Action Summit zentral, da sie Bias umfassend nach Lebenszyklusphasen und Datenprozessen strukturieren.

Das BSI (2025)[25] ordnet eine Auswahl von 11 Bias-Arten den Phasen eines KI-Systems zu und macht damit sichtbar, dass Verzerrungen in jeder Stufe entstehen können. Es unterscheidet Bias bei der Datenerhebung und -präparation, Bias in Modellentwicklung und Training, bei der Modellevaluierung sowie Bias im Ausrollen und beim Monitoring. Die ergänzenden Begriffe – etwa „Emergent Bias“, der durch sich verändernde Datenverteilungen im Einsatz entsteht, oder „Temporal Bias“, der strukturelle Veränderungen über die Zeit abbildet – verdeutlichen, dass Bias dynamisch ist. Zusätzlich berücksichtigt das BSI sozial-psychologische Formen wie Verhaltens-Bias oder sozialem Bias, die zeigen, wie menschliches Verhalten und gruppenbezogene Einflüsse in KI-Systeme hineinwirken.

Der International AI Safety Report (2025)[26] ergänzt diese Perspektive durch eine detaillierte Klassifikation der Verzerrungen im Datenproduktionszyklus. Die dortige Struktur unterscheidet unter anderem Stichproben-, historische-, Erfassungs- und Messverzerrung in der Datenerhebung; unterschiedliche Formen von Kennzeichnungs-Bias wie kulturellem oder „Confirmation“ Bias; „Feature Bias“, das aus der Auswahl oder Gewichtung einzelner Merkmale entsteht; sowie Benchmark- und „Evaluation Bias“, die durch nichtrepräsentative Testdaten hervorgerufen werden. Diese Kategorisierung legt den Schwerpunkt klar auf datengetriebene Mechanismen und zeigt auf, wie Fehler in Datensammlung, Kennzeichnung und Testverfahren zu verzerrten

24 Laurito, Davis, Grietzer et al. (2025), AI–AI bias: Large language models favor communications generated by large language models, Proc. Natl. Acad. Sci. U.S.A. 122 (31) e2415697122, https://doi.org/10.1073/pnas.2415697122.

25 Bundesamt für Sicherheit in der Informationstechnik (BSI) (2025), Bias in der künstlichen Intelligenz, Vers. 1.1, https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Whitepaper_Bias_KI.pdf?__blob=publicationFile&v=5, S. 7-12.

26 AI Action Summit (2025), International AI Safety Report The International Scientific Report on the Safety of Advanced AI, https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf, S. 97.

Ergebnissen führen. Für Bewertungs- und Empfehlungssysteme ist insbesondere der historische Bias[27] relevant, weil Modelle häufig vergangene Präferenz- oder Entscheidungsstrukturen reproduzieren.

27 „Historischer Bias tritt durch Datenverzerrungen auf, die zustande kommen, wenn Daten Sichtweisen und Konzepte widerspiegeln, die in einer modernen Welt nicht mehr zeitgemäß sind.“ (Bundesamt für Sicherheit in der Informationstechnik (BSI) (2025), Bias in der künstlichen Intelligenz, Vers. 1.1, https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Whitepaper_Bias_KI.pdf?__blob=publicationFile&v=5, S. 7).

Lifecycle Stage	Bias Source	Description	Examples
Data Collection	Sampling Bias	Certain perspectives, demographics, or groups are overrepresented or underrepresented in the data.	A dataset for a news aggregator containing primarily sources that favour a particular ideology, leading to skewed results
	Selection Bias	Only certain data types or contexts are included, limiting representativeness.	Language datasets that exclude non-Western languages, limiting model performance in global applications.
Data Annotation	Labeller Bias	Annotators' backgrounds, perspectives, and cultural biases affect their understanding and classification of data, influencing the labelling process.	Annotators label speech by individuals from lower socioeconomic backgrounds as unprofessional or inappropriate, leading to biased decisions.
Data Curation	Historical Bias	Reflecting or perpetuating past societal biases within curated data.	A hiring dataset that favours certain demographics based on historical hiring practices, embedding existing inequalities in AI models.
Data Pre-processing	Feature Selection Bias	Excluding relevant features from a dataset.	Excluding age or gender as features in healthcare models, potentially impacting the relevance of predictions for these demographics.
Model Training	Label Imbalance	Unequal representation in labelled data, leading to biased model outputs.	A classification model trained on 80% male-labelled images might perform poorly when identifying female images.
Deployment Context	Contextual Bias	A model is trained on data from a context that differs from the context of application, leading to worse outcomes for certain groups.	An English-only model deployed in multilingual settings, causing misinterpretations for non-English users.
Evaluation & Validation	Benchmark Bias	Evaluation benchmarks favour certain groups or knowledge bases over others.	AI models evaluated primarily on US-centric datasets fail to generalise well in non-Western settings.
Feedback Mechanisms	Feedback Loop Bias	Models learn from biased user feedback, reinforcing initial biases.	A recommendation system that receives more engagement on certain types of content may reinforce exposure to the same biased content.

Abbildung 1: Bias/Verzerrungen können in verschiedenen Phasen des Datenproduktionszyklus für KI-Systeme auftreten[28]

28 AI Action Summit (2025), International AI Safety Report The International Scientific Report on the Safety of Advanced AI, https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf, S. 97.

IBM (2023)[29] beschreibt praxisnahe Formen wie algorithmische Verzerrung, Ausschluss- und Messverzerrung, stereotype Verzerrung oder Outgroup-Homogenitätsbias. Diese Begriffe zeigen, wie technische und menschliche Faktoren ineinandergreifen – etwa durch inkonsistente Datenkennzeichnung, subjektive Entscheidungen oder die Übertragung gesellschaftlicher Stereotype in Trainingsdaten. SAP (2024)[30] ergänzt dies durch eine vierteilige Einordnung in Datenbias, algorithmischen Bias, menschliche Voreingenommenheit und Bias in generativer KI. Letztere gewinnt insbesondere im Kontext KI-gestützter Inhalte-Erzeugung an Relevanz, da generative Modelle dazu neigen, stereotype oder unterrepräsentierte Darstellungen zu verstärken.

Domänenspezifisch zeigen Hasanzadeh et al. (2025) für den Gesundheitsbereich, dass ein Großteil der Verzerrungen menschlichen Ursprungs ist. Bias kann auch dort in jeder Phase des Lebenszyklus entstehen, von der Konzeption über die Datensammlung bis zur Modellvalidierung und dem Ausrollen (Abbildung 2).[31] Die Autoren betonen zudem zeitabhängige Effekte wie „Concept Shift“, bei dem sich Bedeutungen oder klinische Praktiken verändern und Modelle nicht mehr zur aktuellen Versorgungssituation passen.

29 IBM (2023), Was ist KI-Verzerrung? https://www.ibm.com/de-de/think/topics/ai-bias.

30 SAP (2024), Was ist KI-Bias? Ursachen, Auswirkungen und Strategien zur Bekämpfung, https://www.sap.com/germany/resources/what-is-ai-bias.

31 Hasanzadeh, Josephson, Waters et al. (2025), Bias recognition and mitigation strategies in artificial intelligence healthcare applications, npj Digit. Med, 8, 154, https://doi.org/10.1038/s41746-025-01503-7.

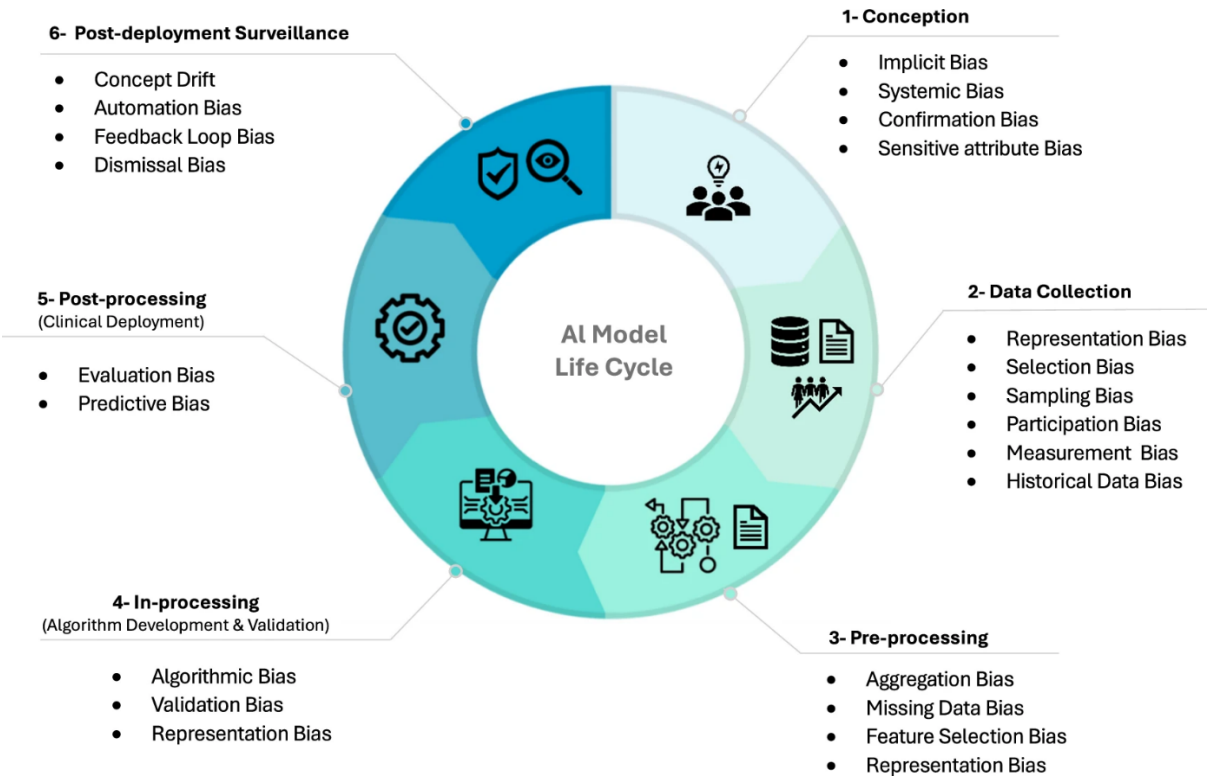

Abbildung 2: Der Lebenszyklus eines KI-Modells und häufige Verzerrungen in den einzelnen Phasen[32]

Die identifizierten Bias-Formen führen in der Anwendung zu verschiedenen Formen algorithmischer Diskriminierung. Die EU Agency for Fundamental Rights (2022)[33] unterscheidet direkte, indirekte, multiple bzw. intersektionale sowie Diskriminierung durch Assoziation. Für moderne Bewertungs- und Empfehlungssysteme ist insbesondere indirekte Diskriminierung zentral, da viele Verzerrungen aus scheinbar neutralen Kriterien resultieren. Systeme können beispielsweise bestimmte demografische Gruppen systematisch schlechter bewerten oder benachteiligen, obwohl diese Merkmale nicht explizit im Modell enthalten sind.

3.4. Ursachen

Bias in KI-Systemen entsteht durch ein Zusammenspiel gesellschaftlicher, technischer und organisatorischer Faktoren. Die Ursachen sind vielfältig und reichen von der Erzeugung und Strukturierung von Daten über algorithmische Entscheidungen bis zur Art und Weise, wie KI-Modelle eingesetzt werden. Verzerrungen bzw. Bias stellen nicht allein technische Fehler dar, sondern sind Ausdruck menschlicher Entscheidungen, kultureller Prägungen und sozialer Ungleichheiten. Für einen illustrativen Überblick zu den Ursachen und Auswirkungen von KI-Bias bietet die

32 Ebd., S. 3.

33 European Union Agency for Fundamental Rights (2022), Bias in Algorithms – Artificial Intelligence and Discrimination, https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf, S. 24.

von Smith und Rustagi (2020) entwickelte „Bias in KI“-Karte ein nachvollziehbares Schema, da sie zentrale Einflusspunkte entlang des gesamten KI-Lebenszyklus visualisiert (Abbildung 3).

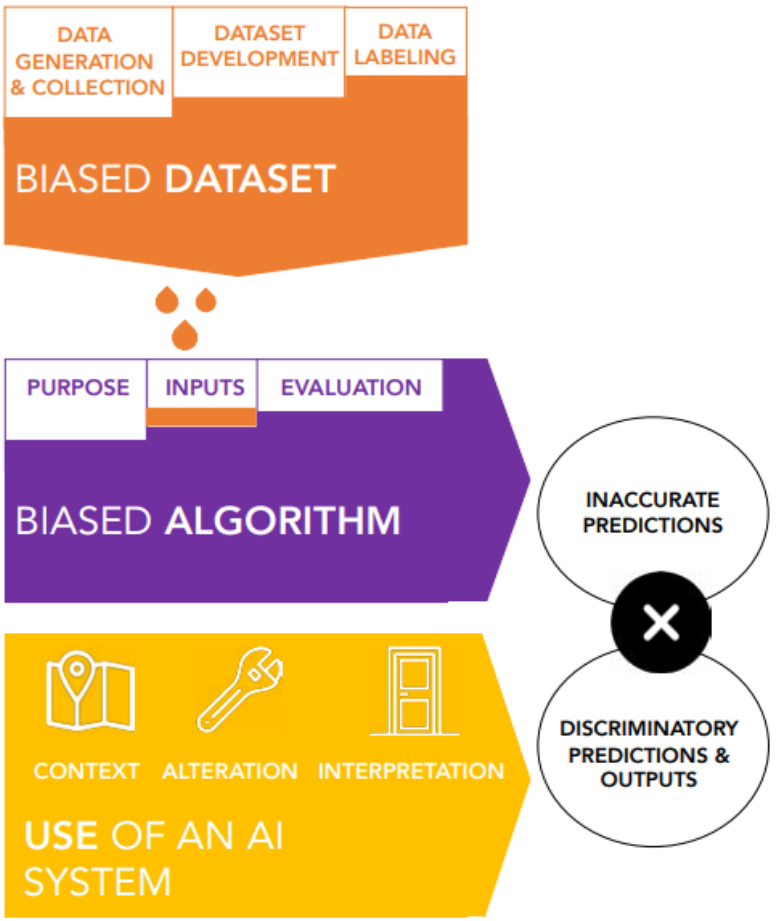

Abbildung 3: Die „Bias in KI“-Karte (Bias in AI Map)[34]

Die Abbildung zeigt aus einer schematischen Perspektive, wie KI-Systeme zu ungenauen oder diskriminierenden Vorhersagen und Ergebnissen führen können.

Ein erster zentraler Ursachenbereich betrifft die **Erhebung, Struktur und Kennzeichnung von Daten**.[35] Daten spiegeln gesellschaftliche Realität wider, sind jedoch stets durch bestehende Ungleichheiten, historische Muster und subjektive Entscheidungen geprägt.[36] Unterrepräsentationen einzelner Gruppen oder unausgewogene Stichproben führen dazu, dass Modelle bestimmte

34 Smith & Rustagi (2020), Mitigating Bias in Artificial Intelligence – An Equity Fluent Leadership Playbook, Berkeley Haas Center for Equity, Gender and Leadership, https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf, S. 23.

35 Ebd. S. 22.

36 Ebd, S. 24.

Bevölkerungsgruppen systematisch schlechter abbilden können. Beispiele hierfür sind Gesichtserkennungsdatensätze, die Personen mit heller Hautfarbe überrepräsentieren und bei Personen anderer Hautfarben höhere Fehlerquoten erzeugen.[37] Auch geografisch konzentrierte Sicherheitsdaten können in polizeilichen Anwendungen rassistische Verzerrungen verstärken, weil sie bestehende Überwachungsmuster reproduzieren. Die Kennzeichnung von Daten stellt eine weitere Fehlerquelle dar. Inkonsistente oder subjektive Labels – etwa in Rekrutierungssystemen – können dazu führen, dass bestimmte Merkmale überbewertet und qualifizierte Bewerberinnen und Bewerber systematisch benachteiligt werden.[38] Demnach sind Daten nie neutral, sondern gehen stets aus sozialen, politischen und ökonomischen Kontexten hervor.[39]

37 IBM (o.D.), KI-Verzerrungen anhand von Beispielen aus der realen Welt beleuchten, https://www.ibm.com/de-de/think/topics/shedding-light-on-ai-bias-with-real-world-examples.

38 Ebd.

39 Schwartz, Vassilev, Greene et. al (2022), Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD, [online], https://doi.org/10.6028/NIST.SP.1270, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf; Smith & Rustagi (2020), Mitigating Bias in Artificial Intelligence – An Equity Fluent Leadership Playbook, Berkeley Haas Center for Equity, Gender and Leadership, https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf, S. 23.

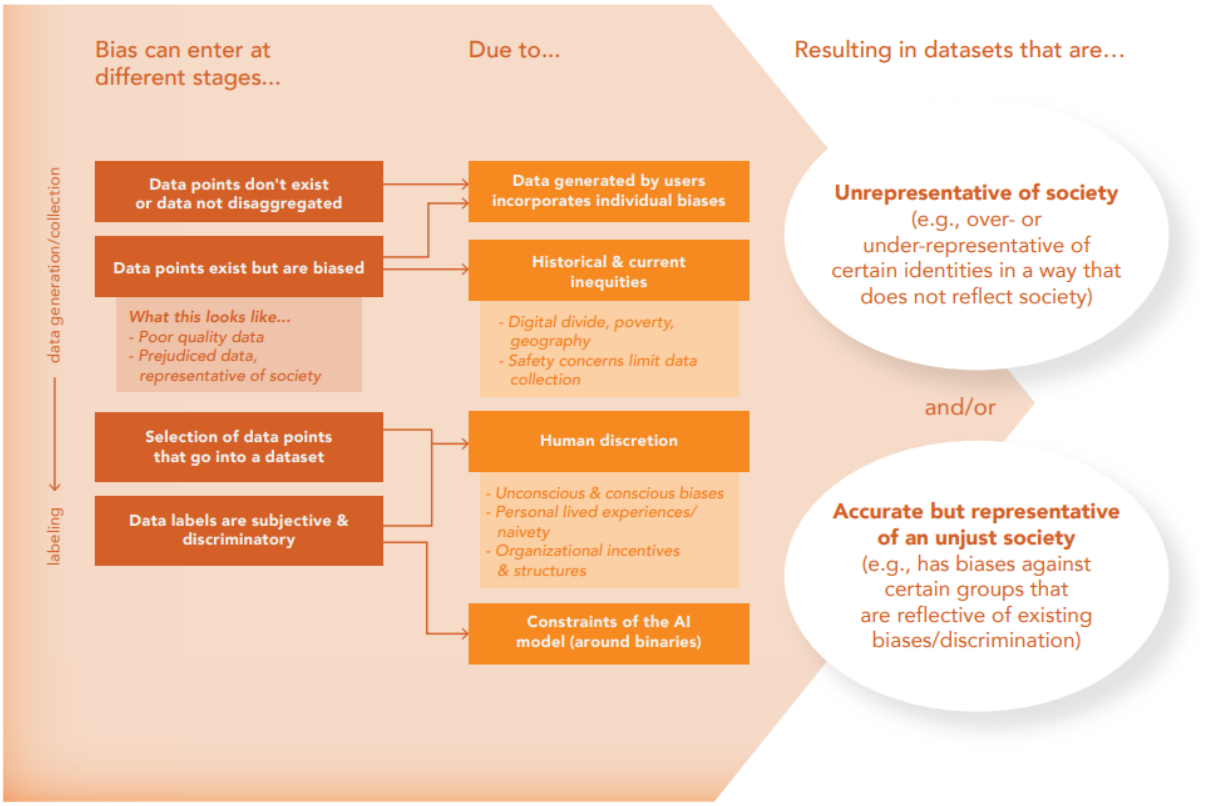

Abbildung 4: Pfade eines voreingenommenen Datensets[40]

Eine zweite Ebene der Ursachen umfasst **algorithmische Verzerrungen**, die sich aus Modellierungsentscheidungen, Parametern und Anwendungsannahmen ergeben. Algorithmen können einseitige Datenmuster nicht nur reproduzieren, sondern auch verstärken, wenn Modellarchitekturen Verzerrungen systematisch gewichten oder bestimmte Signale überbetonen.[41] Menschliche Vorannahmen können zusätzlich in die Modellgestaltung einfließen, etwa wenn Entwicklerteams Variablen unbewusst anhand kultureller Normen priorisieren.[42] Dies zeigt sich beispielsweise bei Hate-Speech-Erkennungssystemen, deren Modelle aufgrund der Trainingsdaten auf bestimmte Begriffe „überreagieren“ können und neutrale Inhalte dann fälschlich als beleidigend

40 Smith & Rustagi (2020), Mitigating Bias in Artificial Intelligence – An Equity Fluent Leadership Playbook, Berkeley Haas Center for Equity, Gender and Leadership, https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf, S. 23.

41 IBM (o.D.), KI-Verzerrungen anhand von Beispielen aus der realen Welt beleuchten, https://www.ibm.com/de-de/think/topics/shedding-light-on-ai-bias-with-real-world-examples.

42 Smith & Rustagi (2020), Mitigating Bias in Artificial Intelligence – An Equity Fluent Leadership Playbook, Berkeley Haas Center for Equity, Gender and Leadership, https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf, S. 23.

markieren.[43] Der Fall verdeutlicht, wie algorithmische Strukturen und Trainingsdaten zusammenwirken und Verzerrungen sowohl technischer als auch sozialer Natur sind.

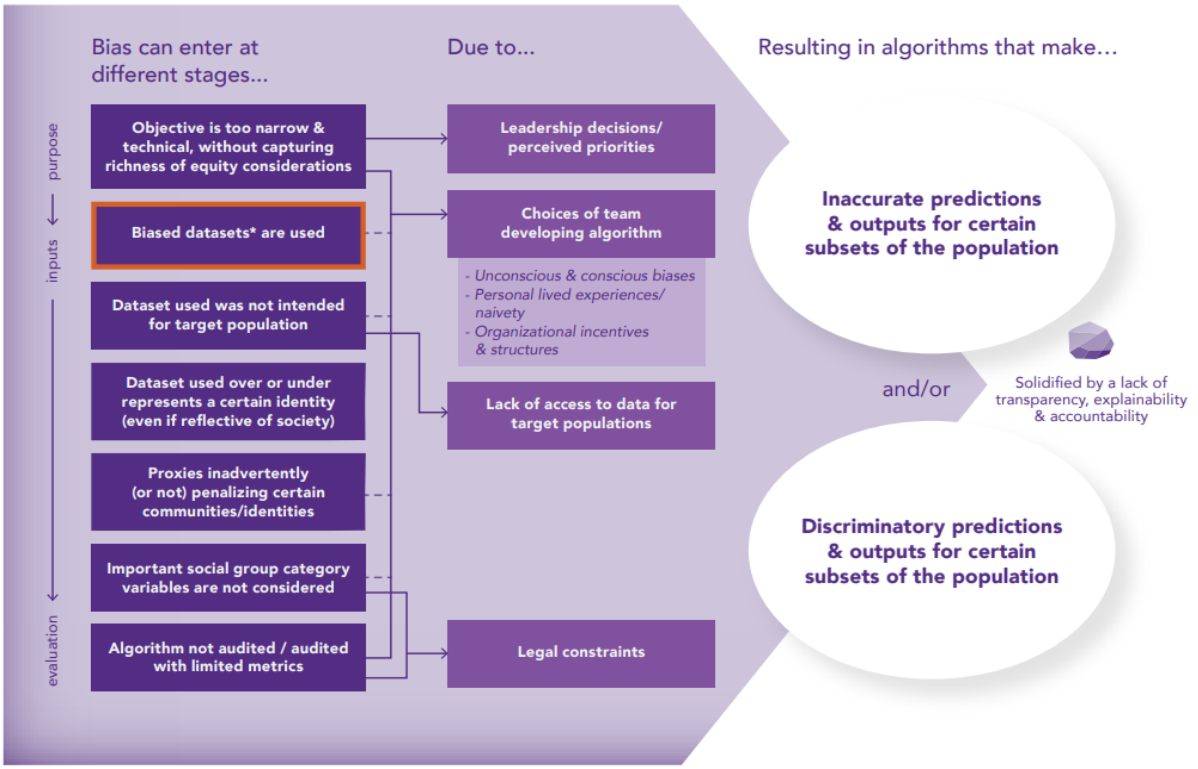

Abbildung 5: Pfade eines voreingenommenen Algorithmus[44]

Die dritte Ursache liegt in der **Nutzung und Einbettung von KI-Systemen** (Abbildung 6). Verzerrungen entstehen nicht nur in Daten und Modellen, sondern auch dort, wo KI-Systeme reale Entscheidungen beeinflussen. Selbst korrekt trainierte Modelle können verzerrte Wirkungen entfalten, wenn sie in Kontexten eingesetzt werden, die sich von den ursprünglichen Trainingsbedingungen unterscheiden, etwa bei abweichenden Zielgruppen, geografischen Räumen oder Anwendungsfällen.[45] Besonders zentral sind in diesem Zusammenhang „Feedback-Loops“, die in der Nutzungsebene entstehen. Sie treten auf, wenn die Vorhersagen eines Systems das Verhalten von Organisationen oder Individuen prägen und diese veränderte Realität anschließend erneut in die Datengrundlage zurückfließt. Dadurch wird der Output eines Modells zum Input derselben

43 European Union Agency for Fundamental Rights (2022), Bias in Algorithms – Artificial Intelligence and Discrimination, https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf, S. 11 f.

44 Smith & Rustagi (2020), Mitigating Bias in Artificial Intelligence – An Equity Fluent Leadership Playbook, Berkeley Haas Center for Equity, Gender and Leadership, https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf, S. 30.

45 Smith & Rustagi (2020), Mitigating Bias in Artificial Intelligence – An Equity Fluent Leadership Playbook, Berkeley Haas Center for Equity, Gender and Leadership, https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf, S. 34 f.

Modelllogik, wodurch bestehende Verzerrungen verstärkt werden können.[46] Im Bereich des „Predictive Policing“ (Strafverfolgung) führt dieser Mechanismus etwa dazu, dass empirische Schwerpunkte polizeilicher Interventionen durch frühere algorithmische Empfehlungen erzeugt oder verschoben werden. Die so entstehenden Daten bestätigen anschließend vermeintliche Risikogebiete und verschärfen den Bias über Zeit, was in sogenannten „Runaway Feedback Loops“ (außer Kontrolle geratene Rückkopplungsschleifen) resultieren kann.[47]

Darüber hinaus können Verzerrungen in der Nutzungsebene auch aus menschlichen Interpretationen der Modellergebnisse entstehen. Wenn Entscheidungsträger algorithmischen Empfehlungen zu viel Autorität zuschreiben oder eigene Vorannahmen bei der Auswertung einfließen lassen, verstärkt dies die Wirkung von Verzerrungen zusätzlich.[48] Damit zeigt sich, dass Bias nicht allein im technischen System verankert ist, sondern durch die Interaktion von KI-Ausgaben, organisationalen Handlungen und sich dynamisch verändernden Datenumwelten entsteht.

46 European Union Agency for Fundamental Rights (2022), Bias in Algorithms – Artificial Intelligence and Discrimination, https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf, S. 8.

47 Ebd.

48 Smith & Rustagi (2020), Mitigating Bias in Artificial Intelligence – An Equity Fluent Leadership Playbook, Berkeley Haas Center for Equity, Gender and Leadership, https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf, S. 35.

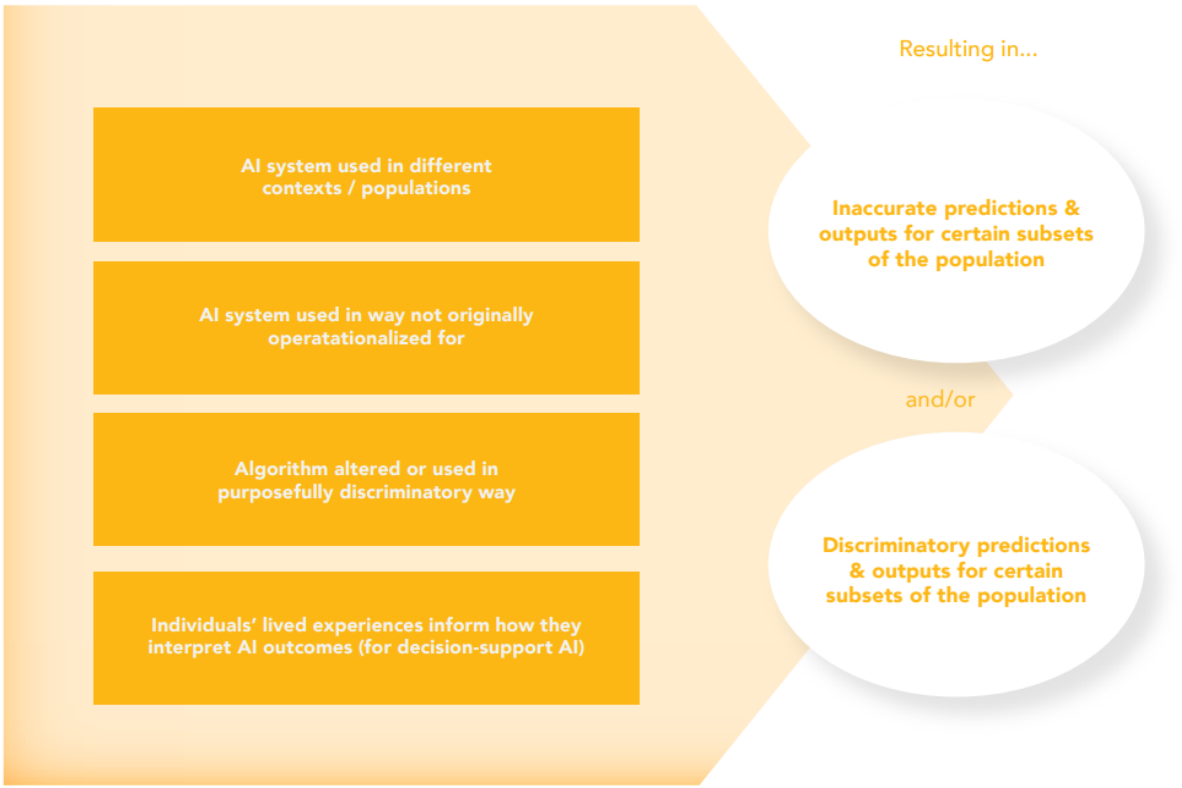

Abbildung 6: Verzerrungen durch den Einsatz von KI-Systemen[49]

3.5. Auswirkungen: Ungenaue Vorhersagen und diskriminierender Output

Bias in KI-gestützten Systemen führt vor allem zu zwei zentralen Problemen: ungenauen Vorhersagen und diskriminierenden Ausgaben. Wie Smith und Rustagi (2020) zeigen, entstehen diese Effekte als Ergebnis des gesamten KI-Lebenszyklus – von der Datengenerierung und -kennzeichnung über die Modellgestaltung bis zum Einsatz der Systeme. Die bereits eingeführte „Bias in KI"-Karte (Kap. 3.4) zeigt, dass Verzerrungen über mehrere Pfade in die Modelle einfließen können, sich im praktischen Einsatz kumulieren und möglicherweise in Kontexten genutzt werden, für die sie nicht entwickelt wurden (**Fehler! Es wurde kein Textmarkenname vergeben.**Abbildung 7).

49 Smith & Rustagi (2020), Mitigating Bias in Artificial Intelligence – An Equity Fluent Leadership Playbook, Berkeley Haas Center for Equity, Gender and Leadership, https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf, S. 35.

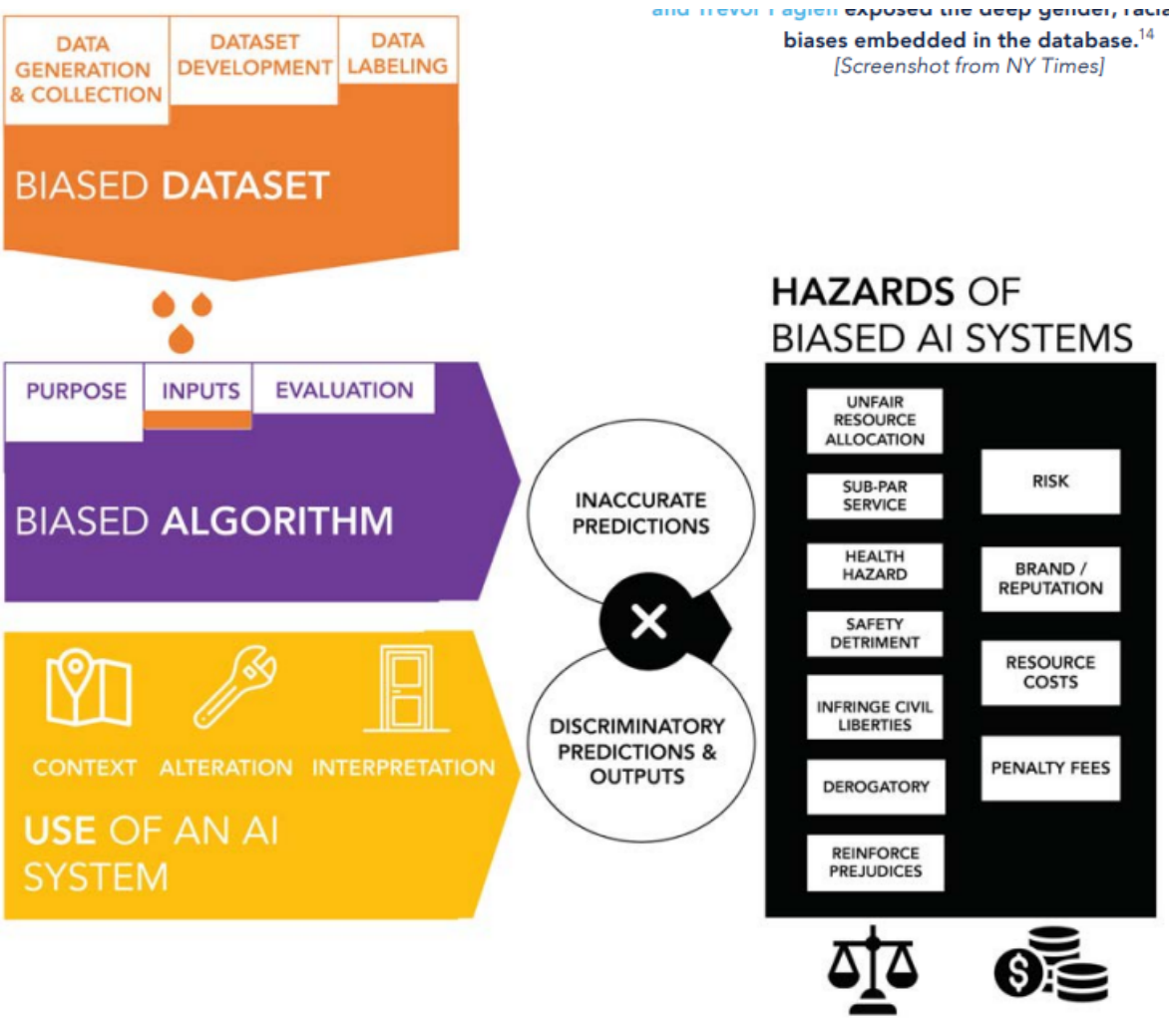

Abbildung 7: „Bias in KI“-Karte[50]

Verzerrte Ausgaben sowie fehlerhafte Vorhersagen stellen nicht nur technische Ungenauigkeiten dar, sondern können zu ernsthaften gesellschaftlichen und wirtschaftlichen Folgen führen. Die Ergebnisse reichen von unpassenden Risikoeinstufungen über eingeschränkte berufliche Chancen bis hin zu problematischen Empfehlungen in sensiblen Bereichen wie Gesundheit oder Bildung.[51] Solche Modelle können soziale Ungleichheiten verstärken, weil marginalisierte Gruppen häufiger von fehlerhaften Entscheidungen betroffen sind. Hinzu kommt, dass KI-Systeme stereotype Muster reproduzieren können, etwa wenn Sprachmodelle Berufe mit bestimmten

50 Smith & Rustagi (2020), Mitigating Bias in Artificial Intelligence – An Equity Fluent Leadership Playbook, Berkeley Haas Center for Equity, Gender and Leadership, https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf, S. 6.

51 SAP (2024), Was ist KI-Bias? Ursachen, Auswirkungen und Strategien zur Bekämpfung, https://www.sap.com/germany/resources/what-is-ai-bias.

Geschlechtern verbinden oder generative Modelle verzerrte Darstellungen sozialer Gruppen erzeugen.

Eine zusätzliche Gefahr kommt durch die Leistungsfähigkeit der technischen Systeme zustande: Diskriminierende Outputs können dadurch in einer Geschwindigkeit und Reichweite entstehen, die traditionelle menschliche Entscheidungsprozesse übertrifft.[52] Werden diese Verzerrungen nach einem Schaden bemerkt, entstehen auch organisationale Risiken wie Reputationsschäden, Vertrauensverlust oder negative wirtschaftliche Folgen.[53]

Mit zunehmendem Einsatz und Leistungsfähigkeit von KI-Systemen nimmt die Forderung nach Fairness, gesellschaftliche Teilhabe und institutionelle Verantwortung zu. Den Risiken, die aus Bias und Diskriminierung entstehen, muss daher aktiv entgegengewirkt werden.

3.6. Minimierung von Bias und logarithmischer Diskriminierung

Die Verhinderung von Bias und algorithmischer Diskriminierung in KI-Systemen ist möglich. Es ist ein mehrschichtiger Prozess, der technische, organisatorische, personelle, nutzerorientierte und gesellschaftliche Maßnahmen erfordert. Wie bereits gezeigt wurde, kann Bias entlang des gesamten Lebenszyklus eines KI-Modells entstehen. Geeignete Maßnahmen sind daher eine Kombination aus strukturellen und technischen Interventionen. In der Praxis können Maßnahmen auf fünf Ebenen gebündelt werden:

Im Zentrum steht die **technische Ebene**, die darauf abzielt, Verzerrungen in Daten, Algorithmen oder Outputs unmittelbar zu mindern. Pre-Processing-Methoden wie Datenbereinigung, Neugewichtung, Re-Annotation oder die Nutzung synthetischer Daten sollen Verzerrungen bereits vor dem Modelltraining reduzieren.[54] In-Processing-Methoden verändern Trainingsprozesse oder Modellarchitekturen und integrieren fairnessorientierte Mechanismen, um diskriminierende Muster während des Lernens abzuschwächen.[55] Post-Processing-Techniken korrigieren Modelloutputs,

---

52 Schwartz, Vassilev, Greene et. al (2022), Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD, [online], https://doi.org/10.6028/NIST.SP.1270, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf.

53 SAP (2024), Was ist KI-Bias? Ursachen, Auswirkungen und Strategien zur Bekämpfung, https://www.sap.com/germany/resources/what-is-ai-bias.

54 AI Action Summit (2025), International AI Safety Report - The International Scientific Report on the Safety of Advanced AI, https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf, S. 95 ff.; SAP (2024), Was ist KI-Bias? Ursachen, Auswirkungen und Strategien zur Bekämpfung, https://www.sap.com/germany/resources/what-is-ai-bias.

55 AI Action Summit (2025), International AI Safety Report - The International Scientific Report on the Safety of Advanced AI, https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf, S. 95 ff.; Bundesamt für Sicherheit in der Informationstechnik (BSI) (2025), Bias in der künstlichen Intelligenz, Vers. 1.1, https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Whitepaper_Bias_KI.pdf?__blob=publicationFile&v=5, S. 7-12.

etwa durch Kalibrierung oder Filter für problematische Ausgaben.[56] Diese Maßnahmen werden durch kontinuierliches Monitoring ergänzt, da im laufenden Betrieb neue Verzerrungen entstehen können – etwa durch Konzeptdrift oder Veränderungen der Zielpopulation.[57] Allerdings wird auch auf die Grenzen dieser technischen Debiasing-Methoden hingewiesen. Sie können neue Verzerrungen erzeugen, aufwendige Re-Annotationen verursachen oder in Zielkonflikte mit Genauigkeit, Effizienz oder Datenschutz geraten:

> "It is difficult to effectively address discrimination concerns, as bias mitigation methods are not reliable. Bias mitigation challenges existed before general-purpose AI systems (…), but current techniques to address bias can unintentionally create new biases despite considerable progress on this front. (…) A significant challenge in mitigating AI risks also lies in defining and measuring effective outcomes, particularly for bias. It remains unclear how to measure bias in general, how to distinguish between data that reflects legitimate demographic differences (e.g. disease prevalence by population) and data that inherently perpetuates bias, and what an ideal, measurable end state looks like. (…) Bias reduction can conflict with other desiderata, and achieving complete algorithmic fairness may not be technically feasible. Many desirable properties of general-purpose AI systems involve trade-offs, such as the four-way trade-off between fairness, accuracy, privacy, and efficiency (…). Attempts to ensure fairness can have downsides."[58]

Die **organisatorische Ebene** schafft die strukturelle Grundlage dafür, dass technische Maßnahmen nachhaltig wirken. Dazu gehören definierte Verantwortlichkeiten für Bias, organisatorische Prozesse zur frühzeitigen Datenevaluierung und klare Vorgaben für den Umgang mit vortrainierten Modellen.[59] Weitergehende Governance-Strukturen verankern sog. „Responsible AI"-Ansätze, „Corporate Social Responsibility"-Elemente und unternehmensweite Leitlinien, die Fairness über den gesamten Entwicklungsprozess hinweg berücksichtigen.[60] Zudem müssen Modelle nicht als

---

56 SAP (2024), Was ist KI-Bias? Ursachen, Auswirkungen und Strategien zur Bekämpfung, https://www.sap.com/germany/resources/what-is-ai-bias; Bundesamt für Sicherheit in der Informationstechnik (BSI) (2025), Bias in der künstlichen Intelligenz, Vers. 1.1, https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Whitepaper_Bias_KI.pdf?__blob=publicationFile&v=5, S. 7-12.

57 Hasanzadeh, Josephson, Waters et al. (2025), Bias recognition and mitigation strategies in artificial intelligence healthcare applications, npj Digit. Med, 8, 154, https://doi.org/10.1038/s41746-025-01503-7.

58 AI Action Summit (2025), International AI Safety Report The International Scientific Report on the Safety of Advanced AI, https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf, S. 96.

59 Bundesamt für Sicherheit in der Informationstechnik (BSI) (2025), Bias in der künstlichen Intelligenz, Vers. 1.1, https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Whitepaper_Bias_KI.pdf?__blob=publicationFile&v=5, S. 7-12.

60 Smith & Rustagi (2020), Mitigating Bias in Artificial Intelligence – An Equity Fluent Leadership Playbook, Berkeley Haas Center for Equity, Gender and Leadership, https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf, S. 50.

statische Produkte, sondern als Systeme mit Bedarf an kontinuierlicher Verbesserung und Feedback gestaltet werden.[61]

Die **Team- und Kompetenzebene** stellt sicher, dass menschliche Entscheidungen nicht selbst zur Quelle von Verzerrungen werden. Divers zusammengesetzte und multidisziplinär arbeitende Teams verbessern die Fähigkeit, Bias in Daten und Modellen zu erkennen und sachgerecht zu bewerten.[62] Ein ausgeprägtes Bias-Wissen der Entwickelnden über relevante Verzerrungsformen und deren Entstehungswege ist zentral.[63] Eine organisationsweite Kultur ethischer Reflexion unterstützt diesen Prozess und fördert kritisches Hinterfragen in der Modellentwicklung.[64]

Auf der **Benutzer- und Anwendungsebene** spielt die Interaktion mit KI-Systemen eine wesentliche Rolle. Nutzerinnen und Nutzer müssen sich der Existenz versteckter Verzerrungen bewusst sein und generierte Inhalte kritisch prüfen.[65] Präzise Prompts, die Formulierung neutraler Anfragen und die Bereitstellung ausreichender Kontexte können Verzerrungen reduzieren. Ebenso hilfreich sind Rückfragen an das System oder das bewusste Zuweisen von Rollen, um reflektierte Ausgaben zu erzeugen.[66] Besonders in Anwendungen, die vulnerable Gruppen betreffen, ist eine kontinuierliche Beobachtung der Systemwirkung maßgeblich.[67]

Die **gesellschaftliche und regulatorische Ebene** definiert schließlich die Rahmenbedingungen, in denen KI-Fairness entsteht. Repräsentative und vielfältige Datensätze sind eine Grundvoraussetzung für faire Modelle; „Daten-Monokulturen“ können hingegen zu systematischen Verzerrungen führen.[68] Gleichzeitig bedarf es klarer Fairnessindikatoren, regulativer Leitlinien und

---

61 Washington (2025), Fragile Foundations: Hidden Risks of Generative AI, Hrsg. Bertelsmann Stiftung https://www.bertelsmann-stiftung.de/fileadmin/files/user_upload/Fragile_foundations_risks_of_generativ_AI_2025.pdf, S. 24.

62 Smith & Rustagi (2020), Mitigating Bias in Artificial Intelligence – An Equity Fluent Leadership Playbook, Berkeley Haas Center for Equity, Gender and Leadership, https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf, S. 48.

63 Bundesamt für Sicherheit in der Informationstechnik (BSI) (2025), Bias in der künstlichen Intelligenz, Vers. 1.1, https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Whitepaper_Bias_KI.pdf?__blob=publicationFile&v=5, S. 3.

64 Smith & Rustagi (2020), Mitigating Bias in Artificial Intelligence – An Equity Fluent Leadership Playbook, Berkeley Haas Center for Equity, Gender and Leadership, https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf, S. 48.

65 IBM (2024), Bias vermeiden: Künstliche Intelligenz richtig steuern, https://de.newsroom.ibm.com/Bias-vermeiden-Kunstliche-Intelligenz-richtig-steuern.

66 Ebd.

67 Washington (2025), Fragile Foundations: Hidden Risks of Generative AI, Hrsg. Bertelsmann Stiftung https://www.bertelsmann-stiftung.de/fileadmin/files/user_upload/Fragile_foundations_risks_of_generativ_AI_2025.pdf, S. 24 ff.

68 Ebd.

transparenter Verfahren, die über einzelne Organisationen hinauswirken.[69] Die Einbindung von Entwickler-Communities und Betroffenen kann dazu beitragen, Auswirkungen frühzeitig zu erkennen und systemische Verzerrungen offenzulegen.[70]

Insgesamt zeigt sich, dass keine einzelne Maßnahme allein ausreicht. Effektive Bias-Minderung entsteht erst durch das Zusammenspiel technischer Lösungen, organisatorischer Verantwortung, diverser Kompetenzen, reflektierter Nutzung und gesellschaftlicher Steuerung. Fairness ist kein Endzustand, sondern ein fortlaufender Prozess begleitet durch kontinuierliche Überwachung, Lernbereitschaft und Anpassung entlang des KI-Lebenszyklus.[71]

4. Halluzinationen von generativen KI-Systemen

4.1. Halluzinationen, Konfabulation und Falschinformationen

Bias ist ein strukturelles Risiko in nahezu allen KI-Systemen, während Halluzinationen vor allem in generativen Modellen wie Large Language Models (LLMs = große Sprachmodelle) auftreten. Sie entstehen aus denselben wahrscheinlichkeitsberechnenden Mechanismen, die flüssige und flexible Textgenerierung ermöglichen und markieren damit ein spezifisches Verhalten generativer Systeme.[72]

Mit der breiten Nutzung von LLMs in Bereichen wie Bildung, Medizin oder Verwaltung wächst der Bedarf an verlässlichen Systemen. Trotz Leistungssteigerungen bleiben heutige Modelle fehleranfällig, weshalb Nutzerinnen und Nutzer Ausgaben kontrollieren und Erwartungen an die Zuverlässigkeit gegebenenfalls nachjustieren müssen.[73] Halluzinationen bezeichnen hierbei

---

69 AI Action Summit (2025), International AI Safety Report - The International Scientific Report on the Safety of Advanced AI, https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf, S. 92.

70 SAP (2024), Was ist KI-Bias? Ursachen, Auswirkungen und Strategien zur Bekämpfung, https://www.sap.com/germany/resources/what-is-ai-bias.

71 Hasanzadeh, Josephson, Waters et al. (2025), Bias recognition and mitigation strategies in artificial intelligence healthcare applications, npj Digit. Med, 8, 154, https://doi.org/10.1038/s41746-025-01503-7; Bundesamt für Sicherheit in der Informationstechnik (BSI) (2025), Bias in der künstlichen Intelligenz, Vers. 1.1, https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Whitepaper_Bias_KI.pdf?__blob=publicationFile&v=5, S. 3.

72 Für eine vereinfachte Beschreibung der Funktionsweise von sog. Basismodellen (engl. „Foundation-Models"), zu denen bspw. die großen Sprachmodellen zählen, siehe Kap. 4.4.

73 Zhou, Schellaert, Martínez-Plumed et al. (2024), Larger and more instructable language models become less reliable, Nature 634, 61–68, https://doi.org/10.1038/s41586-024-07930-y.

Inhalte, die von überprüfbaren Fakten abweichen, darunter erfundene Details, falsche Zitate oder andere unzuverlässige Ergebnisse.[74]

Die Begrifflichkeit hat sich dabei gewandelt. Unter Experten wurde ursprünglich von „Konfabulation“ gesprochen, da die Fehler eher falsch konstruierte Antworten als Halluzinationen (Wahrnehmungsstörungen) darstellen.[75] Öffentlich und zunehmend auch fachlich hat sich jedoch der Begriff Halluzination durchgesetzt, verstärkt durch die gesellschaftliche Aufmerksamkeit für das Phänomen.[76]

Für die Regulierung und den praktischen Umgang mit diesen Systemen ist entscheidend, die Gründe und die Unterschiede in der Halluzinationsanfälligkeit verschiedener Modelle zu verstehen. Das Wissen unterstützt die Risikobewertungen und bestimmt, wie viel menschliche Aufsicht bei der verantwortungsvollen Nutzung erforderlich ist[77], da Halluzinationen zugleich eine Quelle von Falschinformationen bilden. Wenn fehlerhafte Modellinhalte ungeprüft in Anwendungen oder Entscheidungen einfließen, entfaltet sich möglicherweise ein gesellschaftlicher Schaden.

4.2. Auftreten von Halluzinationen

Ein zentrales Kriterium für die Bewertung generativer KI ist die Frage, wie häufig und in welchen Situationen LLMs falsche Inhalte erzeugen. Studien und Praxiserfahrungen zeigen, dass Halluzinationen in unterschiedlichen Kontexten auftreten und sowohl alltägliche Anwendungen als auch sicherheitskritische Bereiche betreffen.

Erste Hinweise auf das Ausmaß liefern empirische Untersuchungen zu faktischen Fehlern.[78] Untersuchungen zeigen ein sehr uneinheitliches Bild. So produzierten laut einer Studie in 2024 verschiedene Chatbots bei Referenzauskünften Fehlerquoten zwischen 30 und 90 Prozent, als etwa Titel, Autorenschaft oder Publikationsjahr abgefragt wurden. Solche Fehlangaben gewinnen an Bedeutung, da Nutzende häufig dazu neigen, Modellantworten ungeprüft zu übernehmen. Ein

---

74 Rawte, Chakraborty, Pathak, et al. (2023), The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, S. 2541–2573, https://aclanthology.org/2023.emnlp-main.155.pdf.; AI Action Summit (2025), International AI Safety Report - The International Scientific Report on the Safety of Advanced AI, https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf, S. 88.

75 https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence).

76 Jones (2025), AI hallucinations can’t be stopped — but these techniques can limit their damage, Nature News Feature, https://www.nature.com/articles/d41586-025-00068-5.

77 Rawte, Chakraborty, Pathak, et al. (2023), The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, S. 2541–2573, https://aclanthology.org/2023.emnlp-main.155.pdf.

78 Vgl. im Folgenden Jones (2025), AI hallucinations can’t be stopped — but these techniques can limit their damage, Nature News Feature, https://www.nature.com/articles/d41586-025-00068-5.

bekanntes Beispiel ist der Fall eines US-Anwalts, der 2023 erfundene Gerichtsentscheidungen in einem Schriftsatz zitierte, nachdem er sich auf ein generatives Modell verlassen hatte.

Aktuelle Benchmarks zeigen jedoch, dass Halluzinationsraten stark modellabhängig sind und sich in konkreten Aufgaben erheblich unterscheiden. Der öffentliche Vectara-Benchmark, der prüft, wie oft Modelle beim Zusammenfassen von Dokumenten erfundene Inhalte einfügen, weist Spannweiten von 0,6 bis 1,9 Prozent aus (Abbildung 8). Ältere Versionen desselben Vergleichs zeigen immer wieder andere Platzierungen verschiedener Sprachmodelle, tendenziell aber eine generelle Abnahme der Halluzinationsraten.

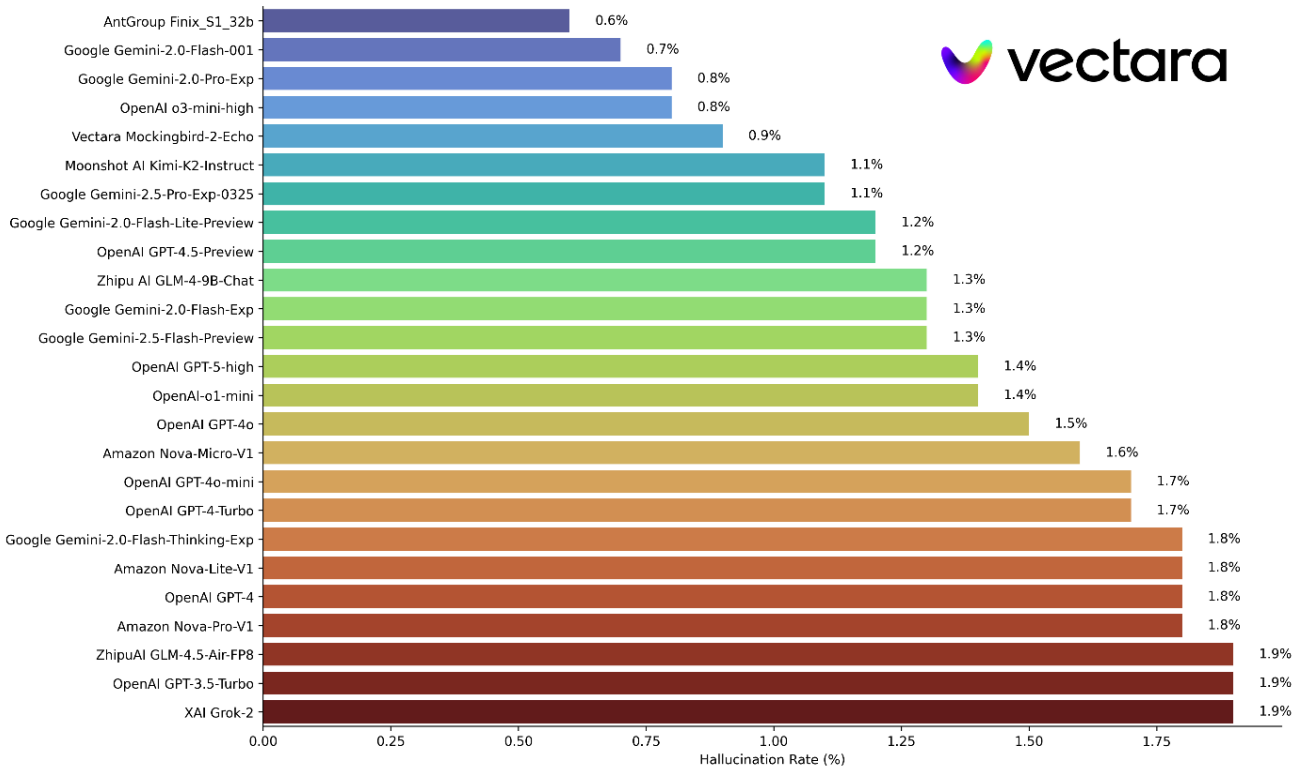

Abbildung 8: Öffentliche Rangliste von Halluzinationsraten verbreiteter Sprachmodelle (Stand 16. Oktober 2025)[79]

Zu den Modellen mit den derzeit niedrigsten Raten gehören AntGroup FiniX S1_32b (0,6 %), Google Gemini-2.0-Flash-001 (0,7 %) und Google Gemini-2.0-Pro-Exp (0,8 %). Am oberen Ende liegen Modelle wie OpenAI GPT-3.5-Turbo, Zhipu GLM-4.5-Air FP8 oder XAI Grok-2 mit jeweils 1,9 Prozent. Die Ergebnisse verdeutlichen, dass auch moderne Systeme bei dokumentenbasierten Aufgaben fehleranfällige Muster zeigen können.

79 https://github.com/vectara/hallucination-leaderboard?tab=readme-ov-file.

Eine gemeinsame Untersuchung von BBC und Europäischer Rundfunkunion zeigt, dass KI-Assistenten Nachrichteninhalte systematisch fehlerhaft wiedergeben.[80] In über 3.000 Antworten aus 18 Ländern und 14 Sprachen fanden sich plattformübergreifend ungenaue Quellen, erfundene Details und veraltete Informationen. Die ergänzende Auswertung zeigt das Ausmaß: 45 Prozent der Antworten enthielten mindestens einen relevanten Fehler, überwiegend aufgrund fehlerhafter Quellenangaben (31 Prozent) oder inhaltlicher Ungenauigkeiten (20 Prozent). Die Ergebnisse weisen auf ein durchgängig hohes Fehlerrisiko bei aktuellen, recherchierten KI-Zusammenfassungen hin.

Der International AI Safety Report (2025)[81] zeigt, dass Halluzinationen in professionellen Aufgabenfeldern auftreten können, in denen generative Modelle komplexe fachliche Arbeiten übernehmen. Im **Software-Engineering** unterstützen LLMs zwar bei Codegenerierung und -tests, erzeugen jedoch häufig fehlerhaften oder sicherheitskritischen Code. Studien belegen, dass Teams mit KI-Unterstützung teils mehr Sicherheitslücken produzieren und diese oft nicht erkennen. Im **juristischen Bereich** erzielen Modelle in standardisierten Tests teilweise Spitzenergebnisse, brechen jedoch unter leicht veränderten Bedingungen oder in realen Fällen deutlich ein. Dies führte bereits zu Fehlentscheidungen, etwa durch ungeprüft übernommene, erfundene Präzedenzfälle. Auch in der **Medizin** zeigt sich ein ähnliches Muster: Gute Testergebnisse kontrastieren mit begrenzter praktischer Verlässlichkeit, etwa durch fehlerhafte Schlussfolgerungen oder fehlendes Kontextverständnis. Wenn menschliche Kontrolle und Umgang mit Systemen unzureichend sind, können Halluzinationen nicht nur sporadisch auftreten, sondern ein strukturelles Risiko darstellen.

4.3. Kategorien von Halluzinationen und Falschinformationen

In der Forschung existieren verschiedene Ansätze, um Halluzinationen generativer KI-Systeme systematisch zu klassifizieren. Die Kategorisierungen setzen jeweils unterschiedliche Schwerpunkte und verdeutlichen, dass Halluzinationen mehrere klar unterscheidbare Formen annehmen können.

Halluzination können entlang zweier Grundtypen beschrieben werden: „Factuality Hallucination“ bezeichnet Abweichungen von überprüfbaren Fakten, während „Faithfulness Hallucination“ Divergenzen gegenüber Nutzeranweisungen, dem bereitgestellten Kontext oder der inneren Konsistenz des Modells umfasst.[82] „Faithfulness“ Halluzinationen werden weiter unterteilt in Instruktionsinkonsistenz, Kontextabweichungen und logische Widersprüche.

---

80 Ipsos & BBC (2025), Audience Use and Perceptions of AI Assistants for News Research Findings, https://www.bbc.co.uk/aboutthebbc/documents/audience-use-and-perceptions-of-ai-assistants-for-news.pdf.

81 AI Action Summit (2025), International AI Safety Report - The International Scientific Report on the Safety of Advanced AI, https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf.

82 Huang, Yu, Ma et al. (2025), A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, ACM Trans. Inf. Syst. 43, 2, Article 42, https://doi.org/10.1145/3703155 (Pre-Print Nov. 2024 https://arxiv.org/pdf/2311.05232).

Darüber hinaus existieren feingranulare Klassifikationen, die typische Fehlermuster der faktischen Ausgabegenerierung adressieren. Die Typologie von Rawte, Chakraborty, Pathak et al. (2023)[83] zielt darauf ab, wiederkehrende Fehlermuster in faktualen Ausgaben systematisch zu erfassen. Sie unterscheidet sechs Kategorien:

- Numeric Nuisance – fehlerhafte numerische Angaben.
- Acronym Ambiguity – falsche oder unpräzise Auflösungen von Akronymen.
- Generated Golem – Erfindung nicht-existierender Akteure oder Rollen.
- Virtual Voice – Zuschreibung nicht belegbarer Aussagen.
- Geographic Erratum – falsche geografische Zuordnungen.
- Time Wrap – Vermischung oder Fehlordnung zeitlicher Abläufe.

In einer weiteren Kategorisierung werden Halluzinationen aus Nutzersicht unterteilt. Diese Perspektive bildet das Spektrum der im praktischen Einsatz am häufigsten wahrgenommenen Ausfallformen ab (Tabelle 2):

Tabelle 2: Taxonomie von durch Nutzer gemeldeten LLM-Halluzinationen in KI-Mobil-Apps (ergänzt durch den Autor)[84]

Hallucination Type	**Distribution of hallucination types**	**Definition**	**Anonymized Example from User Review**
Factual Incorrectness	38 %	LLM provides information that is verifiably false or contradicts established facts relevant to the app's domain or general knowledge.	"The AI travel guide said Paris is the capital of England. That's just wrong!"
Fabricated Information / Invention	15 %	LLM generates details, entities, features, or sources that are entirely non-existent or not present within the app's context or reality.	"My AI recipe app invented a spice called 'solar-salt' for a simple pasta dish."
Nonsensical / Irrelevant Output	25 %	LLM produces responses that are grammatically sound but semantically meaningless, incoherent, repetitive, or	"I asked the AI story generator for a sci-fi plot

83 Rawte, Chakraborty, Pathak, et al. (2023), The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, S. 2541–2573, https://aclanthology.org/2023.emnlp-main.155.pdf.

84 Massenon, Gambo, Khan et al. (2025), My AI is Lying to Me: User-reported LLM hallucinations in AI mobile apps reviews, Scientific Reports, 15(1), 30397, https://doi.org/10.1038/s41598-025-15416-8.

Hallucination Type	Distribution of hallucination types	Definition	Anonymized Example from User Review
		completely off-topic to the user's query or interaction.	and it just gave me a list of farm animals."
Logical Inconsistency / Self-Contradiction	10 %	LLM's output contains statements that contradict each other within the same response, across a short conversational turn, or demonstrates clearly flawed reasoning.	"The AI first told me the event was on Saturday, then later insisted it was on Tuesday."
Persona Deviation / Role Inconsistency	5 %	The AI's responses deviate significantly from its established persona, intended role within the app, or the expected tone, potentially using inappropriate or unexpected language.	"The professional AI assistant for my work app suddenly started using slang and emojis."
Visual Fabrication (Generative AI)	4 %	Specific to generative AI tools (e.g., image, avatar generators), where the output contains elements that are physically impossible, bizarrely malformed, or violate common sense visual constraints.	"The AI avatar creator gave my character three hands and a floating hat. Looked ridiculous."
Repetitive Output (Non-functional)	3 %	LLM gets stuck in a loop, repeating the same phrase, sentence, or set of characters nonsensically and without progression, often indicating a failure state.	"When the AI got confused, it just kept typing 'hello hello hello hello' endlessly."

Deutlich breiter betrachtet der „International AI Safety Report"[85] Halluzinationen als Teil verschiedener Zuverlässigkeitsprobleme von generativen KI-Systemen. Zu diesen Defiziten zählen etwa Konfabulationen, Alltagswissensfehlern oder Probleme im Umgang mit kontextrelevanter, aktueller oder unverzerrter Information (Abbildung 9):

85 AI Action Summit (2025), International AI Safety Report - The International Scientific Report on the Safety of Advanced AI, https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf, S. 88.

Type of reliability issues	Examples
Confabulations or hallucinations	• Citing non-existent precedent in legal briefs *(451)* • Citing non-existent reduced fare policies for bereaved passengers *(452)*
Common-sense reasoning failures	• Failing to perform basic mathematical calculations *(453*)* • Failing to infer basic causal relationships *(454)*
Contextual knowledge failures	• Providing inaccurate medical information *(448)* • Providing outdated information about events *(455)*

Abbildung 9: Allgemeine KI weißt eine Vielzahl von Zuverlässigkeitsproblemen auf[86]

Diese verschiedenen Kategorisierungen verdeutlichen, dass Halluzinationen nicht eindimensional auftreten, sondern in mehreren differenzierten Formen sichtbar werden.

4.4. Ursachen

Die Ursachen von Halluzinationen hängen mit der Funktionsweise von Basismodellen (engl. „Foundation Model“) zusammen. Ein Foundation Model bezeichnet ein großes neuronales Netz, das auf umfangreichen und heterogenen Datensätzen vortrainiert wurde, um anschließend auf bestimmte Wissensgebiete adaptiert („fine-tuned“) oder in unterschiedlichen Anwendungen eingesetzt zu werden.[87]

Stark zusammengefasst bildet die sog. **Transformer-Architektur** das technische Fundament moderner Foundation-Modelle.[88] Sie dient als leistungsfähiges neuronales Netz, das Muster erkennt und Informationen über große Distanzen hinweg verknüpft. Zentrales Element ist die **Selbstaufmerksamkeit (Self-Attention)**. Operativ analysiert „das Modell […] jedes Element (z. B. ein Wort) in Beziehung zu allen anderen Elementen. Dabei gewichtet es, wie wichtig die anderen Elemente für das Verständnis sind. So entstehen Verknüpfungen zwischen zusammenhängenden Elementen.“[89]Am Ende entsteht ein feingliedriges Bedeutungsnetz, das semantische und strukturelle Zusammenhänge präzise abbildet.

Aufgrund der „Black Box“-Eigenschaften generativer KI-Systemen und intransparenten Trainingsbedingungen großer kommerzielle KI-Systeme werden folgende Ursachen für Halluzinationen diskutiert:[90]

---

86 Ebd., S. 90.

87 IBM (o.D.), What are foundation models?, https://www.ibm.com/think/topics/foundation-models.

88 Vgl. im Folgenden Bertelsmann Stiftung (o.D.), Das Ökosystem der KI-Basismodelle: Wie Daten, Energie und menschliche Arbeit KI-Basismodelle formen, reframe[Tech], https://www.reframetech.de/wissensseite-basismodelle/.

89 Ebd.

90 Kalai, Nachum, Vempala, et al. (2025), Why language models hallucinate. arXiv preprint arXiv:2509.04664, https://arxiv.org/pdf/2509.04664.

- Da das Modell Wahrscheinlichkeiten für nächste Token maximiert, entsteht ein Druck auf Vollständigkeit und Konsistenz – es „muss“ eine Antwort liefern, selbst wenn stichhaltige Fakten fehlen.

- In den Trainings- und Auswertungsverfahren werden häufig Raten für „korrekte Antwort“ honoriert, nicht hingegen das Zurückweisen einer Antwort bei Unsicherheit. Das führt zu einem Verhalten, das Spekulationen begünstigt.

- Unterschiedliche Datenqualität, fehlende oder seltene Fakten im Trainingsdatensatz sowie systematische Verzerrungen (Bias) tragen zur Entstehung falscher Aussagen bei.[91]

- Technische Limitationen wie das maximale Kontextfenster (Token-Limit) gelten als weitere Ursache für Halluzinationen. Mit zunehmender Länge des Inputs und das Erreichen der Kontextgrenze nimmt die Genauigkeit vieler Modelle deutlich ab.[92]

4.5. Minimierung von Halluzination

Die Reduzierung von Halluzinationen setzt ein abgestuftes Vorgehen voraus: Erst die verlässliche Erkennung und Bewertung solcher Fehler ermöglicht wirksame Mitigationstechniken. Forschung und Praxis unterscheiden daher konsistent zwischen „Detection“, „Benchmarking“ und verschiedenen Formen von „Mitigation“ (Abbildung 10):

91 IBM (o.D.), What are foundation models?, https://www.ibm.com/think/topics/foundation-models.

92 Huang, Yu, Ma et al. (2025), A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, ACM Trans. Inf. Syst. 43, 2, Article 42, https://doi.org/10.1145/3703155 (Pre-Print Nov. 2024 https://arxiv.org/pdf/2311.05232).

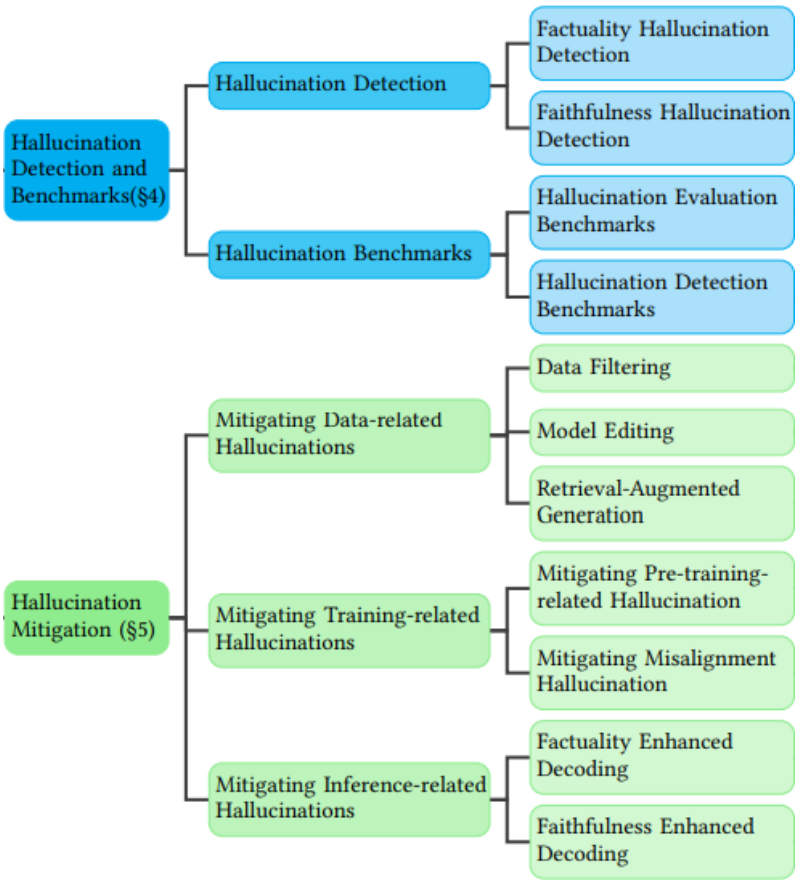

Abbildung 10: Detektion, Benchmarking und Mitigation zur Reduzierung von Halluzinationen[93]

Zentrales Fundament ist die **Detektion**. Methoden zur Erkennung von Halluzinationen orientieren sich an den zugrunde liegenden Fehlertypen. „Factuality Detection" prüft die Übereinstimmung mit verifizierbaren Fakten, typischerweise über externe Faktensysteme oder Unsicherheitsanalysen. „Faithfulness Detection" bewertet dagegen die Kohärenz der Ausgaben mit Nutzeranweisungen, Quellmaterial oder interner Logik. Dabei setzen die Verfahren auf unterschiedliche Ansätze: Sie vergleichen Fakten, nutzen trainierte Klassifikatoren oder Frage-Antwort-Modelle, und sie bewerten Unsicherheitssignale oder arbeiten mit gezielten Promptingstrategien.[94]

Darauf aufbauend ermöglichen **Halluzinationsbenchmarks** eine systematische Vergleichbarkeit. Sie messen entweder die Halluzinationsrate moderner LLMs oder die Leistungsfähigkeit von

93 Huang, Yu, Ma et al. (2025), A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, ACM Trans. Inf. Syst. 43, 2, Article 42, https://doi.org/10.1145/3703155 (Pre-Print Nov. 2024 https://arxiv.org/pdf/2311.05232).

94 Ebd.

Detektionsverfahren. Benchmarks bilden damit eine Grundlage für Modellvergleiche, Fortschrittsmessung und Evaluierung von Gegenmaßnahmen.[95]

Die Mitigation selbst umfasst drei Gruppen technischer Ansätze.[96]

- (1) Datenbezogene Maßnahmen adressieren fehlerhafte oder verzerrte Trainingsdaten. Dazu zählen Datenfilterung, das gezielte Aktualisieren von Modellparametern sowie der Einsatz externer Wissensquellen über Retrieval-Augmented Generation (RAG).

- (2) Trainingsbezogene Maßnahmen setzen direkt am Modellaufbau und am Lernprozess an, etwa durch Verbesserungen im Pre-Training, angepasstes Fine-Tuning oder Verfahren wie RLHF („Reinforcement Learning from Human Feedback“).[97]

- (3) Inferenzbezogene Maßnahmen setzen beim Generierungsprozess an, vor allem durch bessere Decodingstrategien, die die Genauigkeit und Kontexttreue der Ausgaben erhöhen.

Retrieval-Augmented Generation (RAG) stellt dabei einen eigenständigen Ansatz dar. Durch die Einbindung aktueller Wissensquellen kann das RAG das Faktenwissen deutlich erhöhen. Gleichzeitig bleiben Systeme anfällig, wenn Abrufmechanismen versagen oder generative Modelle abgerufene Inhalte nicht angemessen integrieren. Halluzinationen entstehen hier sowohl durch Retrieval-Fehler als auch durch Generierungsengpässe, etwa bei mangelnder Kontextausrichtung.[98]

Darüber hinaus sind **nutzerspezifische Strategien** von wesentlicher Bedeutung. Dazu zählen seitens der Nutzenden zusätzliche Konsistenzprüfungen der Ausgaben, Analysen der internen Modellzustände oder Verfahren zur Abschätzung der eigenen Antwortsicherheit. Solche Methoden helfen, unsichere Ausgaben zu erkennen und die Nutzung generativer Systeme kritischer zu gestalten.[99]

---

95 Huang, Yu, Ma et al. (2025), A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, ACM Trans. Inf. Syst. 43, 2, Article 42, https://doi.org/10.1145/3703155 (Pre-Print Nov. 2024 https://arxiv.org/pdf/2311.05232).; AI Action Summit (2025), International AI Safety Report - The International Scientific Report on the Safety of Advanced AI, https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf, S. 90.

96 Vgl. im Folgenden Huang, Yu, Ma et al. (2025), A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, ACM Trans. Inf. Syst. 43, 2, Article 42, https://doi.org/10.1145/3703155 (Pre-Print Nov. 2024 https://arxiv.org/pdf/2311.05232).

97 Massenon, Gambo, Khan et al. (2025), My AI is Lying to Me: User-reported LLM hallucinations in AI mobile apps reviews, Scientific Reports, 15(1), 30397, https://doi.org/10.1038/s41598-025-15416-8.

98 Huang, Yu, Ma et al. (2025), A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, ACM Trans. Inf. Syst. 43, 2, Article 42, https://doi.org/10.1145/3703155 (Pre-Print Nov. 2024 https://arxiv.org/pdf/2311.05232); Jones (2025), AI hallucinations can’t be stopped — but these techniques can limit their damage, Nature News Feature, https://www.nature.com/articles/d41586-025-00068-5.

99 Jones (2025), AI hallucinations can’t be stopped — but these techniques can limit their damage, Nature News Feature, https://www.nature.com/articles/d41586-025-00068-5.

Diese Verfahren unterstützen sowohl die Erkennung unsicherer Antworten als auch die kritische Nutzung generativer Systeme.

Aktuelle Analysen weisen jedoch auch auf **Grenzen der Mitigation** hin. Es existieren derzeit keine verlässlichen, umfassend wirksamen Methoden zur vollständigen Eliminierung von Halluzinationen. Die Effizienz vieler Techniken ist noch nicht ausreichend belegt, Benchmarks sind heterogen, und das Modellverhalten variiert stark über Aufgaben und Kontexte.

Experten fordern für eine nachhaltige Reduzierung robuste Evaluierungsprozesse, transparente Kommunikation über Modellgrenzen und eine Stärkung nutzerseitiger Kompetenz im Umgang mit generativer KI.[100]

***

100 AI Action Summit (2025), International AI Safety Report - The International Scientific Report on the Safety of Advanced AI, https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf, S. 91.