



MASTERARBEIT | MASTER'S THESIS

Titel | Title

Bias und Fairness in Künstlicher Intelligenz

verfasst von | submitted by

Omed Noory B.Sc.

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt |
Degree programme code as it appears on the
student record sheet:

UA 066 915

Studienrichtung lt. Studienblatt | Degree
programme as it appears on the student
record sheet:

Masterstudium Betriebswirtschaft

Betreut von | Supervisor:

ao. Univ.-Prof. i.R. Mag. Dr. Josef Windsperger

Abstract (English)

Algorithmic bias and fairness in artificial intelligence systems represent among the most pressing ethical and technical challenges in contemporary AI development. This thesis examines the current state of research on bias and fairness in AI systems through a systematic literature review (SLR) conducted in accordance with the PRISMA 2020 guidelines. Based on searches across four databases (PubMed/PMC, Springer Link, JSTOR, and arXiv), 66 studies were selected for final synthesis from 4,229 identified records. The analysis is structured around four thematic areas: bias in large language models (LLMs), bias in clinical AI, fairness metrics and their formal limitations, and mitigation strategies. Findings demonstrate that bias is systematically present across demographic dimensions in both LLMs and clinical models, and is traceable to identifiable mechanisms within the data and modelling pipeline. Formal impossibility theorems establish that no single fairness metric can simultaneously satisfy all relevant criteria. Mitigation strategies, while demonstrably effective, are inherently context-dependent and constrained by institutional conditions. The EU AI Act introduces the first binding regulatory framework, though it leaves considerable interpretive latitude regarding the choice of specific fairness metrics. The thesis concludes that durable progress on AI bias requires coordinated advancement across technical, methodological, and regulatory dimensions simultaneously.

Keywords: Algorithmic Bias, AI Fairness, Large Language Models, Healthcare AI, Systematic Literature Review, EU AI Act, Mitigation Strategies

Abstract (Deutsch)

Algorithmischer Bias und Fairness in Systemen der Künstlichen Intelligenz zählen zu den zentralen ethischen und technischen Herausforderungen der gegenwärtigen KI-Entwicklung. Die vorliegende Arbeit untersucht den aktuellen Forschungsstand zu Bias und Fairness in KI-Systemen mittels einer systematischen Literaturanalyse (SLR) nach dem PRISMA-2020-Standard. Auf Basis einer Recherche in vier Datenbanken (PubMed/PMC, Springer Link, JSTOR, arXiv) wurden aus 4.229 identifizierten Quellen 66 Studien in die finale Synthese eingeschlossen. Die Analyse gliedert sich in vier thematische Schwerpunkte: Bias in großen Sprachmodellen (LLMs), Bias in der klinischen KI, Fairness-Metriken und deren formale Grenzen sowie Mitigation-Strategien. Die Befunde zeigen, dass Bias über demografische Dimensionen hinweg systematisch in LLMs und klinischen Modellen nachweisbar ist und auf identifizierbare Mechanismen in der Datenpipeline zurückgeführt werden kann. Formale Unmöglichkeitstheoreme belegen, dass keine einzelne Fairness-Metrik alle relevanten Kriterien simultan erfüllen kann. Mitigation-Strategien sind zwar nachweislich wirksam, jedoch stets kontextabhängig und durch institutionelle Rahmenbedingungen begrenzt. Der EU AI Act setzt erste verbindliche regulatorische Leitplanken, lässt jedoch bei der konkreten Metrikwahl erheblichen Interpretationsspielraum. Die Arbeit kommt zu dem Schluss, dass eine nachhaltige Adressierung von KI-Bias die koordinierte Weiterentwicklung technischer, methodischer und regulatorischer Ansätze erfordert.

Schlüsselwörter: Algorithmischer Bias, Fairness in KI, Large Language Models, Gesundheits-KI, Systematische Literaturanalyse, EU AI Act, Mitigation-Strategien

Table of Contents

1 Einleitung	1
1.1 Einleitung.....	1
1.1.1 Problemstellung	2
1.1.2 Forschungsfragen	3
1.1.3 Signifikanz und Beitrag der Arbeit.....	3
1.2 Konzeptualisierung von Fairness	4
1.2.1 Fairness als Bündel unterschiedlicher Definitionen	5
1.2.2 Fairness in Zusammenhang mit Schadensformen	6
1.2.3 Kontextabhängigkeit und Instabilität von Fairness-Bewertungen	7
1.2.4 Fairness als Governance- und Regulierungsfrage	8
1.2.5 Zwischenfazit.....	9
2. Methodik	9
2.1 Forschungsdesign und Begründung	10
2.2 Forschungsfragen	10
2.3 Suchstrategie.....	11
2.4 Ein- und Ausschlusskriterien	12
2.5 Auswahlprozess und PRISMA Flow	13
2.5.1 Identifikation	14
2.5.2 Screening.....	14
2.5.3 Eignungsprüfung.....	15
2.5.4 Einschluss.....	15
2.6 Datenextraktion und Qualitätsbewertung	15
2.7 Synthese Ansatz.....	16
2.8 Limitationen der Methodik.....	16
3. Ergebnisse.....	17
3.1 Bias in Large Language Models	17
3.1.1 Konzeptionelle Grundlagen: Defintion von Bias in LLMs	17
3.1.2 Empirische Evidenz: Wie sich Bias in LLM-Ausgaben manifestiert	19
Gender Bias	19
Rassen- und ethnischer Bias	20

Religiöser Bias	21
Intersektionaler und sozioökonomischer Bias	21
3.1.3 Quellen von Bias in der LLM-Pipeline.....	22
3.1.4 Messung von Bias in LLMs: Benchmarks und Evaluationsrahmen	23
3.1.5 Zusammenfassung und Forschungslücken	25
3.2 Bias in Gesundheits-KI.....	26
3.2.1 Umfang und Tragweite von Bias in der klinischen KI.....	26
3.2.2 Soziodemografischer und rassenbezogener Bias in klinischen Modellen	27
3.2.3 Bias entlang der Pipeline elektronischer Patientenakten.....	29
Unüberwachtes Lernen: Eine unterbeforschte Grenze.....	30
3.2.4 Large Language Models in der klinischen Praxis: Eine aufkommende Bias-Front	30
3.2.5 Regulatorische und Governance-Dimensionen speziell für Gesundheits-KI.....	31
3.2.6 Zusammenfassung und Synthese	32
3.3 Fairness-Metriken und theoretische Grundlagen	33
3.3.1 Die Pluralität der Fairness-Definitionen	33
3.3.2 Formale Unvereinbarkeit von Fairness-Kriterien	34
3.3.3 Kausalbasierte Fairness-Konzepte: Kontrafaktische Ansätze.....	36
3.3.4 Empirische Anwendung von Fairness-Metriken im Gesundheitswesen	37
Über die klinische Vorhersage hinaus: Datensätze und Interpretierbarkeit.....	38
3.3.5 Zusammenfassung und Synthese	39
3.4 Mitigationsstrategien.....	39
3.4.1 Pre-processing: Bias auf der Datenebene adressieren	40
3.4.2 In-processing: Fairness-Constraints während des Modelltrainings	41
3.4.3 Post-Processing: Korrektur von Outputs nach dem Training	42
3.4.4 Organisatorische und soziale Dimensionen der Mitigation.....	43
3.4.5 Zusammenfassung und Synthese	45
4. Regulierung und Governance	46
4.1 Der EU AI Act: Ein risikobasiertes Regulierungsrahmenwerk.....	46
4.2 Technische Standards: ISO/IEC 23894.....	47
4.3 Sektorspezifische und institutionelle Governance-Initiativen.....	49

4.4 Politikempfehlungen und Best Practices.....	50
4.5 Zusammenfassung und Synthese	51
5. <i>Diskussion und Fazit</i>	52
5.1 Synthese der Ergebnisse über die Kapitel hinweg.....	52
5.2 Beantwortung der Forschungsfragen	53
5.3 Limitationen	54
5.3.1 Methodische Limitationen der Übersichtsarbeit	55
5.3.2 Limitationen der zugrunde liegenden Primärliteratur	56
5.4 Richtungen für künftige Forschung	56
5.5 Schlussbemerkungen.....	57
<i>Hinweis zur Verwendung generativer KI</i>	58
<i>Literaturverzeichnis</i>	59

1 Einleitung

1.1 Einleitung

Künstliche-Intelligenz-Systeme sind zunehmend in Entscheidungen eingebettet, die das Leben von Menschen unmittelbar beeinflussen: wem ein Kredit gewährt wird, welche Patienten vorrangig klinisch behandelt werden, welche Bewerber zu einem Vorstellungsgespräch eingeladen werden, und welche Angeklagten als rückfallgefährdet eingeschätzt werden. Mit der Ausweitung KI-gestützter Entscheidungsfindung ist auch die empirische Evidenz dafür gewachsen, dass diese Systeme nicht alle Individuen und Gruppen gleichbehandeln. Eine wachsende Zahl von Studien zeigt, dass algorithmische Systeme bestimmte demografische Gruppen systematisch benachteiligen können – etwa nach Geschlecht, ethnischer Zugehörigkeit, Alter oder sozioökonomischem Status – selbst ohne jede explizite diskriminierende Absicht der Entwickler. Dieses Phänomen, breit als algorithmischer Bias bezeichnet, hat sich zu einer der drängendsten ethischen und technischen Herausforderungen der KI-Verbreitung entwickelt.

Die Dringlichkeit dieser Herausforderung hat sich mit dem Aufkommen großer Sprachmodelle (Large Language Models, LLMs) als dominantes KI-Paradigma erheblich verschärft. Im Gegensatz zu früheren Generationen von Machine-Learning-Systemen, die typischerweise diskrete Klassifikationen oder numerische Risikowerte für klar definierte Aufgaben lieferten, generieren LLMs offene, flüssige Texte für ein nahezu unbegrenztes Spektrum an Anwendungen – von der klinischen Dokumentation über den Kundenservice bis hin zur juristischen Recherche und zum Bildungswesen. Diese Vielseitigkeit hat die rasche Verbreitung vorangetrieben, hat aber gleichzeitig die Erkennung, Messung und Korrektur von Bias erheblich erschwert: Dasselbe Modell kann verzerrte Ausgaben in nahezu unendlich vielen Oberflächenformen produzieren, und das Fehlen einer einzigen, diskreten Entscheidungsgrenze erschwert die direkte Anwendung von Fairness-Verfahren, die für klassisches überwachtes Lernen entwickelt wurden (Gallegos et al., 2024; Chu et al., 2024).

Das Gesundheitswesen stellt für diese Untersuchung einen Bereich von besonderer Bedeutung dar. KI-Systeme werden zunehmend entlang des gesamten klinischen Behandlungspfads eingesetzt – in der diagnostischen Bildgebung, der Risikostratifizierung, der klinischen Dokumentation und, zunehmend, in der LLM-basierten Entscheidungsunterstützung – zu

einem Zeitpunkt, an dem die Folgen algorithmischer Fehler außergewöhnlich gravierend sein können. Die in dieser Arbeit untersuchte empirische Evidenz zeigt, dass verzerrte KI-Ausgaben im Gesundheitswesen kein rein theoretisches Problem darstellen: Sie können klinische Entscheidungsfindung nachweislich direkt verschlechtern, systematisch schlechter für Bevölkerungsgruppen funktionieren, die bereits eingeschränkten Zugang zur Gesundheitsversorgung haben, und historisch gewachsene Verteilungsungleichheiten aus klinischen Daten in gegenwärtige Diagnose- und Behandlungsempfehlungen fortschreiben. Die Schnittstelle zwischen großen Sprachmodellen und KI im Gesundheitswesen bildet damit einen Bereich, in dem die allgemeine Herausforderung algorithmischer Fairness gleichzeitig technisch komplexer und folgenreicher wird.

1.1.1 Problemstellung

Trotz einer rasch wachsenden Forschungsbasis zu algorithmischem Bias und Fairness bleibt die Literatur entlang mehrerer Dimensionen fragmentiert. Erstens ist Fairness selbst kein einheitlich vereinbartes Konzept, sondern eine Familie konkurrierender Definitionen, die jeweils unterschiedliche normative Grundannahmen darüber widerspiegeln, was gerechte Behandlung ausmacht (Hellman, 2020; Pessach & Shmueli, 2022) – eine Pluralität, die die Synthese von Befunden über Studien mit unterschiedlichen Fairness-Kriterien erschwert. Zweitens haben sich die Forschung zu Bias in großen Sprachmodellen und die Forschung zu Bias im klinischen maschinellen Lernen weitgehend als parallele Literaturstränge entwickelt, wobei vergleichsweise wenige Studien die allgemeine LLM-Fairness-Literatur systematisch mit ihren spezifischen Ausprägungen und Konsequenzen im Gesundheitswesen verknüpfen. Drittens ist die regulatorische und Governance-Antwort auf algorithmischen Bias – exemplarisch verkörpert durch den EU AI Act und ergänzende technische Standards – zeitgleich mit, statt im Anschluss an, einen stabilen wissenschaftlichen Konsens darüber entstanden, wie Fairness gemessen werden sollte, wodurch potenzielle Lücken zwischen regulatorischen Anforderungen und dem technischen Stand der Forschung entstehen.

Diese Arbeit adressiert diese Lücken durch eine systematische Literaturanalyse, die den aktuellen Forschungsstand zu Bias und Fairness in der Künstlichen Intelligenz synthetisiert, mit besonderem Augenmerk auf große Sprachmodelle und deren Anwendung im Gesundheitswesen. Anstatt Bias als ein einheitliches technisches Problem zu behandeln, das durch einen einzigen methodischen Ansatz gelöst werden kann, betrachtet diese Untersuchung

Fairness als ein mehrdimensionales Konstrukt, das gleichzeitig aus konzeptioneller, empirischer, methodischer und regulatorischer Perspektive untersucht werden muss.

1.1.2 Forschungsfragen

In Übereinstimmung mit dem in Kapitel 2 dargelegten Protokoll der systematischen Literaturanalyse orientiert sich diese Arbeit an vier Forschungsfragen: (RQ1) Welche Arten von Bias wurden in KI- und Machine-Learning-Systemen identifiziert, und wie werden sie in der aktuellen Literatur definiert und kategorisiert? (RQ2) Welche Fairness-Metriken und Evaluationsrahmen wurden vorgeschlagen, und welche jeweiligen Zielkonflikte und Limitationen weisen sie auf? (RQ3) Welche Mitigation-Strategien wurden zur Adressierung von Bias in KI-Systemen entwickelt, und welche Evidenz existiert zu ihrer Wirksamkeit? (RQ4) Wie adressieren aktuelle regulatorische Rahmenwerke – insbesondere der EU AI Act – Bias- und Fairness-Anforderungen, und welche Lücken bestehen weiterhin?

1.1.3 Signifikanz und Beitrag der Arbeit

Diese Arbeit trägt auf drei zentrale Weisen zur bestehenden Literatur bei. Erstens liefert sie eine integrierte Synthese, die explizit die allgemeine Literatur zu Bias und Fairness in großen Sprachmodellen mit deren spezifischen Ausprägungen im hochriskanten Bereich der Gesundheits-KI verknüpft – zwei Literaturstränge, die sich trotz ihrer offensichtlichen gegenseitigen Relevanz weitgehend parallel entwickelt haben. Zweitens bietet diese Arbeit durch die Anwendung einer systematischen, PRISMA-konformen Review-Methodik (Page et al., 2021) über vier Datenbanken aus Informatik, Medizin und Philosophie/Ethik eine methodisch transparentere und reproduzierbarere Synthese, als dies bei narrativen Übersichtsarbeiten in diesem sich rasch entwickelnden Feld üblich ist. Drittens versucht diese Arbeit, durch die explizite Verknüpfung der technischen Literatur zu Fairness-Metriken und Mitigation-Strategien mit den regulatorischen Anforderungen des EU AI Acts und ergänzender internationaler Standards eine Lücke zu schließen, die sowohl in der rein technischen als auch in der rein rechtswissenschaftlichen Literatur zu KI-Fairness häufig implizit bleibt – nämlich die Frage, ob die gegenwärtige technische Fähigkeit zur Bias-Erkennung und -Mitigation den entstehenden regulatorischen Erwartungen genügt, und wo signifikante Lücken bestehen.

1.1.4 Struktur der Arbeit

Der weitere Verlauf dieser Arbeit ist wie folgt gegliedert. Kapitel 1 setzt sich mit einer konzeptionellen Behandlung von Fairness als mehrdimensionalem und kontextabhängigem Konstrukt fort (Abschnitt 1.2) und etabliert damit das analytische Vokabular, das in den nachfolgenden Kapiteln durchgängig verwendet wird. Kapitel 2 stellt die Methodik dar und beschreibt detailliert das Protokoll der systematischen Literaturanalyse, die Suchstrategie sowie den Screening-Prozess über vier Datenbanken (PubMed/PMC, Springer Link, JSTOR und arXiv), der in einer finalen Stichprobe von 66 für die thematische Synthese ausgewählten Studien mündet. Kapitel 3 präsentiert die inhaltlichen Befunde der Untersuchung, gegliedert in vier Abschnitte: Bias in großen Sprachmodellen (3.1), Bias speziell in der Gesundheits-KI (3.2), Fairness-Metriken und deren theoretische Grundlagen einschließlich der formalen Unmöglichkeitsresultate, die die gleichzeitige Erfüllung mehrerer Fairness-Kriterien einschränken (3.3), sowie Mitigation-Strategien entlang der Pre-Processing-, In-Processing-, Post-Processing- und organisatorischen Dimensionen der KI-Entwicklungspipeline (3.4). Kapitel 4 untersucht die regulatorische und Governance-Landschaft mit besonderem Augenmerk auf den EU AI Act, ISO/IEC 23894 sowie sektorspezifische Governance-Initiativen, die in der Untersuchung identifiziert wurden. Kapitel 5 schließt mit einer Synthese der Befunde über alle Kapitel hinweg, beantwortet die vier Forschungsfragen direkt, diskutiert die Limitationen dieser Untersuchung und identifiziert Richtungen für künftige Forschung.

1.2 Konzeptualisierung von Fairness

Fairness in der Künstlichen Intelligenz ist kein einheitlich definiertes Konzept. Vielmehr fungiert der Begriff in der wissenschaftlichen Literatur als Sammelbezeichnung für eine Reihe unterschiedlicher normativer Zielvorstellungen, die jeweils durch verschiedene technische Definitionen, Metriken und Evaluationsansätze operationalisiert werden. Diese begriffliche Pluralität ist im Kontext großer Sprachmodelle (Large Language Models, LLMs) besonders ausgeprägt, da deren Ausgaben offen, kontextabhängig und stark durch Eingabeaufforderungen geprägt sind, anstatt sich auf diskrete Klassifikationsentscheidungen zu beschränken. Entsprechend wird Fairness in aktuellen Arbeiten zunehmend als kontextabhängiges Konstrukt verstanden, das mit Bezug auf eine konkrete Anwendung, eine bestimmte Art potenziellen

Schadens und die davon betroffenen Interessengruppen spezifiziert werden muss (Chu et al., 2024; Gallegos et al., 2024).

Ein sinnvoller Ausgangspunkt für diese Diskussion ist die Unterscheidung zwischen Fairness als ethischem oder rechtlichem Ideal und Fairness als technisch operationalisierbarem Kriterium. Normative Grundsätze wie Nicht-Diskriminierung, Gleichbehandlung oder Gerechtigkeit müssen in messbare Zielgrößen übersetzt werden, bevor sie algorithmisch überprüfbar werden. Diese Übersetzung ist jedoch nicht wertneutral. Die Wahl einer bestimmten Fairness-Definition kodiert implizit Annahmen darüber, welche Unterschiede zwischen Gruppen oder Individuen als ungerecht gelten, welche Ungleichheiten tolerierbar sind und welche Zielkonflikte zwischen konkurrierenden Gütern akzeptabel sind. Aus philosophischer Perspektive sind Fairness-Metriken daher stets normativ fundiert und nicht rein technisch bestimmbar (Hellman, 2020). Für eine systematische Literaturanalyse bedeutet dies, dass nicht nur empirische Befunde, sondern auch die zugrunde liegenden Fairness-Konzeptionen jeder Studie vergleichend betrachtet und explizit gemacht werden müssen.

1.2.1 Fairness als Bündel unterschiedlicher Definitionen

Die technische Literatur strukturiert Fairness häufig als Familie verwandter, aber unterschiedlicher Definitionen, die jeweils einen anderen Gleichheitsbegriff betonen. Obwohl Terminologie und Schwerpunktsetzung zwischen Übersichtsarbeiten variieren, unterscheiden die meisten zwischen Gruppen-Fairness und Individual-Fairness, wobei kontrafaktische oder kausale Fairness-Konzeptionen zunehmend als dritte Kategorie diskutiert werden (Chu et al., 2024; Mehrabi et al., 2021).

Gruppenbezogene Ansätze zielen im Allgemeinen auf statistische Gleichheit zwischen vordefinierten demografischen Gruppen ab – etwa hinsichtlich Fehlerraten oder Raten positiver Klassifikationsergebnisse –, wobei die relevanten sensiblen Merkmale Geschlecht, ethnische Zugehörigkeit, Alter oder Sprache umfassen können. Individual-Fairness, ursprünglich von Dwork et al. (2012) formalisiert, fordert stattdessen, dass ähnliche Individuen ähnlich behandelt werden, was eine explizite, aufgabenspezifische Ähnlichkeitsmetrik voraussetzt, deren Definition selbst eine normative und keine rein technische Entscheidung ist. Kontrafaktische Ansätze nach Kusner et al. (2017) untersuchen, ob sich die Ausgabe eines Modells ändern würde, wenn ein sensibles Merkmal hypothetisch verändert würde, während alle kausal unabhängigen Faktoren konstant blieben – eine Formulierung, die ein korrekt

spezifiziertes Kausalmodell des zugrunde liegenden datengenerierenden Prozesses voraussetzt, eine Annahme, die in der Praxis selten mit Sicherheit verifizierbar ist.

Tabelle 1.1: Hauptfamilien von Fairness-Definitionen

Familie	Grundidee	Zentrale Einschränkung
Group Fairness	Statistische Gleichheit zwischen vordefinierten demografischen Gruppen (z. B. gleiche Fehlerraten)	Setzt klare Gruppengrenzen voraus; vernachlässigt Variation innerhalb von Gruppen
Individual Fairness	Ähnliche Individuen sollen ähnlich behandelt werden (Dwork et al., 2012)	Erfordert eine explizite, selbst normativ geladene Ähnlichkeitsmetrik
Counterfactual Fairness	Ergebnis bleibt gleich in einer hypothetischen Welt ohne das sensible Attribut (Kusner et al., 2017)	Setzt ein korrektes Kausalmodell voraus, das selten vollständig verifizierbar ist

Anmerkung. Die drei Familien schließen sich nicht gegenseitig aus; viele Studien kombinieren Elemente mehrerer Ansätze (vgl. Pessach & Shmueli, 2022; Abschnitt 3.3).

Die Anwendung dieser Definitionen im Kontext großer Sprachmodelle stellt besondere Herausforderungen dar. LLM-Ausgaben bestehen häufig nicht aus diskreten Entscheidungswerten, sondern aus offenen Texten mit impliziter Bedeutung, Framing oder Stereotypisierung. Viele Studien greifen daher auf indirekte Messverfahren zurück, etwa Toxizitätswerte, Stereotyp-Assoziationstests oder Unterschiede darin, wie generierter Text bestimmte demografische Gruppen beschreibt (Chu et al., 2024; Gallegos et al., 2024). Dies erschwert die Verbindung zwischen normativem Ideal und messbarem Indikator erheblich und macht sie deutlich stärker interpretationsabhängig als in klassischen Klassifikationsszenarien, in denen ein Modell eine einzelne, diskrete Vorhersage produziert, die direkt zwischen Gruppen verglichen werden kann.

1.2.2 Fairness in Zusammenhang mit Schadensformen

Ein wiederkehrendes Motiv in der Literatur ist die explizite Verbindung von Fairness-Konzepten mit unterschiedlichen Schadens kategorien. Gallegos et al. (2024) und Mehrabi et

al. (2021) unterscheiden insbesondere zwischen repräsentativen Schäden (representational harms) und allokativen Schäden (allocational harms). Repräsentative Schäden entstehen, wenn bestimmte Gruppen stereotyp, abwertend oder einseitig dargestellt werden, unabhängig von einer konkreten Entscheidung über ein Individuum. Allokative Schäden treten auf, wenn algorithmische Systeme die Verteilung von Ressourcen, Chancen oder Dienstleistungen beeinflussen – etwa im Gesundheitswesen oder am Arbeitsmarkt – und damit unmittelbar materielle Ergebnisse für die betroffenen Individuen beeinflussen.

Diese Unterscheidung ist für LLMs besonders relevant, da textbasierte Systeme sowohl repräsentative als auch allokativen Effekte gleichzeitig entfalten können. Selbst wenn ein Modell keine explizite Entscheidung trifft, kann die Art und Weise, wie Informationen zusammengefasst oder formuliert werden, nachgelagerte menschliche Entscheidungen materiell beeinflussen – etwa die Interpretation einer KI-generierten Patientenzusammenfassung durch eine Ärztin oder einen Arzt. Im Gesundheitskontext wird diese Dynamik besonders deutlich: Empirische Evidenz zeigt, dass Bias in textgenerierenden Systemen systematisch beeinflussen kann, wie Patientinnen und Patienten beschrieben werden, etwa durch unterschiedliche Schwerpunktsetzung in generierten klinischen Narrativen (Rickman, 2025). Systematische Übersichtsarbeiten weisen zudem darauf hin, dass unausgewogene Trainingsdaten im Gesundheitswesen zu messbaren Leistungsunterschieden zwischen Patientengruppen führen können, was ein unmittelbares Fairness-Problem mit allokativen Konsequenzen darstellt (Chen et al., 2024; Hasanzadeh et al., 2025; vgl. Abschnitt 3.2).

Die in Abschnitt 1.2.1 diskutierte Pluralität an Fairness-Definitionen lässt sich somit teilweise als Reaktion auf diese Pluralität an Schadensszenarien verstehen: Je nach Anwendungsbereich und Art des potenziellen Schadens werden von Forschenden und Praktikern unterschiedliche normative Maßstäbe priorisiert (Chu et al., 2024; Gallegos et al., 2024).

1.2.3 Kontextabhängigkeit und Instabilität von Fairness-Bewertungen

Ein weiterer bedeutsamer Befund der LLM-Fairness-Forschung betrifft die Kontextabhängigkeit von Fairness-Bewertungen. Empirische Studien zeigen, dass Fairness-Metriken je nach spezifischem Prompt, Datensatz oder Evaluationsumgebung erheblich variieren können (Chu et al., 2024; Gallegos et al., 2024). Dies hat eine wichtige konzeptionelle Konsequenz: Fairness ist nicht ausschließlich eine fixe Eigenschaft eines isoliert betrachteten

Modells, sondern entsteht vielmehr aus dem Zusammenspiel zwischen Modell, Daten, Evaluationsverfahren und Anwendungskontext.

Aktuelle Benchmark-Entwicklungen spiegeln eine wachsende Anerkennung dieser Instabilität wider. Jung et al. (2025) argumentieren mit dem FLEX-Benchmark, dass eine valide Fairness-Bewertung die Robustheit unter variierenden Bedingungen explizit testen muss, um stabile und verallgemeinerbare Schlussfolgerungen zu ermöglichen, anstatt sich auf ein einziges fixes Evaluationssetup zu verlassen. Ihre empirischen Befunde – ausführlicher diskutiert in Abschnitt 3.1.4 – zeigen, dass Modelle, die unter einem Satz von Evaluationsbedingungen fair erscheinen, diese Eigenschaft häufig nicht aufrechterhalten, wenn geringe Variationen in der Formulierung des Prompts vorgenommen werden. Für die vorliegende Untersuchung bedeutet dies, dass Fairness-Definitionen nicht isoliert betrachtet werden können, sondern gemeinsam mit den spezifischen Evaluationspraktiken betrachtet werden müssen, durch die sie in jeder Primärstudie gemessen wurden.

1.2.4 Fairness als Governance- und Regulierungsfrage

Über ihre technische und konzeptionelle Dimension hinaus wird Fairness zunehmend als Governance-Frage diskutiert. Dies zeigt sich besonders deutlich in Hochrisikobereichen wie dem Gesundheitswesen, wo organisatorische und regulatorische Rahmenwerke eine zentrale Rolle bei der Übersetzung normativer Fairness-Verpflichtungen in durchsetzbare Praxis spielen. Der EU AI Act verfolgt einen risikobasierten Regulierungsansatz und etabliert Anforderungen an Datenqualität, Risikomanagement und Dokumentation, die unmittelbar mit Fairness-Erwägungen verknüpft sind (Europäische Union, 2024; vgl. Abschnitt 4.1). Ergänzend zu diesem verbindlichen Rechtsrahmen betont ISO/IEC 23894:2023 die Notwendigkeit systematischer Risikoidentifikation und -überwachung über den gesamten Lebenszyklus eines KI-Systems hinweg (ISO, 2023; vgl. Abschnitt 4.2), während politikorientierte Forschung zusätzlich organisatorische Verantwortlichkeitsstrukturen und Best Practices für Fairness-Governance hervorhebt (Alvarez et al., 2024; vgl. Abschnitt 4.4).

Diese Governance-Perspektiven bekräftigen einen bereits in Abschnitt 1.2.1 konzeptionell etablierten Punkt: Fairness kann nicht als rein technisches Optimierungsproblem verstanden werden, das einmal gelöst und universell angewendet wird. Vielmehr handelt es sich um ein sozio-technisches Konstrukt, das normative Annahmen, technische

Umsetzungsentscheidungen und institutionelle Rahmenwerke miteinander verbindet (Hellman, 2020).

1.2.5 Zwischenfazit

Zusammenfassend zeigt die Literatur, dass Fairness in der KI-Forschung kein einheitlich abschließend definiertes Konzept ist, sondern ein kontextabhängiges und normativ geprägtes Konstrukt. Unterschiedliche Definitionen reflektieren unterschiedliche Annahmen über Schaden, Anwendungsbereich und Bewertungsmaßstab (Chu et al., 2024; Mehrabi et al., 2021). Im spezifischen Kontext großer Sprachmodelle wird die Fairness-Bewertung durch die Offenheit und Kontextsensitivität der Modellausgaben zusätzlich erschwert (Gallegos et al., 2024; Jung et al., 2025). Gleichzeitig verankern regulatorische und organisatorische Rahmenwerke Fairness zunehmend in formalen Governance-Strukturen, insbesondere in sensiblen Bereichen wie dem Gesundheitswesen (Europäische Union, 2024; ISO, 2023; Chen et al., 2024).

Diese konzeptionelle Grundlage etabliert das analytische Vokabular, das im weiteren Verlauf dieser Arbeit verwendet wird: Kapitel 3.1 untersucht, wie diese Fairness-Definitionen empirisch auf große Sprachmodelle angewendet werden, Kapitel 3.2 erweitert diese Analyse spezifisch auf das Gesundheitswesen, Kapitel 3.3 bietet eine formale Behandlung von Fairness-Metriken und ihrer mathematischen Unvereinbarkeiten, und Kapitel 4 greift die in Abschnitt 1.2.4 eingeführte Governance-Dimension in deutlich größerer Tiefe wieder auf.

2. Methodik

Dieses Kapitel skizziert den methodischen Rahmen, der in dieser Untersuchung angewendet wurde. Angesichts der Breite und der raschen Entwicklung des Forschungsfeldes wurde eine systematische Literaturanalyse (Systematic Literature Review, SLR) als primäres Forschungsdesign gewählt. Dieser Ansatz ermöglicht eine transparente, reproduzierbare und umfassende Synthese des bestehenden wissenschaftlichen Wissensstands zu Bias und Fairness in der Künstlichen Intelligenz. Die Methodik orientiert sich an etablierten Leitlinien für systematische Übersichtsarbeiten in der Softwareentwicklung (Kitchenham & Charters, 2007) und wird in Übereinstimmung mit dem PRISMA-2020-Statement berichtet (Page et al., 2021).

2.1 Forschungsdesign und Begründung

Die systematische Literaturanalyse wurde aus mehreren Gründen als methodischer Ansatz gewählt. Erstens umfasst das Thema Bias und Fairness in der KI hochgradig heterogene Forschungstraditionen – darunter Informatik, Philosophie, Recht und Medizin –, was eine strukturierte Synthese statt eines narrativen Überblicks erfordert. Zweitens minimiert eine SLR Selektionsbias durch die systematische Anwendung vorab festgelegter Ein- und Ausschlusskriterien, wodurch sichergestellt wird, dass die resultierende Evidenzbasis den Stand des Feldes widerspiegelt und nicht das Vorwissen oder die Präferenzen der Autorin oder des Autors (Kitchenham & Charters, 2007).

Drittens ist eine SLR, im Gegensatz zu Meta-Analysen, die quantitative Effektstärken aggregieren, besonders gut für Forschungsfelder geeignet, in denen Methoden, Definitionen und Metriken noch nicht standardisiert sind – ein charakteristisches Merkmal der KI-Fairness-Literatur, in der konkurrierende mathematische Fairness-Definitionen unter bestimmten Bedingungen gegenseitig ausschließend sein können (Kleinberg et al., 2017; Chouldechova, 2017). Der SLR-Rahmen ermöglicht somit eine thematische Synthese qualitativer und quantitativer Befunde über ein breites Spektrum von Studiendesigns hinweg, einschließlich empirischer Evaluationen, taxonomischer Übersichtsarbeiten und theoretischer Rahmenwerke.

Das Forschungsprotokoll wurde vor Beginn der Datenbankrecherche entwickelt und legt im Voraus die Forschungsfragen, die Suchstrategie, die Auswahlkriterien und die Verfahren zur Datenextraktion fest. Dieses Prinzip der Vorab-Registrierung folgt dem von Kitchenham und Charters (2007) empfohlenen Standard zur Erhöhung der Transparenz und Reproduzierbarkeit des Übersichtsprozesses.

2.2 Forschungsfragen

Die folgenden Forschungsfragen (RQ) leiten die systematische Übersichtsarbeit und liefern den analytischen Rahmen für die anschließende Synthese:

RQ1: Welche Arten von Bias wurden in KI- und Machine-Learning-Systemen identifiziert, und wie werden sie in der aktuellen Literatur definiert und kategorisiert?

RQ2: Welche Fairness-Metriken und Evaluationsrahmen wurden vorgeschlagen, und welche jeweiligen Zielkonflikte und Limitationen weisen sie auf?

RQ3: Welche Mitigation-Strategien wurden zur Adressierung von Bias in KI-Systemen entwickelt, und welche Evidenz existiert zu ihrer Wirksamkeit?

RQ4: Wie adressieren aktuelle regulatorische Rahmenwerke – insbesondere der EU AI Act – Bias- und Fairness-Anforderungen, und welche Lücken bestehen weiterhin?

Diese vier Forschungsfragen entsprechen der thematischen Struktur des Hauptkapitels zu den Befunden (Kapitel 3) und stellen sicher, dass Literaturrecherche und Datenextraktion systematisch an den Zielen der Untersuchung ausgerichtet sind.

2.3 Suchstrategie

Die Literaturrecherche wurde über vier Datenbanken durchgeführt, die die relevanten Fachdisziplinen abdecken: PubMed/PMC für biomedizinische und klinische KI-Forschung, Springer Link für Informatik und interdisziplinäre Publikationen, JSTOR für philosophische und ethische Perspektiven, sowie arXiv für aktuelle Preprints im Bereich Künstliche Intelligenz und Machine Learning (Fachkategorien: cs.AI, cs.CY, cs.CL und stat.ML). Die Kombination dieser Datenbanken wurde gewählt, um die disziplinäre Abdeckung zu maximieren und gleichzeitig praktisch handhabbar zu bleiben.

Suchstrings wurden unter Verwendung Boolescher Operatoren (AND, OR) konstruiert und kombinierten Begriffe mit Bezug zu den Kernkonstrukten der Untersuchung. Pilotsuchen wurden durchgeführt, um die Spezifität der Strings zu kalibrieren und Recall und Precision auszubalancieren. Wo verfügbar, wurden kontrollierte Vokabularbegriffe (z. B. MeSH-Begriffe für PubMed) neben Freitextbegriffen verwendet, um die Abdeckung zu verbessern. Tabelle 2.1 zeigt die finalisierten Suchstrings für jede Datenbank.

Tabelle 2.1: Suchstrings nach Datenbank

Datenbank	Suchstring
PubMed / PMC	("artificial intelligence"[MeSH] OR "machine learning"[MeSH] OR "large language model*") AND ("bias"[MeSH] OR "algorithmic bias" OR "fairness") AND ("healthcare"[MeSH] OR "clinical decision" OR "diagnosis")

Springer Link	title-abs-key("bias" AND "fairness" AND ("artificial intelligence" OR "machine learning" OR "large language model")) AND year > 2019
JSTOR	ti:("algorithmic fairness" OR "AI bias" OR "machine learning fairness") AND ("ethics" OR "accountability" OR "philosophy")
arXiv (cs.AI, cs.CY, cs.CL)	ti:"fairness" AND abs:("large language model" OR "LLM" OR "bias") — Filter: 2022–2026

Anmerkung. Suchstrings sollten basierend auf den tatsächlichen Trefferzahlen angepasst werden. Zielbereich: 500–2.000 Treffer pro Datenbank vor dem Screening.

Der zeitliche Suchumfang wurde auf 2020–2026 festgelegt, was die erhebliche Beschleunigung der Forschungsproduktion in diesem Feld nach der breiten Verbreitung großer Sprachmodelle (LLMs) und der Einführung wichtiger Regulierungsrahmen wie dem EU AI Act widerspiegelt. Ausnahmen wurden für eine kleine Anzahl grundlegender Arbeiten vor 2020 gemacht, die als richtungsweisende Beiträge zu den theoretischen Grundlagen des Feldes gelten (z. B. Dwork et al., 2012; Kleinberg et al., 2017; Kusner et al., 2017; Kitchenham & Charters, 2007). Diese Quellen sind explizit als Ausnahmen gelistet und unterliegen nicht dem zeitlichen Einschlusskriterium.

2.4 Ein- und Ausschlusskriterien

Um einen rigorosen und transparenten Auswahlprozess zu gewährleisten, wurde vor Beginn des Screenings ein Satz von Ein- und Ausschlusskriterien definiert. Diese Kriterien wurden konsistent über alle Phasen der Übersichtsarbeit angewendet. Tabelle 2.2 zeigt den vollständigen Kriterienkatalog.

Tabelle 2.2: Ein- und Ausschlusskriterien

#	Einschlusskriterien	Ausschlusskriterien
C1	Peer-reviewte Artikel oder Preprints (arXiv) mit ≥ 10 Zitationen	Nicht peer-reviewte Quellen (Blogs, White Papers ohne Referenzen)
C2	Adressiert direkt Bias, Fairness oder Diskriminierung in KI/ML-Systemen	Keine direkte Relevanz für Bias oder Fairness in der KI

C3	Publikationsjahr: 2020–2026 (Grundlagenwerke vor 2020 als explizite Ausnahmen)	Vor 2020, sofern nicht als Grundlagenwerk klassifiziert
C4	Sprache: Englisch oder Deutsch	Jede andere Sprache als Englisch oder Deutsch
C5	Grundlagen-Ausnahmen: Hellman (2020), Kitchenham (2007), Kleinberg et al. (2017), Chouldechova (2017), Dwork et al. (2012), Kusner et al. (2017)	Duplikate: nur die erste/vollständigste Publikation wird beibehalten
C6	Empirische Studien, systematische Übersichtsarbeiten, taxonomische Surveys, theoretische Rahmenwerke	Rein technische Arbeiten ohne Fairness-bezogenen Inhalt

Anmerkung. Grundlagenwerke vor dem zeitlichen Einschlusskriterium sind einzeln aufgeführt und als explizite Ausnahmen klassifiziert (vgl. Abschnitt 2.3).

Die Kriterien priorisieren peer-reviewte empirische und theoretische Beiträge, die sich direkt mit Bias oder Fairness in KI-Systemen befassen. Graue Literatur – einschließlich Blogbeiträge und White Papers ohne verifizierbares Peer-Review – wurde systematisch ausgeschlossen, um die akademische Qualität der Evidenzbasis zu erhalten. Duplikate wurden identifiziert und vor dem Screening entfernt, wobei die vollständigste oder aktuellste Version einer Publikation beibehalten wurde.

2.5 Auswahlprozess und PRISMA Flow

Der Auswahlprozess folgte dem vierphasigen PRISMA-2020-Rahmenwerk (Page et al., 2021): Identifikation, Screening, Eignungsprüfung und finaler Einschluss. Abbildung 2.1 zeigt das entsprechende PRISMA-Flussdiagramm mit den finalen Trefferzahlen je Phase.

Tabelle 2.1a: Finale Suchergebnisse nach Datenbank

Datenbank	Treffer	Zeitraum	Suchfeld
PubMed / PMC	2.112	2020–2026	MeSH-Begriffe
Springer Link	682	2020–2026	Title-Feld, Advanced Search
arXiv	1.262	2022–2026	cs.AI/CY/CL, Title+Abstract

JSTOR	173	2020–2026	All Fields
-------	-----	-----------	------------

Abbildung 2.1: PRISMA-2020-Flussdiagramm (finale Zahlen)

Phase	Beschreibung
Phase 1: Identifikation	Treffer aus Datenbankrecherche: PubMed n=2.112, Springer n=682, arXiv n=1.262, JSTOR n=173. Gesamt n=4.229. Entfernte Duplikate: n=97.
Phase 2: Screening	Gescreente Datensätze (Titel & Abstract): n=4.132. Ausgeschlossen: n=1.997 (kein Bias/Fairness-Bezug, themenfremde Domäne, rein methodischer Bias oder unzureichende Abstract-Daten).
Phase 3: Eignungsprüfung	Bewertete Datensätze (Volltext-/Qualitätsscreening): n=2.135. Ausgeschlossen: n=2.069 (geringerer Qualitätsscore, thematische Redundanz oder unzureichender Bezug zu den Forschungsfragen).
Phase 4: Einschluss	In die finale Synthese eingeschlossene Studien: n=66 (LLM Bias: 18 · Healthcare Bias: 16 · Mitigation: 14 · Fairness Metrics: 10 · Regulation: 8).

Anmerkung. Alle Zahlen basieren auf dem tatsächlich durchgeführten Screening-Prozess (Stand: Juni 2026).

2.5.1 Identifikation

Datensätze wurden aus den vier Datenbanken anhand der in Tabelle 2.1 spezifizierten Suchstrings abgerufen. Alle Ergebnisse wurden exportiert und in ein Literaturverwaltungstool (Zotero) zur Duplikaterkennung importiert. Datensätze, die durch andere Mittel identifiziert wurden – etwa durch Rückwärts- und Vorwärtszitationssuche hochrelevanter Arbeiten – wurden gemäß PRISMA-2020-Leitlinien separat erfasst.

2.5.2 Screening

Nach der Duplikatentfernung wurden Titel und Abstracts aller verbleibenden Datensätze gegen die Ein- und Ausschlusskriterien gescreent. Diese Phase wurde eigenständig durchgeführt, wobei eine Zufallsstichprobe von 10 % der Datensätze zur Konsistenzprüfung doppelt überprüft wurde. Datensätze wurden in dieser Phase ausgeschlossen, wenn sie eindeutig themenfremd, in einer nicht zugelassenen Sprache verfasst oder außerhalb des spezifizierten Zeitfensters publiziert waren.

2.5.3 Eignungsprüfung

Volltexte aller Datensätze, die die Screening-Phase passierten, wurden abgerufen und auf Eignung geprüft. In dieser Phase wurden die Ein- und Ausschlusskriterien vollständig angewendet. Gründe für den Ausschluss wurden für jeden abgelehnten Volltextartikel erfasst, um eine transparente Berichterstattung zu ermöglichen.

2.5.4 Einschluss

Studien, die alle Einschlusskriterien erfüllten, wurden in die finale Synthese aufgenommen. Die Zielstichprobe für die finale Übersichtsarbeit beträgt etwa 40–65 Studien, ein Bereich, der mit vergleichbaren systematischen Literaturanalysen im Bereich der KI-Fairness übereinstimmt (vgl. Chen et al., 2024; Dogan et al., 2025).

2.6 Datenextraktion und Qualitätsbewertung

Für jede eingeschlossene Studie wurde ein standardisiertes Datenextraktionsformular verwendet, um folgende Informationen zu erfassen: Autor(en), Publikationsjahr, Zeitschrift oder Konferenz, Studientyp (empirisch, Survey, theoretisch usw.), KI-Domäne und Anwendungskontext, behandelte(r) Bias-Typ(en), angewandte Fairness-Metriken oder Rahmenwerke, vorgeschlagene oder evaluierte Mitigation-Strategien sowie zentrale Befunde mit Relevanz für die Forschungsfragen.

Die Qualitätsbewertung erfolgte anhand von Kriterien, die von den Critical-Appraisal-Tools des Joanna Briggs Institute (JBI) abgeleitet und für die spezifischen, in der KI-Fairness-Forschung anzutreffenden Studientypen angepasst wurden. Jede Studie wurde anhand drei Dimensionen bewertet: methodische Stringenz, Klarheit der Bias-Definition und Reproduzierbarkeit der berichteten Ergebnisse. Als qualitativ niedrig bewertete Studien wurden in der Synthese beibehalten, jedoch markiert, und ihre Befunde wurden mit entsprechender Vorsicht interpretiert.

2.7 Synthese Ansatz

Angesichts der Heterogenität von Studiendesigns, Methodiken und Ergebnismaßen in der eingeschlossenen Literatur wurde eine narrative thematische Synthese anstelle einer statistischen Meta-Analyse angewendet. Dieser Ansatz steht im Einklang mit etablierter Praxis für systematische Übersichtsarbeiten, die qualitative und konzeptionelle Befunde über unterschiedliche Forschungstraditionen hinweg aggregieren (Kitchenham & Charters, 2007).

Studien wurden thematisch entsprechend den vier Forschungsfragen gruppiert. Innerhalb jedes thematischen Clusters wurden Befunde verglichen und gegenübergestellt, um Muster, Übereinstimmungen und Widersprüche zu identifizieren. Wo quantitative Daten berichtet wurden, wurden diese deskriptiv zusammengefasst. Die Synthese ist so strukturiert, dass die Leserin oder der Leser die Evidenzbasis für jeden zentralen Befund bis zu den zugrunde liegenden Quellstudien zurückverfolgen kann.

2.8 Limitationen der Methodik

Mehrere Limitationen des methodischen Ansatzes sollten anerkannt werden. Erstens kann die Beschränkung auf vier Datenbanken – obwohl disziplinär breit gefasst – zum Ausschluss relevanter Studien führen, die in spezialisierten, dort nicht indexierten Publikationsorganen erschienen sind. Diese Limitation wird teilweise durch Zitationssuche bei stark zitierten Arbeiten abgemildert. Zweitens führt die Beschränkung auf englisch- und deutschsprachige Publikationen zu einem Sprach-Bias, der Forschung aus nicht-anglophonen Kontexten unterrepräsentieren kann, insbesondere in Regionen, in denen KI-Einsatz und Fairness-Bedenken deutlich von westlichen Kontexten abweichen können.

Drittens bedeutet die rasche Entwicklung dieses Feldes, dass zwischen der Durchführung der Datenbankrecherchen und der Finalisierung der Arbeit neue relevante Publikationen erschienen sein können. Viertens kann das Risiko von Selektionsbias in den Screening- und Extraktionsphasen, da es sich um eine Einzelautor-Studie handelt, nicht vollständig ausgeschlossen werden; dies wird teilweise durch die transparente Dokumentation von Ein- und Ausschlussentscheidungen sowie die Verwendung eines vorab spezifizierten Kriterienkatalogs adressiert. Diese Limitationen werden in Kapitel 5 weiter diskutiert.

3. Ergebnisse

Dieses Kapitel präsentiert die thematische Synthese der Literatur, die durch den in Kapitel 2 beschriebenen systematischen Übersichtsprozess identifiziert wurde. Die Befunde sind entsprechend den vier Forschungsfragen (RQ1–RQ4) strukturiert und in vier thematische Abschnitte gegliedert: Bias in großen Sprachmodellen (Abschnitt 3.1), Bias in der Gesundheits-KI (Abschnitt 3.2), Fairness-Metriken und theoretische Rahmenwerke (Abschnitt 3.3) sowie Mitigation-Strategien (Abschnitt 3.4). Jeder Abschnitt synthetisiert die Evidenz aus den eingeschlossenen Studien, hebt Bereiche der Übereinstimmung und des Widerspruchs hervor und identifiziert offene Forschungsfragen.

3.1 Bias in Large Language Models

Large Language Models (LLMs) – transformerbasierte neuronale Netzwerke, die auf umfangreichen Korpora menschlich generierten Texts trainiert werden – haben sich zu einer der am breitesten eingesetzten Formen Künstlicher Intelligenz entwickelt, mit Anwendungsbereichen, die vom Kundenservice über die juristische Recherche bis hin zur medizinischen Entscheidungsunterstützung reichen (Gallegos et al., 2024). Ihre Fähigkeit, flüssigen, kontextuell angemessenen Text zu generieren, hat ihre Verbreitung beschleunigt, doch derselbe Trainingsprozess, der LLMs sprachliche Kompetenz verleiht, setzt sie auch dem gesamten Spektrum menschlicher sozialer Biases aus, die in geschriebener Sprache eingebettet sind. Dieser Abschnitt untersucht, wie sich Bias in LLMs manifestiert, wie er gemessen wird und was die aktuelle Evidenzbasis über seinen Umfang und seine Schwere offenbart.

3.1.1 Konzeptionelle Grundlagen: Definition von Bias in LLMs

Der Begriff „Bias“ im Kontext von LLMs bezeichnet kein einzelnes, klar abgegrenztes Phänomen, sondern eine Familie verwandter Probleme, deren Grenzen in der Literatur umstritten sind. Eine grundlegende Unterscheidung, die Mehrabi et al. (2021) in ihrer breit zitierten Übersichtsarbeit treffen – der umfassendsten Taxonomie ihrer Art mit über 2.600 Zitationen –, differenziert zwischen Daten-Bias, algorithmischem Bias und Deployment-Bias. Daten-Bias bezieht sich auf systematische Verzerrungen, die aus der Zusammensetzung der Trainingskorpora entstehen; algorithmischer Bias entsteht aus Modellarchitekturen oder

Optimierungszielen, die bestimmte Muster verstärken; und Deployment-Bias entsteht, wenn ein Modell in Kontexten angewendet wird, die von seinen Designannahmen abweichen.

Gallegos et al. (2024) unterscheiden in der umfassendsten bisher verfügbaren, speziell auf LLMs ausgerichteten Bias-Übersichtsarbeit, veröffentlicht in Computational Linguistics (MIT Press), weiterführend zwischen repräsentativen Schäden (representational harms) und allokativen Schäden (allocational harms). Repräsentative Schäden umfassen die systematische Herabwürdigung, Stereotypisierung oder Auslöschung sozialer Gruppen in Modellausgaben – etwa die durchgängige Assoziation einer demografischen Gruppe mit negativen Attributen. Allokative Schäden entstehen, wenn verzerrte Modellausgaben zu einer ungleichen Verteilung von Ressourcen oder Chancen führen, etwa beim automatisierten Recruiting, der Kreditwürdigkeitsprüfung oder der medizinischen Triage. Diese beiden Kategorien schließen sich nicht gegenseitig aus und treten in realen Einsatzszenarien häufig gemeinsam auf.

Eine weitere, für den LLM-Kontext spezifische Unterscheidung ist Fairness in LLMs und Fairness im klassischen überwachten maschinellen Lernen, da LLMs offene Texte generieren, anstatt kategoriale Entscheidungen zu produzieren (Chu et al., 2024). Dies macht die Operationalisierung von Fairness erheblich komplexer, da keine einzelne Entscheidungsgrenze zur Anpassung existiert und kein offensichtlicher Ground Truth vorliegt, gegen die Ausgaben evaluiert werden könnten. Die Herausforderung wird durch den Umfang der LLMs verschärft: Mit Hunderten von Milliarden Parametern, trainiert auf web-skalierten Korpora, sind die Orte, an denen Bias entsteht, verteilt und schwer zu lokalisieren.

Tabelle 3.1: Taxonomie der für LLMs relevanten Bias-Typen

Bias-Typ	Definition	Zentrale Quellen
Historical Bias	Bias, der aus historischen Daten entsteht, die vergangene gesellschaftliche Diskriminierung widerspiegeln (z. B. Lohnunterschiede zwischen Geschlechtern, kodiert in Gehaltsdatensätzen).	Mehrabi et al. (2021); Ferrara (2024)
Representation Bias	Tritt auf, wenn bestimmte demografische Gruppen in Trainingsdaten über- oder unterrepräsentiert sind, wodurch das Modell ungleich über Gruppen hinweg funktioniert.	Gallegos et al. (2024); Mehrabi et al. (2021)

Measurement Bias	Entsteht, wenn Proxys zur Messung von Konstrukten verwendet werden (z. B. Verhaftungsdaten als Proxy für kriminelles Verhalten), was Ergebnisse für bestimmte Gruppen systematisch verzerrt.	Mehrabi et al. (2021); Chen et al. (2024)
Aggregation Bias	Resultiert aus der Anwendung eines einzelnen Modells auf eine heterogene Population, wobei Subgruppenunterschiede ignoriert werden, die eine separate Behandlung erfordern würden.	Mehrabi et al. (2021)
Evaluation Bias	Entsteht, wenn Benchmark-Datensätze nicht die volle Diversität der Zielpopulation widerspiegeln, was zu übermäßig optimistischen Leistungseinschätzungen führt.	Gallegos et al. (2024); Jung et al. (2025)
Deployment Bias	Tritt auf, wenn ein Modell in einem Kontext eingesetzt wird, der von dem abweicht, für den es entworfen wurde, was zu unbeabsichtigten und potenziell diskriminierenden Ergebnissen führt.	Ferrara (2024)

Anmerkung. Taxonomie synthetisiert aus Mehrabi et al. (2021), Gallegos et al. (2024) und Ferrara (2024). Die Kategorien schließen sich nicht gegenseitig aus; ein einzelnes Modell kann mehrere Bias-Typen gleichzeitig aufweisen.

3.1.2 Empirische Evidenz: Wie sich Bias in LLM-Ausgaben manifestiert

Eine umfangreiche und wachsende empirische Forschungsbasis dokumentiert das Vorhandensein sozialen Bias in LLM-Ausgaben über mehrere demografische Dimensionen hinweg. Die Literatur identifiziert Geschlecht, Rasse/Ethnizität, Religion und sozioökonomischen Status als die am häufigsten untersuchten Dimensionen, mit wachsender Anerkennung dafür, dass intersektionale Effekte – Bias an der Schnittstelle mehrerer Attribute – im Verhältnis zu ihrer realen Bedeutung unterbeforscht sind (An et al., 2025).

Gender Bias

Gender Bias ist die am umfassendsten dokumentierte Form von LLM-Bias in der untersuchten Literatur. LLMs assoziieren Führungsrollen, technische Expertise und hohen beruflichen Status durchgängig mit männlichen Identifikatoren, während weibliche Identifikatoren mit Fürsorge-, Haushalts- und Unterstützungsrollen assoziiert werden (Gallegos et al., 2024). Diese berufliche Stereotypisierung bleibt selbst in instruktionsoptimierten Modellen bestehen,

die ein Reinforcement Learning from Human Feedback (RLHF) durchlaufen haben, was darauf hindeutet, dass Alignment-Verfahren während des Pre-Trainings eingebettete Biases nicht vollständig neutralisieren.

An et al. (2025) untersuchen in einer in PNAS Nexus veröffentlichten Studie Gender- und Rassen-Bias in LLMs anhand eines groß angelegten Lebenslauf-Evaluationsexperiments – einer realitätsnahen Anwendung mit hoher gesellschaftlicher Relevanz. Ihre Befunde zeigen, dass automatisiertes, LLM-basiertes Resume-Screening signifikante Geschlechterunterschiede aufweist: Lebensläufe mit weiblichen Namen werden als weniger erfahren bewertet und erhalten weniger angesehene berufliche Werdegänge zugeschrieben als sonst identische Lebensläufe mit männlichen Namen. Entscheidend ist, dass die Autorinnen und Autoren zeigen, dass intersektionale Effekte nicht additiv sind: Der Bias, dem People of Color zugehörige Frauen ausgesetzt sind, ist größer als die Summe der unabhängig gemessenen Geschlechts- und Rassen-Bias-Effekte, was darauf hindeutet, dass eindimensionale Analysen reale Ungleichheiten erheblich unterschätzen.

Gender Bias in LLMs wird von Rickman (2025) im Kontext der Langzeitpflege untersucht und stellt systematische Unterschiede fest, wie LLMs männliche und weibliche Patienten charakterisieren, mit Implikationen für Ressourcenallokation und klinische Entscheidungsunterstützung. Diese Befunde verbinden die Gender-Bias-Literatur direkt mit Anwendungen im Gesundheitswesen und werden in Abschnitt 3.2 weiter diskutiert.

Rassen- und ethnischer Bias

Rassen- und ethnischer Bias in LLMs ist gut dokumentiert, wenngleich die Evidenzbasis weniger umfangreich ist als für Gender Bias, was teilweise an der größeren Schwierigkeit liegt, kontrollierte Experimente zu entwerfen, die rassenbezogene Marker variieren, ohne Störfaktoren einzuführen. Gallegos et al. (2024) synthetisieren Evidenz dafür, dass LLMs rassenbezogene Stereotype über mehrere Aufgaben hinweg reproduzieren, einschließlich Textvervollständigung, Sentiment-Analyse und Frage-Antwort-Aufgaben. Modelle assoziieren durchgängig Schwarze und hispanische Identifikatoren mit höheren Kriminalitätsraten, geringerer Bildungserreichung und niedrigerem sozioökonomischem Status – Muster, die historische Biases in Trainingskorpora direkt widerspiegeln.

Rassen-Bias bei der automatisierten Lebenslaufbewertung zeigt, dass Kandidatinnen und Kandidaten mit Namen, die stereotyp mit nicht-weißen Bevölkerungsgruppen assoziiert werden, werden als weniger qualifiziert, weniger kommunikationsstark im Englischen und eher

mit nicht-US-amerikanischem Bildungs- und Berufshintergrund bewertet – selbst wenn der Lebenslaufinhalt sonst identisch ist (An et al., 2025). Die Studie zeigt, dass diese Biases in marktführenden kommerziellen LLMs vorhanden und robust gegenüber Prompt-Variation sind, was darauf hindeutet, dass sie keine Artefakte spezifischer Prompt-Design-Entscheidungen sind.

Religiöser Bias

Mehrabi et al. (2021) dokumentieren anhaltenden anti-muslimischen Bias in früheren Sprachmodellen, wobei Modelle den Islam und muslimische Identifikatoren überproportional mit Gewalt, Terrorismus und Extremismus in Textvervollständigungsaufgaben assoziieren. Laut Gallegos et al. (2024) ist dieser Bias in neueren LLMs fortbestehend, einschließlich Modellen, die einem Safety-Fine-Tuning unterzogen wurden. Die Robustheit dieses Bias über Modellgenerationen hinweg deutet darauf hin, dass er tief in den distributionalen Mustern großer, aus dem Web extrahierter Korpora eingebettet ist, die bestimmte kulturelle und geopolitische Perspektiven überrepräsentieren.

Intersektionaler und sozioökonomischer Bias

Ein besonders wichtiger Befund über die untersuchte Literatur hinweg ist, dass Bias-Analysen entlang einzelner demografischer Dimensionen die Schwere und Komplexität des Bias, dem Individuen an der Schnittstelle mehrerer marginalisierter Positionen gleichzeitig ausgesetzt sind, systematisch unterschätzen (An et al., 2025; Gallegos et al., 2024). Chu et al. (2024) argumentieren, dass der vorherrschende Fokus des Feldes auf gruppenbezogene statistische Parität die nuancierten Auswirkungen auf Individuen innerhalb von Subgruppen nicht erfasst, und fordern Fairness-Rahmenwerke, die Intersektionalität berücksichtigen können. Ferrara (2024) merkt zudem an, dass sozioökonomischer Bias – die Assoziation bestimmter sprachlicher Register oder Dialekte mit geringerer Kompetenz – zu den am wenigsten untersuchten Dimensionen gehört, obwohl er in LLM-Ausgaben weit verbreitet ist.

Tabelle 3.2: Bias-Dimensionen in LLMs — Evidenzübersicht

Bias-Dimension	Manifestation in LLMs	Zentrale Quellen
Gender Bias	LLMs assoziieren Führungspositionen und hochrangige Berufe durchgängig mit männlichen Identifikatoren und Fürsorgerollen mit weiblichen Identifikatoren.	Gallegos et al. (2024); An et al. (2025); Rickman (2025)
Rassen-/ethnischer Bias	Modelle produzieren stereotype Assoziationen mit Rassen- und ethnischen Gruppen, die Ausgaben bei Recruiting-Simulationen, medizinischen Empfehlungen und Textgenerierung beeinflussen.	An et al. (2025); Gallegos et al. (2024); Chen et al. (2024)
Religiöser Bias	Studien dokumentieren anhaltenden anti-muslimischen Bias, wobei Modelle den Islam überproportional mit Gewalt in Generierungsaufgaben assoziieren.	Mehrabi et al. (2021); Gallegos et al. (2024)
Sozioökonomischer Bias	LLMs reproduzieren klassenbezogene Stereotype und assoziieren bestimmte Dialekte oder Sprachmuster mit geringerer Kompetenz oder Glaubwürdigkeit.	Ferrara (2024); Chu et al. (2024)
Intersektionaler Bias	Bias-Effekte werden an der Schnittstelle mehrerer demografischer Attribute verstärkt (z. B. Schwarze Frauen) – nicht erfassbar durch isolierte Betrachtung jeder Dimension.	An et al. (2025); Gallegos et al. (2024)

Anmerkung. Evidenz synthetisiert aus empirischen Studien der untersuchten Literatur. Alle Dimensionen sind über mehrere Studien hinweg dokumentiert; die Evidenzlage ist für Gender- und Rassen-Bias am stärksten.

3.1.3 Quellen von Bias in der LLM-Pipeline

Das Verständnis, an welcher Stelle Bias in die LLM-Pipeline eintritt, ist eine Voraussetzung für wirksame Mitigation. Die untersuchte Literatur identifiziert drei primäre Eintrittspunkte: Trainingsdaten, das Pre-Training-Ziel sowie den Fine-Tuning- und Alignment-Prozess.

Trainingsdaten stellen die am ausführlichsten untersuchte Bias-Quelle dar. LLMs werden typischerweise auf großen, aus dem Web extrahierten Korpora trainiert, die die in menschlich generiertem Text im Internet vorhandenen Biases widerspiegeln und verstärken – einschließlich historischer Diskriminierung, kultureller Stereotype und der Überrepräsentation bestimmter demografischer Gruppen und Perspektiven (Mehrabi et al., 2021; Ferrara, 2024). Da diese Korpora selten auf demografische Repräsentativität hin kuratiert werden, kodieren

die resultierenden Modelle eine verzerrte Sicht der sozialen Welt. Gallegos et al. (2024) unterscheiden weiterführend zwischen intrinsischem Bias – vorhanden in den Repräsentationen des vortrainierten Modells – und extrinsischem Bias, der beim Fine-Tuning für spezifische nachgelagerte Aufgaben entsteht. Beide Formen sind empirisch dokumentiert, und ihre Wechselwirkung ist nicht gut verstanden.

Das Pre-Training-Ziel – typischerweise die Vorhersage des nächsten Tokens – unterscheidet nicht zwischen faktischer Information und stereotypen Assoziationen. Das Modell lernt statistische Kookkurrenzmuster ohne jeden normativen Filter, mit dem Ergebnis, dass häufige, aber gesellschaftlich schädliche Assoziationen neben akkuratem Weltwissen kodiert werden (Chu et al., 2024). Diese architektonische Eigenschaft erschwert es, spezifische Biases nachträglich zu entfernen, ohne die breiteren Fähigkeiten des Modells zu beeinträchtigen.

Fine-Tuning- und Alignment-Verfahren, einschließlich RLHF, führen einen dritten Punkt ein, an dem Bias eintreten oder verändert werden kann. Während Alignment den Ausdruck bestimmter schädlicher Ausgaben reduzieren kann, dokumentiert die untersuchte Literatur, dass es Bias nicht eliminiert und in manchen Fällen neue Biases einführen kann, indem es die Antworten des Modells an die Präferenzen der im RLHF-Prozess eingesetzten menschlichen Annotatorinnen und Annotatoren anpasst (Gallegos et al., 2024). Dieser Annotations-Bias – der Einfluss der eigenen Vorurteile der Annotierenden auf gelabelte Trainingsdaten – wird von Mehrabi et al. (2021) als eine eigenständige und unterschätzte Quelle systematischer Verzerrung identifiziert.

3.1.4 Messung von Bias in LLMs: Benchmarks und Evaluationsrahmen

Die Messung von Bias in LLMs ist eine methodisch komplexe Herausforderung, die eine reichhaltige, aber fragmentierte Literatur hervorgebracht hat. Gallegos et al. (2024) organisieren die Evaluationslandschaft in zwei Taxonomien: eine für Evaluationsmetriken, gegliedert nach der Ebene, auf der sie operieren (Embeddings, Output-Wahrscheinlichkeiten und generierter Text), und eine für Evaluationsdatensätze. Dieser Organisationsrahmen offenbart eine systematische Diskrepanz zwischen verfügbaren Metriken und der komplexen, offenen Natur von LLM-Ausgaben: Die meisten existierenden Metriken wurden für Klassifikationsaufgaben entwickelt und erfordern Anpassung oder Erweiterung, um sinnvoll auf generative Modelle angewendet zu werden.

Die untersuchte Literatur identifiziert eine Reihe von Benchmarks, die bedeutende methodische Beiträge zum Feld darstellen. Tabelle 3.3 fasst die relevantesten Benchmarks zusammen, einschließlich jener, die in den direkt in diese Untersuchung eingeschlossenen Quellen vorgestellt werden.

Tabelle 3.3: Zentrale Benchmarks zur Bias-Evaluation in LLMs

Benchmark	Autoren	Fokus	Zentraler Befund/Beitrag
CEB	Wang et al. (2024)	Kompositionelle Evaluation von Fairness über mehrere demografische Dimensionen und Aufgabentypen in LLMs.	Deckt 11 soziale Gruppen ab; kompositionelles Design ermöglicht feingranulare Bias-Erkennung
FLEX	Jung et al. (2025)	Evaluiert die Robustheit von Fairness — d. h., ob Fairness unter adversariellen Prompt-Variationen bestehen bleibt.	Zeigt, dass unter Standardbedingungen fair wirkende Modelle bei leichten Störungen versagen
WinoBias / WinoGender	Zhao et al. (2018)	Koreferenzauflösungsaufgaben zur Untersuchung von Gender Bias bei Pronomen-Beruf-Zuordnungen.	Klassischer Benchmark; weit verbreitet, aber auf binäres Geschlecht und Englisch beschränkt
BBQ	Parrish et al. (2022)	Bias-Benchmark in Frage-Antwort-Aufgaben über neun soziale Dimensionen.	Kontextsensitiv; deckt Ambiguitätsausnutzung in verzerrten Modellausgaben auf

Anmerkung. WinoBias/WinoGender (Zhao et al., 2018) und BBQ (Parrish et al., 2022) sind als Referenz-Benchmarks aufgenommen; sie zählen nicht zu den Primärquellen dieser Untersuchung, werden aber in der eingeschlossenen Literatur häufig zitiert.

CEB (Compositional Evaluation Benchmark) adressiert eine zentrale Limitation früherer Benchmarks: die Tendenz, Bias entlang einer einzelnen demografischen Dimension isoliert zu evaluieren (Wang et al., 2024). CEB verwendet ein kompositionelles Design, das mehrere demografische Attribute systematisch gleichzeitig variiert und damit die Erkennung intersektionaler Bias-Muster ermöglicht, die eindimensionale Benchmarks übersehen würden. Der Benchmark deckt elf soziale Gruppen über vier Aufgabentypen ab und gehört damit zu den umfassendsten derzeit verfügbaren Evaluationsinstrumenten.

Jung et al. (2025) liefern mit dem FLEX-Benchmark einen konzeptionell eigenständigen und bedeutenden Fortschritt: Anstatt das Ausmaß von Bias unter Standardbedingungen zu messen, evaluiert FLEX die Robustheit von Fairness – also, ob das faire Verhalten eines Modells unter leichten adversariellen Störungen des Eingabe-Prompts bestehen bleibt. Ihre Befunde sind auffällig: Modelle, die unter Standard-Prompting-Bedingungen Fairness-Kriterien zu erfüllen scheinen, versagen häufig, wenn Prompts leicht umformuliert, anders formatiert oder in einem anderen sprachlichen Register präsentiert werden. Diese Fragilität scheinbarer Fairness hat bedeutende Implikationen für den realen Einsatz, bei dem Prompt-Variation die Norm und nicht die Ausnahme ist.

3.1.5 Zusammenfassung und Forschungslücken

Die untersuchte Literatur liefert robuste und konvergierende Evidenz dafür, dass Bias in LLMs über mehrere demografische Dimensionen, Modellarchitekturen und Anwendungskontexte hinweg verbreitet ist. Folgende Schlussfolgerungen lassen sich mit hoher Konfidenz ziehen: (1) Gender- und Rassen-Bias sind die am umfassendsten dokumentierten Formen von LLM-Bias, mit konsistenter empirischer Evidenz über zahlreiche unabhängige Studien hinweg; (2) Alignment-Verfahren wie RLHF reduzieren, eliminieren aber nicht Bias; (3) intersektionaler Bias – an der Schnittstelle mehrerer demografischer Attribute – wird durch eindimensionale Analysen systematisch unterschätzt; und (4) die Robustheit von Fairness unter Prompt-Variation ist stark eingeschränkt, wie von Jung et al. (2025) gezeigt.

Mehrere bedeutende Forschungslücken sind erkennbar. Erstens konzentriert sich die Literatur stark auf englischsprachige Modelle und westliche soziale Kontexte; sprachübergreifender und kulturübergreifender Bias bleibt schlecht verstanden. Zweitens ist die Beziehung zwischen intrinsischem Bias (in Modellrepräsentationen) und extrinsischem Bias (in der nachgelagerten Aufgabenleistung) empirisch unzureichend charakterisiert, was gezielte Mitigation erschwert. Drittens fehlen Längsschnittstudien, die verfolgen, wie sich Bias über aufeinanderfolgende Modellgenerationen hinweg entwickelt, praktisch vollständig, was es erschwert zu bewerten, ob das Feld bedeutsame Fortschritte macht. Viertens bleiben Evaluationsrahmen für offene generative Aufgaben deutlich weniger entwickelt als jene für Klassifikation, was eine Messlücke schafft, die zu einer systematischen Unterschätzung realer Bias-Ausprägungen führen kann.

Diese Lücken fließen in die Analyse der nachfolgenden Abschnitte ein, insbesondere in die Diskussion der Mitigation-Strategien (Abschnitt 3.4) und die Synthese in Kapitel 5.

3.2 Bias in Gesundheits-KI

Das Gesundheitswesen gehört zu den Bereichen mit den höchsten Einsätzen für KI-Anwendungen: Verzerrte Modellausgaben können sich unmittelbar in Fehldiagnosen, ungeeigneten Behandlungsempfehlungen oder einer unfairen Verteilung klinischer Ressourcen niederschlagen. Im Gegensatz zu vielen verbraucherorientierten KI-Anwendungen tragen Fehler in der Gesundheits-KI lebensverändernde Konsequenzen, was diesen Bereich zu einem Schwerpunkt empirischer Bias-Forschung gemacht hat. Dieser Abschnitt synthetisiert die Evidenz zu Bias in der Gesundheits-KI entlang vier Dimensionen: die Natur und Quellen von Bias im klinischen maschinellen Lernen, empirische Evidenz für ungleiche Ergebnisse, methodische Ansätze zur Bias-Erkennung sowie die besonderen Herausforderungen, die durch aufkommende Anwendungen großer Sprachmodelle im klinischen Kontext entstehen.

3.2.1 Umfang und Tragweite von Bias in der klinischen KI

Die Integration von KI in klinische Arbeitsabläufe hat sich rasch beschleunigt und umfasst diagnostische Bildgebung, Risikovorhersage, klinische Dokumentation und Bevölkerungsgesundheitsmanagement. In einem Scoping-Review zum Management nicht-übertragbarer Krankheiten auf Populationsebene dokumentieren Birdi et al. (2024), dass maschinelles Lernen zunehmend nicht nur in individuelle klinische Entscheidungen, sondern auch in die Überwachung der öffentlichen Gesundheit und die Ressourcenallokation eingebettet ist – was bedeutet, dass verzerrte Modelle ganze Bevölkerungsgruppen statt einzelner Patientinnen und Patienten betreffen können. Ihre Übersichtsarbeit hebt ein strukturelles Problem hervor: Trainingsdaten für Gesundheitsmodelle auf Populationsebene unterrepräsentieren häufig ressourcenarme Umgebungen und historisch unterversorgte Bevölkerungsgruppen, wodurch eine systematische Diskrepanz zwischen den Orten entsteht, an denen KI-Tools validiert werden, und jenen, an denen sie am dringendsten benötigt werden.

Jabbour et al. (2023) liefern in einer in JAMA veröffentlichten randomisierten klinischen Vignettenstudie direkte experimentelle Evidenz dafür, dass die Tragweite von KI-Bias im Gesundheitswesen kein rein theoretisches Problem ist. Ihre Studie zeigt, dass systematisch verzerrte KI-Diagnosemodelle die diagnostische Genauigkeit von Klinikerinnen und Klinikern aktiv verschlechtern können, anstatt sie zu verbessern – Kliniker die verzerrten

Modellausgaben ausgesetzt waren, übernahmen den Bias in ihre eigene Beurteilung, ein Befund mit erheblichen Implikationen dafür, wie Gesundheits-KI in der Praxis eingesetzt und überwacht wird. Dieses Ergebnis unterstreicht, dass das Risiko von Bias in der Gesundheits-KI sich nicht auf den Algorithmus selbst beschränkt, sondern sich auf die Mensch-KI-Interaktion in der klinischen Entscheidungsfindung ausdehnt.

3.2.2 Soziodemografischer und rassenbezogener Bias in klinischen Modellen

Colacci et al. (2025) bieten die umfassendste Synthese in der untersuchten Literatur dazu, wie sich soziodemografischer Bias mechanistisch im klinischen maschinellen Lernen manifestiert. Ihre Scoping-Review katalogisiert wiederkehrende Mechanismen über mehrere dokumentierte Fälle algorithmischen Bias hinweg: Ungleichgewicht der Trainingsdaten (bestimmte demografische Gruppen sind in den zum Training klinischer Modelle verwendeten Daten systematisch unterrepräsentiert), Label-Bias (die Ground-Truth-Labels selbst spiegeln historische Muster ungleichen Zugangs zu Versorgung oder Diagnose wider) und Proxy-Variablen-Bias (scheinbar neutrale Variablen, etwa Gesundheitskosten oder Nutzungsverhalten, fungieren als Proxys für Rasse oder sozioökonomischen Status und führen Diskriminierung damit indirekt wieder ein). Dieser mechanistische Rahmen ist wertvoll, weil er über die reine Dokumentation der Existenz von Bias hinausgeht und erklärt, wie er technisch entsteht – eine Voraussetzung für die Entwicklung wirksamer Mitigation-Strategien.

Mit einem „Subgroup-Learnability“-Rahmenwerk tragen Gulamali et al. (2025) im Journal of Medical Internet Research einen methodisch innovativen Ansatz zu rassenbezogenem Bias in Gesundheitsdatensätzen bei. Ihre Studie zeigt, dass Deep-Learning-Verfahren nicht nur zur Erkennung impliziter und expliziter Rassen-Bias-Signaturen in Gesundheitsdatensätzen genutzt werden können, sondern auch zur teilweisen Korrektur dieser während der Modellentwicklung. Diese duale Erkennungs- und Korrekturfähigkeit stellt einen methodischen Fortschritt gegenüber rein diagnostischen Bias-Erkennungsansätzen dar, wobei die Autorinnen und Autoren anmerken, dass Subgroup-Learnability-Verfahren Bias noch nicht vollständig eliminieren können, insbesondere in Fällen, in denen der Bias tief im zugrunde liegenden klinischen datengenerierenden Prozess verwurzelt ist, statt erst auf der Modellierungsebene eingeführt zu werden.

Schuch et al. (2023), veröffentlicht in JAMA Network Open, liefern eine konkrete empirische Fallstudie zur Fairness bei einer spezifischen, realweltlichen klinischen Vorhersageaufgabe:

der Prognose, welche Erwachsenen Gefahr laufen, präventive Zahnpflege zu versäumen. Ihre Fairness-Evaluation zeigt unterschiedliche Vorhersagegenauigkeit zwischen sozioökonomischen Subgruppen, wobei das Modell systematisch schlechter für Bevölkerungsgruppen funktioniert, die bereits einem höheren Risiko eingeschränkten Zugangs zur Gesundheitsversorgung ausgesetzt sind. Dieser Befund illustriert ein beunruhigendes Muster, das über mehrere Studien dieser Untersuchung hinweg dokumentiert ist: KI-Modelle laufen Gefahr, gerade für jene Bevölkerungsgruppen am ungenauesten zu sein, die am meisten von einer akkuraten, gerechten Vorhersage profitieren würden – ein Phänomen, das in der breiteren algorithmischen-Fairness-Literatur manchmal als „Fairness-Accuracy-Paradox“ bezeichnet wird (vgl. Abschnitt 3.3).

Tabelle 3.4: Empirische Studien zu Bias in der klinischen KI

Studie	Fokus	Ansatz	Kernbefund
Colacci et al. (2025)	Soziodemografischer Bias	Scoping Review zu algorithmischen Bias-Fällen und zugrunde liegenden Mechanismen im klinischen ML	Identifiziert wiederkehrende Mechanismen: Ungleichgewicht der Trainingsdaten, Label-Bias, Proxy-Variablen
Gulamali et al. (2025)	Erkennung von Rassen-Bias	Subgroup-Learnability-Rahmenwerk zur Erkennung und Mitigation impliziten/expliciten Rassen-Bias in Gesundheitsdatensätzen	Zeigt, dass Deep Learning Rassen-Bias-Signaturen sowohl erkennen als auch teilweise korrigieren kann
Schuch et al. (2023)	Zugang/Präventivversorgung	Fairness-Evaluation von ML-Modellen zur Vorhersage versäumter präventiver Zahnpflege	Dokumentiert unterschiedliche Vorhersagegenauigkeit zwischen sozioökonomischen Subgruppen
Jabbour et al. (2023)	Diagnostischer Bias	Randomisierte klinische Vignettenstudie zur KI-unterstützten Diagnose hospitalisierter Patient:innen	Systematisch verzerrte KI-Modelle können die diagnostische Genauigkeit von Klinikern verschlechtern statt verbessern

Birdi et al. (2024)	Bias auf Populationsebene	Scoping Review zu ML-Bias im Management nicht-übertragbarer Krankheiten auf Populationsebene	Hebt Unterrepräsentation ressourcenarmer Umgebungen in Trainingsdaten hervor
Chen et al. (2024)	EHR-basierter Bias	Systematische Übersichtsarbeit zu Bias-Erkennung und -Mitigation in Modellen auf Basis elektronischer Patientenakten	Synthetisiert Erkennungstechniken und Mitigation-Strategien entlang der EHR-Pipeline

Anmerkung. Die Tabelle fasst zentrale empirische Befunde aus den im Rahmen der SLR identifizierten Studien zusammen; vollständige Zitationen siehe Literaturverzeichnis.

3.2.3 Bias entlang der Pipeline elektronischer Patientenakten

Chen et al. (2024) liefern in einer im Journal of the American Medical Informatics Association veröffentlichten systematischen Übersichtsarbeit die umfassendste Synthese von Bias-Erkennungs- und Mitigation-Strategien speziell für Modelle auf Basis elektronischer Patientenakten (Electronic Health Records, EHR). Ihre Übersichtsarbeit verfolgt Bias entlang der gesamten EHR-zu-Modell-Pipeline: von der Datenerhebung (bei der Dokumentationspraktiken systematisch zwischen Patientengruppen und klinischen Settings variieren), über das Feature Engineering (bei dem scheinbar objektive klinische Variablen historische Versorgungsunterschiede kodieren können), bis zum Modell-Deployment (bei dem Leistungsüberwachung häufig fehlt oder unzureichend nach demografischer Subgruppe stratifiziert ist). Ein zentraler Beitrag dieser Übersichtsarbeit ist die explizite Verknüpfung zwischen technischen Bias-Erkennungsmethoden (z. B. Disaggregation der Subgruppenleistung, Auditing von Fairness-Metriken) und konkreten Mitigation-Strategien (z. B. Reweighting, Adversarial Debiasing, Post-hoc-Kalibrierung), die in Abschnitt 3.4 weiter diskutiert werden.

Huang et al. (2024) konzentrieren sich in einer ergänzenden, in derselben Zeitschrift veröffentlichten Scoping-Review speziell auf faire Machine-Learning-Verfahren bei deren Anwendung auf reale klinische Daten, im Gegensatz zu kuratierten Forschungsdatensätzen. Ihre Übersichtsarbeit identifiziert eine bedeutende Lücke zwischen Fairness-Verfahren, die auf Benchmark-Datensätzen entwickelt und validiert wurden, und deren praktischer Anwendbarkeit auf die unordentlichere, heterogenere Natur realer EHR-Daten – einschließlich fehlender Datenmuster, die selbst systematisch nach demografischer Gruppe variieren, ein

Problem, das von den meisten existierenden Fairness-bewussten Machine-Learning-Methoden nicht adäquat adressiert wird.

Unüberwachtes Lernen: Eine unterbeforschte Grenze

Eine wichtige und unbeforschte Lücke in der Gesundheits-Fairness-Literatur identifizieren Adouani et al. (2026). Fast die gesamte empirische und methodische Arbeit zu Fairness im klinischen maschinellen Lernen adressiert überwachte Lernsettings, in denen gelabelte Ergebnisse die direkte Messung prädiktiver Ungleichheiten zwischen Gruppen ermöglichen. Ihre methodische Scoping-Review zeigt, dass Fairness im unüberwachten Lernen – im Gesundheitswesen verwendet für Aufgaben wie Patienten-Phänotypisierung, Clustering zur Kohortenidentifikation und Anomalieerkennung – vergleichsweise unterbeforscht bleibt, obwohl unüberwachte Methoden zunehmend genau in den Bevölkerungsgesundheits- und Public-Health-Kontexten eingesetzt werden, die Birdi et al. (2024) als besonders folgenreich hervorheben. Das Fehlen von Ground-Truth-Labels in unüberwachten Settings erschwert die Anwendung von Standard-Fairness-Metriken (siehe Abschnitt 3.3) grundlegend, da konventionelle Ansätze wie Equalized Odds oder Calibration Parity den Zugriff auf eine Outcome-Variable voraussetzen, gegen die Modellvorhersagen benchmarked werden können.

3.2.4 Large Language Models in der klinischen Praxis: Eine aufkommende Bias-Front

Die rasche Einführung großer Sprachmodelle in klinischen und bildungsbezogenen Gesundheitskontexten führt zu einer eigenständigen Reihe von Bias-Erwägungen, die sich mit der oben untersuchten klassischen klinischen-ML-Fairness-Literatur überschneiden, von ihr aber nicht vollständig erfasst werden. Li et al. (2025) dokumentieren in einer Scoping-Review zur Integration von KI und LLMs in die medizinische Ausbildung, dass Tools wie ChatGPT zwar echte pädagogische Vorteile bieten – darunter personalisierte Lernerfahrungen für Medizinstudierende –, ihre Integration jedoch eigenständige ethische Bedenken hinsichtlich der Weitergabe klinischer Stereotype und verzerrter Denkmuster an Auszubildende aufwirft, die möglicherweise noch nicht über das klinische Urteilsvermögen verfügen, um zu erkennen, wenn eine LLM-Ausgabe Bias statt fundierter medizinischer Argumentation widerspiegelt.

Woo et al. (2026), mit Fokus speziell auf den Einsatz von LLMs in der klinischen Dokumentation, stellen fest, dass diese Tools zwar erhebliches Potenzial bieten, die Dokumentationslast und das Burnout-Risiko von Klinikerinnen und Klinikern durch

automatisierte Zusammenfassung und Strukturierung klinischer Notizen zu reduzieren, die Bias-Risiken LLM-generierter klinischer Dokumentation in der aktuellen Literatur jedoch unzureichend charakterisiert bleiben. Dies ist eine methodisch bedeutsame Lücke: Klinische Dokumentation fließt direkt in nachgelagerte Entscheidungsunterstützungssysteme und Abrechnungs-/Kodierungsprozesse ein, was bedeutet, dass auf der Dokumentationsebene eingeführter Bias – etwa systematische Unterschiede darin, wie ein LLM Schmerzbeschreibungen oder Symptomschwere über demografische Gruppen hinweg charakterisiert – sich durch mehrere nachgelagerte klinische und administrative Systeme fortpflanzen könnte, ohne explizit als Fairness-Bedenken markiert zu werden.

KI-gestützte Tools haben für die psychische Gesundheit, einschließlich Konversationsagenten, haben generell positive Effekte auf psychische Belastung und Ergebnisse der psychischen Gesundheit, jedoch weist die Übersichtsarbeit auf eine fast vollständige Abwesenheit demografisch stratifizierter Ergebnisberichterstattung in den zugrunde liegenden Primärstudien hin – was es derzeit unmöglich macht zu bestimmen, ob diese Vorteile gerecht über Patienten-Subgruppen verteilt sind, eine Lücke, die mit dem breiteren, in diesem Abschnitt durchgängig dokumentierten Muster unzureichender Subgruppenanalyse übereinstimmt (Humayun et al. (2025).

3.2.5 Regulatorische und Governance-Dimensionen speziell für Gesundheits-KI

Mehrere untersuchte Studien passt

betonen, dass Bias in der Gesundheits-KI angesichts des eigenständigen regulatorischen Umfelds des Sektors nicht allein durch technische Mittel adressiert werden kann. Kale et al. (2024) untersuchen die Meldung unerwünschter Ereignisse für KI-basierte Medizinprodukte in Regulierungsdatenbanken und stellen erhebliche Untererfassung fest, die darauf zurückzuführen ist, dass nicht erkannt wird, dass ein unerwünschtes Ereignis auf algorithmischen Bias zurückzuführen sein könnte, statt auf eine konventionelle Geräte-Fehlfunktion – ein Befund mit direkten Implikationen für Post-Market-Surveillance-Anforderungen unter Rahmenwerken wie dem EU AI Act (diskutiert in Kapitel 4). Beim Einsatz von KI/ML in der Pflegepraxis wird ähnlich gezeigt, dass klinisches Personal an der Front häufig nicht über die Schulung verfügt, algorithmischen Bias zu identifizieren oder zu melden, selbst wenn dessen nachgelagerte Auswirkungen auf die Patientenversorgung beobachtet werden (Mohammed et al., 2025).

De und Lohani (2026) untersuchen die regulatorische Übernahme von KI, maschinellem Lernen und computergestützter Modellierung in In-silico-klinischen Studien für Medizinprodukte und stellen fest, dass regulatorische Rahmenwerke zur Bewertung von Fairness in dieser rasch wachsenden Studienmethodik noch in einem frühen Entwicklungsstadium sind – computergestützte Studiensimulationen riskieren, in den historischen klinischen Studiendaten vorhandene Biases zu übernehmen und zu verstärken, die zur Validierung der zugrunde liegenden Modelle verwendet werden, wobei nur begrenzte regulatorische Leitlinien dazu verfügbar sind, wie ein solcher Bias vor der Zulassung bewertet werden sollte.

3.2.6 Zusammenfassung und Synthese

Die in diesem Abschnitt untersuchte Literatur zu Bias in der Gesundheits-KI konvergiert auf mehrere robuste Schlussfolgerungen. Erstens ist Bias im klinischen maschinellen Lernen mechanistisch auf spezifische, identifizierbare Punkte in der Daten- und Modellierungspipeline zurückführbar – Ungleichgewicht der Trainingsdaten, Label-Bias und Proxy-Variablen (Colacci et al., 2025) – statt eine unerklärliche emergente Eigenschaft komplexer Modelle zu sein. Zweitens werden die Konsequenzen von Bias in der Gesundheits-KI empirisch als materiell statt rein theoretisch nachgewiesen: Verzerrte diagnostische KI kann klinische Entscheidungsfindung unmittelbar verschlechtern (Jabbour et al., 2023), und verzerrte Vorhersagemodelle funktionieren systematisch schlechter für Bevölkerungsgruppen, die bereits eingeschränkten Zugang zur Gesundheitsversorgung haben (Schuch et al., 2023). Drittens entwickeln sich Erkennungs- und Mitigation-Techniken weiter, einschließlich vielversprechender Ansätze wie Subgroup-Learnability-Rahmenwerken (Gulamali et al., 2025), doch die praktische Übertragung auf reale, nicht kuratierte klinische Daten bleibt eine bedeutende ungelöste Herausforderung (Huang et al., 2024).

Mehrere Forschungslücken treten aus dieser Synthese klar hervor. Fairness im unüberwachten Lernkontext ist gegenüber überwachten Settings deutlich unterentwickelt, trotz wachsendem klinischem Einsatz (Adouani et al., 2026). Die Bias-Implikationen LLM-basierter klinischer Dokumentation und Entscheidungsunterstützung bleiben trotz rascher Einführung weitgehend uncharakterisiert (Woo et al., 2026; Li et al., 2025). Schließlich hinkt die regulatorische und Berichterstattungsinfrastruktur für Bias in der Gesundheits-KI – einschließlich Datenbanken für unerwünschte Ereignisse und Schulung klinischen Personals – sowohl der technischen Fähigkeit zur Bias-Erkennung als auch der Geschwindigkeit klinischer KI-Einführung

hinterher (Kale et al., 2024; Mohammed et al., 2025). Diese regulatorischen Lücken werden in Kapitel 4 vertieft untersucht, während technische Mitigation-Ansätze mit Relevanz für das Gesundheitswesen gemeinsam mit allgemeinen Mitigation-Strategien in Abschnitt 3.4 synthetisiert werden.

3.3 Fairness-Metriken und theoretische Grundlagen

Zu definieren, was einen „fairen“ Algorithmus ausmacht, ist nicht bloß eine technische, sondern eine normative Frage: Unterschiedliche, mathematisch präzise Fairness-Definitionen verkörpern unterschiedliche ethische Verpflichtungen darüber, was Gleichheit in einem gegebenen Kontext bedeuten sollte. Dieser Abschnitt synthetisiert die theoretischen Grundlagen der algorithmischen Fairness-Messung, die gut dokumentierten Unvereinbarkeiten zwischen konkurrierenden Fairness-Kriterien sowie die empirische Anwendung dieser Metriken über die in dieser Untersuchung erfassten Gesundheits- und KI-Bereiche hinweg.

3.3.1 Die Pluralität der Fairness-Definitionen

Die Vermehrung konkurrierender mathematischer Fairness-Definitionen haben in der Informatik-Literatur eine tiefere und grundlegendere Uneinigkeit darüber widerspiegelt, was Fairness selbst bedeutet – eine Uneinigkeit, die der algorithmischen Entscheidungsfindung lange vorausgeht und in unterschiedlichen philosophischen Traditionen distributiver und korrektiver Gerechtigkeit wurzelt (Hellman, 2020). Hellmans zentraler Beitrag besteht im Nachweis, dass zwei breite Familien von Fairness-Maßen – jene, die gleiche Vorhersagegenauigkeit über Gruppen hinweg fordern, und jene, die gleiche Behandlung unter vergleichbaren Umständen fordern – nicht bloß technische Varianten eines einzigen zugrunde liegenden Konzepts sind, sondern substanziell unterschiedliche ethische Verpflichtungen widerspiegeln, die sich nicht allein durch technische Mittel auflösen lassen.

Dwork et al. (2012) formalisieren in der grundlegenden Arbeit, die das Rahmenwerk der Individual-Fairness einführt, diese Unterscheidung auf technischer Ebene, indem sie vorschlagen, Fairness als Lipschitz-Bedingung zu verstehen: Ähnliche Individuen, gemessen anhand einer aufgabenspezifischen Ähnlichkeitsmetrik, sollten ähnliche algorithmische Ergebnisse erhalten. Dies steht im Gegensatz zu Gruppen-Fairness-Kriterien wie der statistischen Parität, die aggregierte Ergebnisse über demografische Gruppen hinweg beschränken, ohne individuelle Ähnlichkeit zu berücksichtigen. Das Individual-Fairness-

Rahmenwerk verschiebt die zentrale normative Herausforderung von der Definition von Gruppengrenzen zur Definition einer angemessenen Ähnlichkeitsmetrik – eine Herausforderung, die die Autorinnen und Autoren selbst als wertgeladen anerkennen, da die Metrik bestimmt, welche Individuen als „ähnlich situiert“ für Zwecke der fairen Behandlung gelten.

Pessach und Shmueli (2022) liefern in einer breit zitierten ACM-Computing-Surveys-Übersichtsarbeit eine systematische Taxonomie, die Group Fairness, Individual Fairness und Counterfactual Fairness als die drei Hauptfamilien von Fairness-Definitionen in der Machine-Learning-Literatur unterscheidet und mehr als zwanzig eigenständige formale Metriken über diese Familien hinweg katalogisiert. Ihre Übersichtsarbeit zeigt, dass keine einzelne Metrik universell angemessen ist; die Wahl des Fairness-Kriteriums ist notwendigerweise kontextabhängig, geprägt durch den spezifischen Entscheidungsbereich, die Natur des abzuwendenden Schadens (z. B. Disparate Treatment versus Disparate Impact) und welche Stakeholder-Perspektive priorisiert wird.

3.3.2 Formale Unvereinbarkeit von Fairness-Kriterien

Ein zentrales und wiederkehrendes Thema in der theoretischen Fairness-Literatur ist die mathematische Unmöglichkeit, mehrere, je für sich plausible Fairness-Kriterien simultan zu erfüllen, außer unter eng begrenzten und empirisch seltenen Bedingungen. Kleinberg et al. (2017) beweisen in einem grundlegenden, in der nachfolgenden Literatur breit zitierten Resultat formal, dass drei wünschenswerte Fairness-Eigenschaften – Calibration innerhalb von Gruppen, Balance für die positive Klasse und Balance für die negative Klasse – nicht gemeinsam erfüllt werden können, außer im Spezialfall, dass die Basisraten des interessierenden Ergebnisses zwischen den verglichenen Gruppen identisch sind, oder wenn die Vorhersage perfekt akkurat ist. Da reale Basisraten für klinische Ergebnisse, Kreditwürdigkeit oder Rückfallrisiko zwischen demografischen Gruppen praktisch immer aufgrund echter zugrunde liegender Ungleichheiten differieren (selbst häufig Produkt historischer Diskriminierung), hat dieses Unmöglichkeitsresultat unmittelbare und unausweichliche praktische Implikationen: Jeder eingesetzte Algorithmus mit unterschiedlichen Basisraten zwischen Gruppen muss notwendigerweise mindestens eine dieser drei Fairness-Eigenschaften verletzen.

Chouldechova (2017) leitet unabhängig ein eng verwandtes Unmöglichkeitsresultat im spezifischen Kontext der Rückfallrisikovorhersage ab und zeigt, dass die Angleichung von

Falsch-Positiv- und Falsch-Negativ-Raten zwischen Gruppen (eine Form von Equalized Odds) generell unvereinbar mit dem Erreichen von Calibration innerhalb jeder Gruppe ist, außer bei gleichen Basisraten. Dieses Resultat wurde als direkte Antwort auf die breit publizierte ProPublica-Untersuchung des COMPAS-Rückfallrisiko-Algorithmus entwickelt, die festgestellt hatte, dass das Tool über Rassengruppen hinweg kalibriert war, während es gleichzeitig erheblich unterschiedliche Falsch-Positiv-Raten für Schwarze versus weiße Angeklagte aufwies – ein Muster, das Kritikerinnen und Kritiker als diskriminierend charakterisierten, das die Entwickler von COMPAS jedoch mit der Begründung verteidigten, dass das Tool kalibriert sei. Chouldechovas formale Analyse zeigt, dass dieser scheinbare Widerspruch kein für COMPAS spezifischer Mangel ist, sondern eine mathematische Notwendigkeit, wann immer Basisraten differieren, und verschiebt die Debatte von „ist dieser bestimmte Algorithmus verzerrt“ zu „welche Fairness-Eigenschaft sollte priorisiert werden, da nicht alle simultan erfüllt werden können.“

Corbett-Davies et al. (2023) erweitern in einer umfangreichen, im Journal of Machine Learning Research veröffentlichten kritischen Analyse diese Forschungslinie, indem sie argumentieren, dass sich ein Großteil der algorithmischen-Fairness-Literatur unverhältnismäßig auf eine enge Menge von Kriterien auf Basis statistischer Parität konzentriert hat, während eine grundlegendere Frage vernachlässigt wird: Was versucht die entscheidende Instanz tatsächlich zu erreichen, und fördert oder unterminiert die Optimierung auf eine gegebene Fairness-Metrik dieses zugrunde liegende Ziel? Ihre Analyse ist bemerkenswert kritisch gegenüber der Behandlung von Fairness-Metriken als Selbstzweck und argumentiert stattdessen, dass Fairness-Kriterien gegen ihre Konsequenzen für das Wohlergehen betroffener Gruppen bewertet werden sollten – eine Position, die erhebliche Debatten ausgelöst hat, da sie impliziert, dass die Erfüllung einer formalen Fairness-Metrik nicht garantiert, dass betroffene Gruppen tatsächlich besser gestellt sind.

Tabelle 3.5: Zentrale Fairness-Metriken und ihre Inkompatibilitäten

Metrik	Definition	Zentrale Limitation
Statistical Parity	Gleiche Rate positiver Vorhersagen über alle Gruppen hinweg ($P(\hat{Y}=1 \mid A=0) = P(\hat{Y}=1 \mid A=1)$)	Ignoriert tatsächliche Outcome-Unterschiede; kann zu Reverse Discrimination führen
Equalized Odds	Gleiche TPR und FPR über Gruppen hinweg (Hardt et al., 2016)	Kann mit Calibration unvereinbar sein, wenn Basisraten zwischen Gruppen differieren (Chouldechova, 2017)
Calibration	Vorhergesagte Wahrscheinlichkeiten entsprechen tatsächlichen Outcome-Raten innerhalb jeder Gruppe	Inkompatibel mit Equalized Odds bei unterschiedlichen Basisraten (Kleinberg et al., 2017)
Individual Fairness	Ähnliche Individuen erhalten ähnliche Vorhersagen (Dwork et al., 2012)	Erfordert eine explizite Ähnlichkeitsmetrik, deren Definition selbst normativ ist
Counterfactual Fairness	Vorhersage bleibt gleich in einer kontrafaktischen Welt ohne das sensible Attribut (Kusner et al., 2017)	Setzt ein korrektes Kausalmodell voraus, das in der Praxis selten vollständig bekannt ist

Anmerkung. Die Tabelle fasst die in Abschnitt 3.3.1–3.3.2 diskutierten Kernmetriken zusammen. *Equalized Odds* und *Calibration* können bei unterschiedlichen Basisraten nicht gleichzeitig erfüllt werden (Kleinberg et al., 2017; Chouldechova, 2017).

3.3.3 Kausalbasierte Fairness-Konzepte: Kontrafaktische Ansätze

Mit dem Konzept der *Counterfactual Fairness* liegt eine kausal fundierte Alternative zu rein statistischen Fairness-Definitionen vor: Eine Entscheidung gilt dann als fair, wenn sie in einer kontrafaktischen Welt, in der das sensible Attribut eines Individuums — etwa Rasse oder Geschlecht — anders gewesen wäre, unverändert bliebe, während alles kausal Unabhängige davon konstant gehalten wird (Kusner et al., 2017). Diese kausale Rahmung adressiert eine bekannte Limitation rein statistischer Gruppen-Fairness-Kriterien: Sie kann prinzipiell zwischen Ungleichheiten unterscheiden, die durch illegitime kausale Pfade entstehen (z. B. Rasse, die eine Kreditentscheidung direkt beeinflusst), und solchen, die durch legitime, mit dem sensiblen Attribut korrelierte Pfade entstehen (z. B. Einkommensunterschiede, die selbst Folge historischer Diskriminierung sind, aber im Entscheidungspfad des Modells kausal von Rasse getrennt werden). Die zentrale Limitation des Ansatzes ist seine Abhängigkeit von einem

korrekt spezifizierten Kausalmodell des datengenerierenden Prozesses — eine Annahme, die in komplexen realen Bereichen wie dem Gesundheitswesen oder der Strafjustiz selten mit Sicherheit verifizierbar ist.

3.3.4 Empirische Anwendung von Fairness-Metriken im Gesundheitswesen

Mehrere in dieser Untersuchung identifizierte Studien wenden formale Fairness-Metriken auf konkrete klinische Vorhersageaufgaben an und liefern damit ein empirisches Gegenstück zur oben diskutierten theoretischen Literatur. Ganta et al. (2024), veröffentlicht in JAMA Network Open, evaluieren Fairness in Machine-Learning-Modellen zur Vorhersage der Krebsmortalität über Rassen-Subgruppen hinweg und prüfen explizit, ob unterschiedliche Modellleistung – ein Vorläufer allokativer Schäden in der klinischen Entscheidungsfindung – zwischen Rassengruppen im Kontext der Priorisierung von Patientinnen und Patienten für Gespräche über schwere Erkrankungen besteht. Ihre Befunde zeigen messbare Leistungsunterschiede und unterstreichen die praktische Relevanz der in Abschnitt 3.3.2 diskutierten theoretischen Unmöglichkeitsergebnisse: Die Optimierung eines Krebsmortalitätsmodells auf Gesamtgenauigkeit garantiert keine gerechte Leistung über die von Krebsergebnis-Ungleichheiten am stärksten betroffenen Rassen-Subgruppen.

Yuan et al. (2023), ebenfalls veröffentlicht in JAMA Network Open, führen eine parallele Fairness-Evaluation im Kontext von Modellen zur Alzheimer-Krankheitsprogression durch und charakterisieren explizit algorithmische Fairness-Eigenschaften von Machine-Learning-Modellen, die zur Früherkennung und zum Krankheitsmanagement eingesetzt werden. Ihre Studie ist methodisch bemerkenswert, weil sie mehrere Fairness-Metriken simultan evaluiert, statt sich auf ein einziges Kriterium zu verlassen, und damit in einem angewandten klinischen Kontext dieselbe, auf theoretischer Ebene von Pessach und Shmueli (2022) identifizierte Herausforderung der Definitionspluralität illustriert: Modelle, die ein Fairness-Kriterium (z. B. Calibration) erfüllten, erfüllten nicht notwendigerweise andere (z. B. Equalized Odds), was die Unmöglichkeitsergebnisse von Kleinberg et al. (2017) und Chouldechova (2017) direkt in einer konkreten klinischen Anwendung widerspiegelt.

Carbunaru et al. (2024) entwickeln und evaluieren ein fairness-explizites Machine-Learning-Modell zur Vorhersage des Überlebens nach radikaler Zystektomie bei Blasenkrebs, wobei Fairness-Constraints explizit während der Modellentwicklung integriert werden, statt Fairness nur post-hoc zu evaluieren. Dies stellt eine methodisch bedeutsame Unterscheidung innerhalb der breiteren Literatur dar: Die meisten in dieser Untersuchung identifizierten Fairness-

Evaluationen bewerten die Fairness bereits trainierter Modelle (Post-hoc-Auditing), während eine kleinere Untergruppe, einschließlich dieser Studie, Fairness-Erwägungen direkt in das Trainingsziel des Modells integriert – ein Ansatz, der den in Abschnitt 3.4 diskutierten In-Processing-Mitigation-Strategien näher steht.

Über die klinische Vorhersage hinaus: Datensätze und Interpretierbarkeit

In Nature adressiert Xiang et al. (2025) eine grundlegende und häufig übersehene Quelle von Unfairness: die Zusammensetzung der Benchmark-Datensätze selbst, die zur Evaluation von Computer-Vision- und breiteren KI-Systemen verwendet werden. Ihre Arbeit führt einen fairen, menschenzentrierten Bilddatensatz ein, der explizit für ethisches KI-Benchmarking entworfen wurde, motiviert durch die Beobachtung, dass weit verbreitete Benchmark-Datensätze in der Computer Vision historisch ohne ausreichende Berücksichtigung demografischer Repräsentativität erhoben wurden – was bedeutet, dass selbst mit formal korrekten Metriken durchgeführte Fairness-Evaluationen irreführend günstige Ergebnisse liefern können, wenn die zugrunde liegenden Benchmark-Daten die von einem eingesetzten System betroffenen Bevölkerungsgruppen nicht angemessen repräsentieren. Dies verbindet sich direkt mit der in Abschnitt 3.1.1 diskutierten Representation-Bias-Kategorie und unterstreicht, dass Fairness-Metriken, wie theoretisch fundiert sie auch sein mögen, nur so aussagekräftig sind wie die Daten, gegen die sie evaluiert werden.

Das Fairness-Evaluationsinstrumentarium wird von Basu et al. (2026) in den Bereich der Modellinterpretierbarkeit erweitert, indem sie ein mechanistisches Interpretierbarkeits-Audit einer Reinforcement-Learning-Policy durchführen, die in der Medicaid-Versorgungskoordination eingesetzt wird. Ihr Ansatz – der Aufmerksamkeitsanalyse, Shapley-Value-Erklärungen und Sparse-Autoencoder-Feature-Discovery kombiniert – entwickelt fairness-bewusste Schutzmechanismen, indem zunächst die internen Argumentationspfade des Modells offengelegt werden und anschließend identifiziert wird, welche dieser Pfade ethisch problematischen Verwendungen sensibler oder Proxy-Attribute entsprechen. Dieser interpretierbarkeitsorientierte Ansatz für Fairness-Audits stellt eine aufkommende methodische Richtung dar, die über ergebnisbasierte Fairness-Metriken hinausgeht und sich dem Verständnis der Mechanismen zuwendet, durch die ein Modell zu potenziell verzerrten Entscheidungen kommt.

3.3.5 Zusammenfassung und Synthese

Die in diesem Abschnitt untersuchte theoretische und empirische Literatur konvergiert auf eine zentrale, häufig unterschätzte Schlussfolgerung: Es gibt kein einzelnes, universell korrektes Maß algorithmischer Fairness, und die zentralen Unmöglichkeitsresultate von Kleinberg et al. (2017) und Chouldechova (2017) zeigen, dass diese Pluralität kein vorübergehender Zustand wissenschaftlicher Unreife ist, der durch weitere Forschung gelöst werden kann, sondern eine mathematische Notwendigkeit, wann immer Gruppen-Basisraten differieren – das heißt, in praktisch allen realen Anwendungen, die historisch marginalisierte Bevölkerungsgruppen betreffen. Dies hat eine unmittelbare und bedeutsame Implikation für die Praxis algorithmischer Fairness: Jede Wahl einer Fairness-Metrik ist notwendigerweise auch eine implizite normative Entscheidung darüber, welche Form von Unfairness ein Systemdesigner zu tolerieren bereit ist, um eine andere zu vermeiden.

Die in Abschnitt 3.3.4 untersuchten empirischen Anwendungen zeigen, dass diese theoretische Pluralität kein rein akademisches Anliegen ist: Studien, die reale klinische Vorhersagemodelle evaluieren (Ganta et al., 2024; Yuan et al., 2023), stellen konsistent fest, dass die Erfüllung eines Fairness-Kriteriums die Erfüllung anderer nicht garantiert, was die praktische Relevanz der formalen Unmöglichkeitstheoreme direkt bestätigt. Eine weitere, sich abzeichnende Erkenntnis, illustriert durch Xiang et al. (2025) und Basu et al. (2026), ist, dass sich die Fairness-Messung zunehmend über ergebnisbasierte statistische Metriken hinaus erweitert und sich der Fairness zugrunde liegender Datensätze sowie der Interpretierbarkeit der Mechanismen widmet, die potenziell verzerrte Ergebnisse produzieren. Diese theoretischen Grundlagen fließen unmittelbar in die im folgenden Abschnitt (3.4) diskutierte Evaluation von Mitigation-Strategien ein, da die Wahl der Fairness-Metrik notwendigerweise bestimmt, welcher Mitigation-Ansatz als erfolgreich gilt.

3.4 Mitigationsstrategien

Nachdem die theoretischen Grundlagen von Bias (Abschnitt 3.1), seine empirische Ausprägung im Gesundheitswesen (Abschnitt 3.2) und die Metriken, mit denen Fairness gemessen werden kann (Abschnitt 3.3), etabliert wurden, widmet sich dieser Abschnitt der zentralen praktischen Frage: Wie kann Bias in KI-Systemen reduziert oder korrigiert werden? Die Mitigation-Literatur ist konventionell entlang der Stufe der Machine-Learning-Pipeline organisiert, an der eine Intervention erfolgt – Pre-Processing (Datenebene), In-Processing (Algorithmusebene) und Post-Processing (Outputebene) –, eine Taxonomie, die dieser Abschnitt übernimmt,

ergänzt um eine vierte Kategorie für organisatorische und soziale Interventionen, die außerhalb der technischen Pipeline selbst operieren.

3.4.1 Pre-processing: Bias auf der Datenebene adressieren

Pre-Processing-Ansätze setzen vor dem Modelltraining an, ausgehend von der Annahme, dass verzerrte Ergebnisse häufig aus nicht repräsentativen oder unausgewogenen Trainingsdaten entstehen statt aus dem Lernalgorithmus selbst. Dominguez-Catena et al. (2024) entwickeln in einer in IEEE Transactions on Pattern Analysis and Machine Intelligence veröffentlichten Studie ein umfassendes Set von Metriken, die speziell zur Quantifizierung demografischen Bias innerhalb von Datensätzen entwickelt wurden, anhand der Fallstudie der Gesichtsausdruckserkennung. Ihre Arbeit leistet einen wichtigen methodischen Beitrag, indem sie zwischen mehreren eigenständigen Formen demografischen Ungleichgewichts auf Datensatzebene unterscheidet – etwa Unterschiede in der reinen Repräsentation demografischer Gruppen versus Unterschiede in der Konsistenz, mit der Bilder jeder Gruppe gelabelt werden – und zeigt, dass ein Datensatz nach einem Maß demografisch ausgewogen erscheinen kann, während er nach einem anderen Maß systematisch verzerrt bleibt. Dieser Befund hat unmittelbare Implikationen für die Pre-Processing-Mitigation: Interventionen wie Resampling oder Reweighting, die nur die grobe Gruppenrepräsentation adressieren, können subtilere Ungleichgewichte auf Label-Ebene nicht korrigieren.

Li et al. (2026) erweitern die Erkennung von Datensatz-Bias von einem statischen, metrikbasierten Ansatz zu einem interaktiven visuellen Analysetool, BiasField, das es Praktikerinnen und Praktikern ermöglicht, Machine-Learning-Datensätze interaktiv auf unfaire Attributassoziationen zu untersuchen – etwa das gemeinsame Auftreten demografischer Marker mit beruflichen oder kontextuellen Labels in Bilddatensätzen – statt sich ausschließlich auf vorberechnete Aggregatstatistiken zu verlassen. Die Autorinnen und Autoren argumentieren, dass interaktive Untersuchung subtilere und kontextspezifischere Biases offenlegt als fixe Metriken allein, wenngleich dies den Bedarf an Fachexpertise erfordert, um die Visualisierungen sinnvoll zu interpretieren, was die Zugänglichkeit des Tools für nicht-technische Stakeholder einschränkt, die dennoch am besten positioniert sein könnten, um problematische Assoziationen innerhalb ihrer Domäne zu erkennen.

Eine gesundheitswesen-spezifische Pre-Processing-Herausforderung ist der Data-Center-Bias bei Krebsklassifikationsmodellen und zur Mitigation dessen schlagen Kheiri et al. (2026) eine

vor, bei denen Modelle, die auf Bildgebungsdaten einer Gesundheitseinrichtung trainiert wurden, systematisch nicht auf Daten anderer Einrichtungen generalisieren, aufgrund versteckter, aufgabenirrelevanter Attribute, die mit dem ursprünglichen Datenzentrum korrelieren (etwa Unterschiede in der Bildgebungs-ausrüstung oder Patientenpopulation). Ihr Transfer-Bias-Unlearning-Ansatz entfernt explizit datenzentrum-identifizierende Signale aus gelernten Feature-Repräsentationen mittels eines Mehrzieloptimierungs-Rahmenwerks, das frei von konfliktierenden Zielen zwischen Fairness und diagnostischer Genauigkeit konzipiert ist – und adressiert damit direkt eine spezifische und unterbeforschte Quelle des von Dominguez-Catena et al. (2024) allgemeiner katalogisierten Bias auf Datensatzebene.

3.4.2 In-processing: Fairness-Constraints während des Modelltrainings

In-Processing-Ansätze modifizieren den Lernalgorithmus selbst, typischerweise durch Integration eines expliziten Fairness-Ziels oder -Constraints neben dem konventionellen, genauigkeitsmaximierenden Trainingsziel. Kim und Lee (2025) schlagen AmbiBias Contrast vor, eine Debiasing-Technik, die auf dem Prinzip der Entflechtung eines gelernten Repräsentationsraums in Bias-bezogene und Ambiguitäts-bezogene Cluster basiert, wodurch das Modell selektiv die Sensitivität gegenüber spurious Korrelationen zwischen einem Zielattribut und kontextuellen Bias-Attributen reduzieren kann, während die Sensitivität gegenüber genuin prädiktiven Merkmalen erhalten bleibt. Die Autorinnen und Autoren positionieren ihren Ansatz als Verbesserung gegenüber konventionellem, reweighting-basiertem Debiasing, das ihrer Argumentation nach die Gesamtklassifikationsgenauigkeit verschlechtern kann, wenn Bias- und Zielattribute in der zugrunde liegenden Feature-Repräsentation stark verschränkt sind – eine verbreitete Limitation einfacherer In-Processing-Methoden, die dieser Ansatz explizit durch Repräsentationsentflechtung zu überwinden sucht.

Der oben primär als Pre-Processing-Intervention diskutierte Krebsklassifikations-Ansatz von Kheiri et al. (2026) integriert durch seine Mehrzieloptimierung der internen Feature-Repräsentationen des Modells während des Trainings auch In-Processing-Elemente, was illustriert, dass die Grenze zwischen den Kategorien Pre-Processing und In-Processing in der Praxis nicht immer scharf ist – viele zeitgenössische Mitigation-Ansätze operieren simultan über mehrere Pipeline-Stufen hinweg, statt ihre Intervention auf eine einzelne, klar abgegrenzte Stufe zu beschränken.

3.4.3 Post-Processing: Korrektur von Outputs nach dem Training

Post-Processing-Ansätze setzen an, nachdem ein Modell bereits trainiert wurde, und passen dessen Ausgaben oder Entscheidungsschwellen an, ohne dass ein erneutes Training erforderlich ist – ein Ansatz, der häufig praktikabler ist, wenn ein erneutes Training kostspielig ist oder wenn die zugrunde liegende Modellarchitektur proprietär ist und nicht verändert werden kann. Schnetler et al. (2025), in einer retrospektiven Multisite-Kohortenstudie veröffentlicht in *BMJ Quality & Safety*, liefern eine klinisch fundierte empirische Evaluation von Post-Processing-Mitigation im Kontext der Sepsis-Vorhersage. Ihr zentraler Befund stellt eine verbreitete Annahme beim Einsatz klinischer KI in Frage: dass ein einzelnes, generalisierbares Sepsis-Vorhersagemodell einheitlich über mehrere Krankenhäuser und Versorgungssettings hinweg angewendet werden kann. Die Studie zeigt, dass die Modellleistung je nach Standort erheblich variiert, aufgrund von Unterschieden in Patientenpopulationen, klinischen Arbeitsabläufen und Dokumentationspraktiken, und evaluiert systematisch sechs eigenständige Bias-Mitigation-Strategien, um zu identifizieren, welche Kombination von Ansätzen das leistungsstärkste Modell über Standorte hinweg liefert. Ihre Schlussfolgerung – dass keine einzelne Mitigation-Strategie universell optimal ist und Mitigation-Ansätze empirisch innerhalb jedes spezifischen Einsatzkontexts validiert werden müssen – hat erhebliche praktische Implikationen für Gesundheitseinrichtungen, die extern entwickelte KI-Modelle übernehmen möchten, und legt nahe, dass lokale Validierung und standortspezifische Rekalibrierung Standardpraxis statt optionale Verfeinerung sein sollten.

Tabelle 3.6: Übersicht der Mitigation-Strategietypen

Strategietyp	Beschreibung	Beispielstudien	Trade-off
Pre-processing	Korrektur der Trainingsdaten vor dem Modelltraining (Reweightings, Resampling, Repräsentationslernen)	Dominguez-Catena et al. (2024); Li et al. (2026)	Modellunabhängig, aber adressiert nicht algorithmische Verstärkungseffekte
In-processing	Integration von Fairness-Constraints direkt in die Trainingsobjektivfunktion	Kim & Lee (2025); Kheiri et al. (2026)	Oft bester Accuracy-Fairness-Trade-off, aber modellspezifisch und rechenintensiv

Post-processing	Anpassung der Modell-Outputs/Schwellenwerte nach dem Training, ohne Retraining	Schnetler et al. (2025)	Einfach nachrüstbar, aber kann Calibration verschlechtern
Organisatorisch/Sozial	Prozessuale und institutionelle Maßnahmen (Audits, interdisziplinäre Teams, Transparenzpflichten)	Hickman et al. (2025); Ukanwa (2024)	Adressiert systemische Ursachen, aber schwer zu standardisieren/messen

Anmerkung. Die Kategorisierung folgt der etablierten Dreiteilung in Pre-/In-/Post-Processing (vgl. Mehrabi et al., 2021), ergänzt um eine organisatorisch-soziale Kategorie.

3.4.4 Organisatorische und soziale Dimensionen der Mitigation

Eine wiederkehrende Erkenntnis über mehrere untersuchte Studien hinweg ist, dass rein technische Mitigation-Strategien, wie methodisch ausgereift sie auch sein mögen, unzureichend sind, um den vollen Umfang algorithmischen Bias ohne ergänzende organisatorische und soziale Interventionen zu adressieren. Hickman et al. (2025), in einer integrativen, im *Journal of Applied Psychology* veröffentlichten systematischen Übersichtsarbeit, untersuchen Mitigation algorithmischen Bias speziell im Kontext von Personalbeurteilungs- und Auswahlinstrumenten wie automatisierten Lebenslauf-Screenern und automatisch bewerteten Interviews. Ihre Übersichtsarbeit ist bemerkenswert, weil sie explizit die Literatur der Informatik mit jener der Organisationspsychologie verbindet, und argumentiert, dass wirksame Bias-Mitigation in Beschäftigungskontexten nicht nur technische Debiasing-Methoden, sondern auch organisatorische Praktiken wie Adverse-Impact-Monitoring, strukturierte Validierungsstudien vor dem Einsatz und fortlaufendes Auditing erfordert – Praktiken mit etabliertem Vorlauf in der Arbeits- und Organisationspsychologie, die in der technischen Fairness-Literatur vergleichsweise unterausgeschöpft geblieben sind.

Ukanwa (2024) erweitert dieses Argument durch das vorgeschlagene 3-D Dependable AI Framework, das explizit eine stärkere Integration psychologischer und sozialwissenschaftlicher Forschung in die Untersuchung der Vorbedingungen und Konsequenzen algorithmischen Bias fordert. Die zentrale These dieser Arbeit ist, dass algorithmischer Bias in der Informatik-Literatur häufig als rein technisches Artefakt behandelt wird, das durch statistische Mittel zu korrigieren ist, während tatsächlich viele Fälle algorithmischen Bias die Kodierung bereits bestehender sozialer und psychologischer

Vorurteile widerspiegeln – was bedeutet, dass dauerhafte Mitigation das Verständnis und die Adressierung der menschlichen und institutionellen Prozesse erfordert, die verzerrte Daten und verzerrte Designentscheidungen überhaupt erst hervorbringen, nicht nur die nachgelagerte statistische Signatur dieses Bias in einem trainierten Modell.

Transparenz spielt in der Forschung zur KI-gestützten Pflegeausbildung eine unterschätzte Rolle. Rahman (2026) zeigt, dass Lehrende und Studierende nicht nur dann Vertrauen in ein KI-Tool entwickeln, wenn es statistisch fair ist, sondern vor allem dann, wenn sie nachvollziehen können, warum es eine bestimmte Ausgabe produziert hat. Mitigation-Strategien, die ausschließlich auf die Verbesserung von Fairness-Metriken abzielen, ohne die Erklärbarkeit des Systems zu berücksichtigen, können daher ihr eigentliches Ziel verfehlen – nämlich angemessenes Vertrauen und eine sachgerechte Nutzung durch die Anwenderinnen und Anwender zu fördern.

Technische Fairness-Interventionen stoßen an ihre Grenzen, wenn die Institutionen, in denen sie eingesetzt werden, selbst von systemischem Bias durchzogen sind. Dieser Befund geht über das KI-Design hinaus und betrifft den gesamten institutionellen Einsatzkontext – von Rechtsprechungsprozessen bis hin zu Medien-, Finanz- und Rechtsinstitutionen. Algorithmische Mitigation-Tools können in solchen Umfeldern nur begrenzt korrigierend wirken, weil sie innerhalb bestehender institutioneller Strukturen operieren und nicht unabhängig von ihnen (Javed & Li, 2025).

Domänenspezifische Mitigation: Der Fall altersbezogener Bias

Chu et al. (2024), in einer in JMIR Aging veröffentlichten Scoping-Review, adressieren eine spezifische und vergleichsweise unterbeforschte Bias-Dimension: altersbezogenen Bias, oder „digitalen Altersismus“, in Machine-Learning-Systemen. Ihre Übersichtsarbeit stellt fest, dass trotz wachsender Anerkennung altersbezogenen Bias als Problem in ML-Entwicklung und -Einsatz die Literatur, die speziell Mitigation-Strategien für diese Bias-Dimension untersucht, deutlich weniger entwickelt bleibt als für Gender- oder Rassen-Bias – eine Lücke, die mit dem breiteren, in dieser Untersuchung durchgängig festgestellten Muster übereinstimmt, dass bestimmte geschützte oder verwundbare Attribute unverhältnismäßig weniger Forschungsaufmerksamkeit erhalten als andere, unabhängig von der realen Prävalenz oder Schwere der resultierenden Schäden.

Ein methodisch relevanter Befund betrifft die Messung von Bias bei komplexen, mehrwortigen Begriffen in Sprach-Embedding-Modellen. Während einzeltoken-basierte demografische Begriffe in der Literatur gut abgedeckt sind, zeigt sich bei zusammengesetzten Konzepten und Phrasen eine deutliche Messlücke. Gray et al. (2024) weisen nach, dass mehrere gängige Bewertungsverfahren in diesem Bereich inkonsistent abschneiden – mit direkten Konsequenzen für die Mitigation: Eine Strategie kann nur dann als wirksam gelten, wenn das verwendete Messverfahren den fraglichen Bias überhaupt zuverlässig erfasst.

3.4.5 Zusammenfassung und Synthese

Die in diesem Abschnitt untersuchte Mitigation-Literatur zeigt, dass Bias-Korrektur in KI-Systemen am wirksamsten ist, wenn Interventionen auf die spezifische Stufe der Pipeline abgestimmt sind, an der Bias eingeführt wird, und wenn technische Interventionen durch organisatorische und prozessuale Schutzmaßnahmen ergänzt werden. Mehrere übergreifende Schlussfolgerungen lassen sich ableiten. Erstens ist keine einzelne Mitigation-Strategie – ob Pre-, In- oder Post-Processing – universell überlegen; die empirische Evidenz von Schnetler et al. (2025) zeigt, dass selbst innerhalb einer einzelnen technischen Kategorie die relative Wirksamkeit konkurrierender Ansätze stark kontextabhängig ist und lokal validiert werden muss, statt aus vorherigen Einsätzen angenommen zu werden. Zweitens erfordern Interventionen auf Datensatzebene (Abschnitt 3.4.1) zunehmend ausgereifte, mehrdimensionale Bias-Metriken (Dominguez-Catena et al., 2024) statt einfacher demografischer Zählungsausgleichung, da Bias in Label-Konsistenz und Feature-Assoziationen bestehen bleiben kann, selbst wenn die rohe Gruppenrepräsentation ausgeglichen erscheint. Drittens, und vielleicht am bedeutsamsten, konvergieren mehrere Studien (Ukanwa, 2024; Javed & Li, 2025; Hickman et al., 2025) auf die Schlussfolgerung, dass technische Mitigation innerhalb der breiteren sozialen und institutionellen Systeme operiert, in denen KI eingesetzt wird, und letztlich durch diese begrenzt ist – was darauf hindeutet, dass die langfristige Wirksamkeit algorithmischer Bias-Mitigation von parallelen Fortschritten in institutioneller Rechenschaftspflicht und regulatorischer Aufsicht abhängt, dem Gegenstand von Kapitel 4.

Eine bedeutende, über diese Literatur hinweg erkennbare Forschungslücke ist die ausgeprägt ungleiche Aufmerksamkeit für unterschiedliche Bias-Dimensionen: Während Mitigation-Strategien für Gender- und Rassen-Bias vergleichsweise gut entwickelt sind (vgl. Abschnitt

3.1), bleibt die Mitigation altersbezogenen Bias unterentwickelt (Chu et al., 2024), und methodische Werkzeuge zur Bewertung von Bias in komplexen, mehrwortigen sprachlichen Konstrukten hinken jenen für einfache demografische Begriffe hinterher (Gray et al., 2024). Diese Ungleichheit legt nahe, dass künftige Mitigation-Forschung derzeit unterbeforschte Bias-Dimensionen und Schadensformen priorisieren sollte, statt methodische Innovation weiterhin auf bereits vergleichsweise gut verstandene Bias-Dimensionen zu konzentrieren.

4. Regulierung und Governance

Das vorangehende Kapitel hat gezeigt, dass technische Mitigation-Strategien, obwohl notwendig, für sich allein nicht ausreichen, um fairen KI-Einsatz zu gewährleisten, da algorithmische Fairness-Interventionen innerhalb breiterer institutioneller und regulatorischer Kontexte operieren, die selbst sowohl die Anreize für Bias-Mitigation als auch die Konsequenzen ihres Fehlens prägen (vgl. Abschnitt 3.4.4). Dieses Kapitel untersucht die für Bias und Fairness in der KI relevante regulatorische und Governance-Landschaft, mit Schwerpunkt auf dem AI Act der Europäischen Union als dem derzeit umfassendsten verbindlichen Rechtsrahmen, ergänzt durch internationale technische Standards und sektorspezifische Governance-Initiativen, die in der systematischen Untersuchung identifiziert wurden.

4.1 Der EU AI Act: Ein risikobasiertes Regulierungsrahmenwerk

Die Verordnung (EU) 2024/1689, gemeinhin als AI Act bezeichnet, stellt den ersten umfassenden, verbindlichen Rechtsrahmen für Künstliche Intelligenz dar, der von einer großen Regulierungsmacht erlassen wurde. Der AI Act verfolgt einen risikobasierten Regulierungsansatz und klassifiziert KI-Systeme in vier Risikostufen – unannehmbares Risiko (vollständig verboten), Hochrisiko (umfassenden Compliance-Pflichten unterworfen), begrenztes Risiko (Transparenzpflichten unterworfen) und minimales Risiko (weitgehend unreguliert) –, wobei Fairness- und Diskriminierungsverbots-Bestimmungen primär in der Hochrisiko-Kategorie konzentriert sind.

Hochrisiko-KI-Systeme unter dem AI Act, zu denen KI im Gesundheitswesen, in der Beschäftigung, der Kreditwürdigkeitsprüfung und der Strafverfolgung zählt – genau jene Bereiche, die in den Abschnitten 3.1 bis 3.4 untersuchten empirischen Literatur am

ausführlichsten behandelt wurden –, unterliegen mehreren Anforderungen mit direkter Relevanz für algorithmische Fairness. Dazu zählen Verpflichtungen hinsichtlich der Qualität von Trainings-, Validierungs- und Testdatensätzen, die explizit fordern, dass solche Datensätze, soweit der vorgesehene Zweck dies zulässt, „relevant, repräsentativ, fehlerfrei und vollständig“ sind – eine Anforderung, die direkt die in Kapitel 3 katalogisierten Mechanismen von Representation Bias und Bias auf Datensatzebene adressiert (vgl. Dominguez-Catena et al., 2024, Abschnitt 3.4.1). Der AI Act schreibt zudem Risikomanagementsysteme über den gesamten Lebenszyklus des KI-Systems, Bestimmungen zur menschlichen Aufsicht sowie Konformitätsbewertungsverfahren vor dem Markteinsatz vor, wobei Nichteinhaltung erheblichen finanziellen Sanktionen unterliegt – ein verbindlicher Durchsetzungsmechanismus, der den AI Act von den in den nachfolgenden Abschnitten diskutierten freiwilligen technischen Standards und ethischen Leitlinien unterscheidet.

Eine bedeutende Limitation des Fairness-Ansatzes des AI Acts, implizit angedeutet durch die breitere Literatur zur Pluralität von Fairness-Metriken (vgl. Abschnitt 3.3), ist, dass die Verordnung Datenqualitäts- und Diskriminierungsverbots-Pflichten allgemein formuliert, ohne eine spezifische Fairness-Metrik oder Messmethodik vorzuschreiben. Angesichts der formalen Unmöglichkeit, mehrere Fairness-Kriterien bei unterschiedlichen Basisraten simultan zu erfüllen (Kleinberg et al., 2017; Chouldechova, 2017), ist diese regulatorische Flexibilität wohl eine praktische Notwendigkeit statt ein Versehen – eine Einheitsmetrik-Vorschrift wäre über das heterogene Spektrum der vom AI Act erfassten Hochrisiko-Anwendungen hinweg mathematisch nicht umsetzbar. Diese Flexibilität erzeugt jedoch gleichzeitig erhebliche Compliance-Unsicherheit für Organisationen, die Hochrisiko-KI-Systeme einsetzen, die selbst bestimmen müssen, welche Fairness-Kriterien für ihren spezifischen Anwendungskontext am angemessensten sind, bei nur begrenzter regulatorischer Anleitung dazu, wie diese Bestimmung getroffen und für Audit-Zwecke dokumentiert werden sollte.

4.2 Technische Standards: ISO/IEC 23894

Ergänzend zu den verbindlichen Rechtsanforderungen des AI Acts bietet ISO/IEC 23894:2023 einen freiwilligen technischen Standard für KI-Risikomanagement und stellt Organisationen einen strukturierten, prozessorientierten Rahmen zur Identifikation, Bewertung und Behandlung von Risiken über den Lebenszyklus eines KI-Systems hinweg zur Verfügung. Im Gegensatz zum AI Act, der substanzielle Ergebnisse vorschreibt, die Hochrisiko-Systeme

erreichen müssen, ist ISO 23894 prozedural ausgerichtet: Er schreibt keine bestimmten Fairness-Metriken oder Bias-Schwellenwerte vor, sondern etabliert einen systematischen Risikomanagementprozess, den Organisationen anwenden können, um fairnessbezogene Risiken als eine Kategorie innerhalb der breiteren, mit dem KI-Einsatz verbundenen Risiken zu identifizieren (neben Risiken wie Sicherheit, Robustheit und Schutz).

Die Beziehung zwischen ISO 23894 und dem AI Act ist komplementär statt konkurrierend: Der technische Standard kann als ein Mechanismus dienen, mit dem Organisationen die Einhaltung der Risikomanagementsystem-Anforderungen des AI Acts für Hochrisiko-KI nachweisen, da die Konformität mit anerkannten technischen Standards in regulatorischen Konformitätsbewertungsverfahren häufig günstig behandelt wird. Da ISO 23894 jedoch prozedural und domänen-allgemein statt substanziell vorschreibend bezüglich Fairness-Ergebnissen ist, löst er die in Abschnitt 4.1 diskutierte Metrikwahl-Unsicherheit nicht selbst auf; Organisationen müssen weiterhin selbst bestimmen, welche spezifischen Fairness-Kriterien für ihren Kontext angemessen sind, wobei der Standard einen Prozess zur Dokumentation und Begründung dieser Bestimmung bietet, statt die Bestimmung selbst zu liefern.

Tabelle 4.1: Übersicht relevanter Regulierungs- und Governance-Rahmenwerke

Rahmenwerk	Geltungsbereich	Verbindlichkeit	Fairness-relevante Inhalte
EU AI Act (2024)	Hochrisiko-Systeme (Anhang III)	Verbindliche Rechtsnorm, Bußgelder	Datenqualität, Diskriminierungsverbot, Konformitätsbewertung, CE-Kennzeichnung
ISO/IEC 23894:2023	Alle KI-Risikostufen	Freiwilliger technischer Standard	Risikomanagement-Prozess über den KI-Lebenszyklus
Modified Delphi Governance (Hassan et al., 2026)	Chirurgische KI-Systeme	Konsens-basierte Fachgesellschaftsempfehlung	Implementierungslücken, ethische/rechtliche Leitplanken für CDSS

Globale Ethik-Mapping (Korukoglu et al., 2025)	Gesundheits-KI generell	Deskriptive Synthese bestehender Leitlinien	Gemeinsame Prinzipien, aber keine einheitliche Durchsetzung
---	----------------------------	--	---

Anmerkung. Die Tabelle vergleicht das einzige verbindliche Rahmenwerk (EU AI Act) mit freiwilligen technischen Standards und sektorspezifischen Governance-Initiativen aus der systematischen Literaturrecherche.

4.3 Sektorspezifische und institutionelle Governance-Initiativen

Über den EU AI Act und ISO 23894 hinaus identifizierte die systematische Untersuchung mehrere sektorspezifische Governance-Initiativen, vorwiegend im Gesundheitswesen konzentriert, die illustrieren, wie regulatorische und ethische Leitlinien in feinerer Granularität entwickelt werden, als breite horizontale Gesetzgebung erreichen kann. Hassan et al. (2026) entwickeln unter Anwendung einer modifizierten Delphi-Konsensmethodik ein Governance-Rahmenwerk speziell für die sichere und ethische Implementierung KI-gestützter klinischer Entscheidungsunterstützungssysteme in der chirurgischen Praxis. Ihr Konsensprozess, unter Einbeziehung relevanter klinischer und technischer Stakeholder, identifiziert eine bedeutende Implementierungslücke zwischen der nachgewiesenen technischen Leistungsfähigkeit chirurgischer KI-Systeme und der Reife der Governance-Strukturen, die für deren sicheren und gerechten Einsatz erforderlich sind – was das breitere, in Abschnitt 3.2.5 identifizierte Muster direkt bestätigt, wo Kale et al. (2024) und Mohammed et al. (2025) ähnlich feststellen, dass Governance- und Berichterstattungsinfrastruktur im Gesundheitswesen hinter der technischen KI-Fähigkeit zurückbleibt.

Korukoglu et al. (2025) liefern eine globale vergleichende Kartierung ethischer Leitlinien für KI im Gesundheitswesen, die Rahmenwerke aus mehreren Rechtsräumen und Berufsverbänden synthetisiert. Ihre Analyse stellt fest, dass, während gemeinsame ethische Prinzipien – einschließlich Mitigation algorithmischen Bias, Patientendatenschutz und Rechenschaftspflicht – über die untersuchten Leitlinien hinweg konsistent wiederkehren, gesundheitswesen-spezifische KI-Ethik-Rahmenwerke sich weiterhin in aktiver Entwicklung befinden, mit begrenztem Konsens über Durchsetzungsmechanismen oder messbare Compliance-Kriterien. Dieser Befund bekräftigt eine für die Gesamtargumentation dieser Arbeit zentrale Unterscheidung: Ethische Leitlinien und verbindliche Regulierung sind nicht gleichwertig, und die Vermehrung ersterer kompensiert nicht die vergleichsweise Knappheit letzterer außerhalb des verbindlichen EU AI Acts.

Hussain et al. (2025), in einer nach PRISMA-ScR-Leitlinien durchgeführten Scoping-Review, die speziell rassistischen und ethnischen Bias in KI-Gesundheitsalgorithmen untersucht, erweitern die Governance-Analyse auf die Frage der Stakeholder-Aufsicht und untersuchen, welche Akteure Verantwortung dafür tragen, algorithmischen Bias zu identifizieren und zu korrigieren, sobald Gesundheits-KI-Systeme im Einsatz sind. Ihre Übersichtsarbeit stellt fest, dass Aufsichtsverantwortung häufig über mehrere Stakeholder verteilt ist – darunter KI-Entwicklerinnen und -Entwickler, Gesundheitseinrichtungen und Regulierungsbehörden –, ohne klare Abgrenzung, welcher Akteur verantwortlich ist, wenn bias-bedingter Schaden eintritt, eine Governance-Lücke mit direkten Implikationen für die praktische Durchsetzbarkeit der im EU AI Act (Abschnitt 4.1) enthaltenen Datenqualitäts- und Diskriminierungsverbots-Bestimmungen, die auch durch die breitere, in Abschnitt 3.4.4 untersuchte Literatur zu institutionellem Bias unterstrichen wird (Javed & Li, 2025).

Chamouni et al. (2025), mit Fokus speziell auf KI-Anwendungen bei der Diagnose und Behandlung von Lungenkrebs, identifizieren ähnlich eine komplexe Landschaft ethischer und rechtlicher Bedenken, die Haftung für KI-unterstützte Diagnosefehler, informierte Einwilligung für KI-unterstützte Versorgung und Daten-Governance umfasst – Bedenken, die, wie jene von Hussain et al. (2025) identifizierten, häufig über den Umfang derzeit kodifizierter Leitlinien und Regulierung hinausgehen, was darauf hindeutet, dass domänenspezifische Governance-Rahmenwerke wahrscheinlich weiterhin neben, statt vollständig subsumiert durch, horizontale Regulierungen wie den EU AI Act entstehen werden.

4.4 Politikempfehlungen und Best Practices

Alvarez et al. (2024) synthetisieren in einer explizit auf die Politikentwicklung ausgerichteten Studie Best Practices für Bias und Fairness in der KI, gerichtet sowohl an Politikgestaltende als auch an Praktikerinnen und Praktiker. Ihre Empfehlungen betonen die Bedeutung kontextspezifischer Fairness-Bewertung – in direkter Übereinstimmung mit den in Abschnitt 3.3 durchgängig diskutierten Befunden zur Metrikpluralität – neben Forderungen nach verpflichtenden algorithmischen Folgenabschätzungen vor dem Einsatz von Hochrisiko-Systemen, einer Empfehlung, die innerhalb der Konformitätsbewertungsverfahren des EU AI Acts (Abschnitt 4.1) bereits weitgehend umgesetzt ist, jedoch in Rechtsräumen ohne vergleichbare verbindliche Gesetzgebung noch keine Standardpraxis darstellt. Die Autorinnen und Autoren betonen zudem die Bedeutung sinnvoller Stakeholder-Beteiligung –

einschließlich der Repräsentation von Gruppen, die am wahrscheinlichsten von algorithmischem Bias betroffen sind – bei der Gestaltung und Evaluation von Hochrisiko-KI-Systemen, eine Empfehlung, die das von Hussain et al. (2025) identifizierte Problem der diffusen Aufsicht adressiert, indem sie einen konkreteren Mechanismus (strukturierte Stakeholder-Beteiligung) zur Zuweisung von Verantwortlichkeit vorschlägt.

4.5 Zusammenfassung und Synthese

Die in diesem Kapitel untersuchte regulatorische und Governance-Literatur offenbart eine Landschaft, die durch eine entstehende, jedoch unvollständige Architektur gekennzeichnet ist: ein einzelnes umfassendes, verbindliches Rechtsrahmenwerk (der EU AI Act), das innerhalb der Europäischen Union operiert, ergänzt durch freiwillige technische Standards (ISO 23894), die Compliance unterstützen, aber nicht eigenständig sicherstellen, sowie ein sich vermehrender Satz sektorspezifischer und institutioneller Governance-Initiativen, die Lücken füllen, die horizontale Regulierung hinterlässt, aber die verbindliche Durchsetzbarkeit und rechtsraumübergreifende Konsistenz formalen Rechts vermissen lassen. Ein durchgängiges Thema dieser Literatur ist die anhaltende Lücke zwischen technischer Fähigkeit zur Bias-Erkennung und -Mitigation (Kapitel 3) und der Governance-Infrastruktur, die erforderlich ist, um sicherzustellen, dass diese Fähigkeit verlässlich und rechenschaftspflichtig in der Praxis angewendet wird (Hassan et al., 2026; Hussain et al., 2025).

Zwei bedeutende Lücken treten aus dieser Synthese mit direkter Relevanz für die Gesamtargumentation der Arbeit hervor. Erstens schafft das Fehlen vorgeschriebener Fairness-Metriken innerhalb verbindlicher Regulierung – wenngleich wohl eine notwendige Anpassung an die in Abschnitt 3.3.2 diskutierten formalen Unmöglichkeitsergebnisse – erhebliche Compliance-Unsicherheit, die aktuelle freiwillige Standards nicht vollständig auflösen. Zweitens bleiben Aufsicht und Rechenschaftspflicht für algorithmischen Bias über mehrere Stakeholder verteilt, ohne klare Zuweisung von Verantwortlichkeit (Hussain et al., 2025), eine strukturelle Governance-Lücke, die die praktische Wirksamkeit selbst gut konzipierter regulatorischer Anforderungen einschränken kann. Diese Befunde fließen unmittelbar in die in Kapitel 5 präsentierte Diskussion politischer Implikationen und Empfehlungen ein.

5. Diskussion und Fazit

Dieses abschließende Kapitel synthetisiert die in den Kapiteln 1 bis 4 präsentierten Befunde, beantwortet explizit die vier Forschungsfragen, die diese systematische Literaturanalyse leiteten, diskutiert die Limitationen des Übersichtsprozesses und identifiziert Richtungen für künftige Forschung. Das Kapitel schließt mit einer kurzen Reflexion über die breiteren Implikationen dieser Untersuchung für die verantwortungsvolle Entwicklung und den Einsatz von KI-Systemen, insbesondere im Gesundheitswesen.

5.1 Synthese der Ergebnisse über die Kapitel hinweg

Die für diese Arbeit durchgeführte systematische Übersichtsarbeit, die 66 über vier Datenbanken (PubMed/PMC, Springer Link, JSTOR und arXiv) identifizierte Studien nach dem PRISMA-2020-Protokoll umfasst, offenbart eine Forschungslandschaft, die durch erhebliche technische Ausgereiftheit neben anhaltender konzeptioneller Fragmentierung und Übertragungslücken gekennzeichnet ist. Drei übergreifende Muster treten hervor, wenn die Befunde der Kapitel 3 und 4 gemeinsam betrachtet werden, statt innerhalb ihrer jeweiligen thematischen Silos.

Erstens hat die in Kapitel 1.2 etablierte begriffliche Pluralität von Fairness-Definitionen nicht bloß eine philosophische Feinheit, sondern unmittelbare, empirisch nachgewiesene Konsequenzen in den nachfolgenden Kapiteln. Die in Abschnitt 3.3.2 diskutierten formalen Unmöglichkeitsergebnisse (Kleinberg et al., 2017; Chouldechova, 2017) sind keine abstrakten mathematischen Kuriositäten: Sie werden direkt durch die in Abschnitt 3.3.4 untersuchten empirischen klinischen Studien (Ganta et al., 2024; Yuan et al., 2023) bestätigt, die feststellen, dass Modelle, die ein Fairness-Kriterium in realen diagnostischen Kontexten erfüllen, andere nicht erfüllen. Dies deutet darauf hin, dass die theoretische und die empirische Literatur, obwohl weitgehend von unterschiedlichen Forschungsgemeinschaften mit unterschiedlichen Methoden entwickelt, auf eine konsistente Schlussfolgerung konvergieren: Fairness-Zielkonflikte sind ein strukturelles Merkmal algorithmischer Entscheidungsfindung unter realen Bedingungen ungleicher Basisraten, kein Artefakt mangelhaften Modelldesigns, das weitere technische Verfeinerung letztlich beseitigen wird.

Zweitens ist ein wiederkehrendes Thema über die Abschnitte 3.2, 3.4 und 4.3 hinweg, dass der primäre Engpass, der Fairness-Verbesserungen einschränkt, zunehmend organisatorisch und

institutionell statt rein technisch ist. In Abschnitt 3.4 untersuchte Mitigation-Techniken – insbesondere die neueren In-Processing- und Datensatz-Bias-Erkennungsmethoden (Kim & Lee, 2025; Dominguez-Catena et al., 2024) – zeigen, dass sich die technische Fähigkeit, Bias zu erkennen und teilweise zu korrigieren, erheblich weiterentwickelt hat. Doch Abschnitt 3.2.5 und Kapitel 4 zeigen gemeinsam, dass Gesundheitseinrichtungen häufig die Berichterstattungsinfrastruktur (Kale et al., 2024), die Personalschulung (Mohammed et al., 2025) und die klar abgegrenzte Aufsichtsverantwortung (Hussain et al., 2025) vermissen lassen, die zur Nutzung dieser technischen Fähigkeit erforderlich sind. Dieses Muster legt nahe, dass künftiger Fortschritt bei der Fairness der Gesundheits-KI wahrscheinlich weniger durch algorithmische Innovation als durch institutionellen und regulatorischen Kapazitätsaufbau bestimmt wird – eine Schlussfolgerung mit direkten Implikationen dafür, wie Forschungsaufwand und politische Aufmerksamkeit künftig zugewiesen werden sollten.

Drittens weisen die Literatur zu großen Sprachmodellen (Abschnitt 3.1) und die Literatur zur Gesundheits-KI (Abschnitt 3.2) – obwohl sie sich, wie in der Problemstellung (Abschnitt 1.1.1) angemerkt, weitgehend parallel entwickelt haben – ein strukturell ähnliches Bias-Muster auf: Bias ist technisch auf spezifische, identifizierbare Mechanismen zurückführbar (Zusammensetzung der Trainingsdaten, Label-Bias, Zusammensetzung der Evaluationsbenchmarks), statt eine unerklärliche emergente Eigenschaft von Modellgröße oder -komplexität zu sein. Diese strukturelle Ähnlichkeit, die deutlich wird, wenn die Abschnitte 3.1.3 und 3.2.2 gemeinsam gelesen werden, legt nahe, dass Mitigation-Erkenntnisse, die in einer Domäne entwickelt wurden, möglicherweise stärker auf die andere übertragbar sind, als die gegenwärtige Aufspaltung der Forschungsliteratur impliziert – eine methodische Gelegenheit, die die Synthese dieser Arbeit explizit macht, die jedoch von wenigen der einzeln untersuchten Primärstudien genutzt zu werden scheint.

5.2 Beantwortung der Forschungsfragen

Tabelle 5.1 fasst den zentralen Befund dieser Untersuchung in Bezug auf jede der vier in Abschnitt 1.1.2 etablierten Forschungsfragen zusammen, mit Querverweisen auf die relevanten Abschnitte der Kapitel 3 und 4, in denen die unterstützende Evidenz vollständig präsentiert wird.

Tabelle 5.1: Zusammenfassung der Antworten auf die Forschungsfragen

RQ	Zentrales Ergebnis
RQ1	Bias in AI/ML-Systemen lässt sich entlang sechs wiederkehrender Mechanismen kategorisieren (historical, representation, measurement, aggregation, evaluation, deployment bias), die sich empirisch über LLMs und klinische ML-Systeme hinweg replizieren (Abschnitt 3.1.1, 3.2.2).
RQ2	Es existiert kein einzelnes, universell korrektes Fairness-Maß. Formale Unmöglichkeitstheoreme (Kleinberg et al., 2017; Chouldechova, 2017) zeigen, dass mehrere plausible Fairness-Kriterien bei unterschiedlichen Basisraten nicht gleichzeitig erfüllbar sind — eine mathematische Notwendigkeit, nicht ein Mangel an Forschung (Abschnitt 3.3.2).
RQ3	Mitigation-Strategien lassen sich in Pre-/In-/Post-Processing sowie organisatorisch-soziale Maßnahmen einteilen; keine einzelne Strategie ist universell überlegen, und Wirksamkeit muss kontextspezifisch validiert werden (Abschnitt 3.4.5).
RQ4	Der EU AI Act bietet als einziges verbindliches Rahmenwerk Datenqualitäts- und Risikomanagementpflichten, spezifiziert jedoch keine konkrete Fairness-Metrik. Diese regulatorische Flexibilität ist angesichts der Unmöglichkeitstheoreme nachvollziehbar, erzeugt aber Compliance-Unsicherheit und diffuse Verantwortlichkeitsstrukturen (Abschnitt 4.5).

Zusammengenommen stützen die Antworten auf diese vier Forschungsfragen eine zentrale Schlussfolgerung auf Ebene der gesamten Arbeit: Bias und Fairness in der KI können nicht allein durch technische Intervention, allein durch Metrikstandardisierung oder allein durch Regulierung angemessen adressiert werden. Jeder dieser drei Hebel – untersucht jeweils in den Kapiteln 3.1–3.2 (technisch/empirisch), 3.3 (metrisch/theoretisch) und 4 (regulatorisch) – adressiert eine notwendige, jedoch für sich allein unzureichende Komponente der gesamten Fairness-Herausforderung. Diese Schlussfolgerung ist selbst eine direkte Konsequenz der in Kapitel 1.2 etablierten konzeptionellen Grundlage: Da Fairness ein sozio-technisches und kein rein technisches Konstrukt ist, erfordert dauerhafter Fortschritt koordinierten Fortschritt entlang aller drei Hebel simultan, statt sequenzieller Optimierung eines einzelnen.

5.3 Limitationen

Mehrere Limitationen dieser systematischen Literaturanalyse sollten anerkannt werden, die sowohl die Übersichtsmethodik selbst als auch die zugrunde liegende synthetisierte Primärliteratur umfassen.

5.3.1 Methodische Limitationen der Übersichtsarbeit

Wie in Abschnitt 2.8 ausführlich dargestellt, unterliegt diese Untersuchung mehreren methodischen Einschränkungen. Die Beschränkung auf vier Datenbanken kann, obwohl disziplinär breit gefasst, relevante Studien ausgeschlossen haben, die in spezialisierten, dort nicht indexierten Publikationsorganen erschienen sind. Die Beschränkung auf englisch- und deutschsprachige Publikationen führt zu einem Sprach-Bias, der KI-Fairness-Forschung aus nicht-anglophonen Kontexten wahrscheinlich unterrepräsentiert, insbesondere Forschung zu Fairness-Bedenken, die spezifisch für nicht-westliche Bevölkerungsgruppen und Gesundheitssysteme sind – eine Ironie angesichts des inhaltlichen Schwerpunkts dieser Untersuchung auf Representation Bias und Unterrepräsentation als wiederkehrendes Thema (Abschnitt 3.1.1, 3.2.2). Eine für die Durchführung dieser Untersuchung spezifische technische Limitation, dokumentiert in Abschnitt 2.5.5, ist, dass 677 von arXiv abgerufene Datensätze aufgrund einer exportbedingten Kürzung des Abstract-Texts nicht verlässlich gescreent werden konnten; diese Datensätze wurden konservativ ausgeschlossen statt auf Basis des Titels allein eingeschlossen, was zum Ausschluss relevanter aktueller Preprints geführt haben könnte, insbesondere angesichts der Rolle von arXiv als primärer Publikationsort für die aktuellste LLM-Fairness-Forschung (Abschnitt 3.1).

Eine weitere Limitation betrifft die in dieser Untersuchung durchgeführte Eignungsbewertung (Abschnitt 2.5.3): Aufgrund der praktischen Undurchführbarkeit einer Volltextprüfung von über 2.000 geeigneten Datensätzen wurde die finale Auswahl von 66 Studien primär anhand strukturierter Metadaten (Publikationsorgan, Erscheinungsjahr, Studientyp und Abstract-Substanz) getroffen, statt durch vollständige Volltextlektüre jedes geeigneten Kandidaten vor dem finalen Einschluss. Während dieser Ansatz mit den Zeitbeschränkungen einer Masterarbeit konsistent ist und das Screening mit expliziten, dokumentierten Kriterien zur Maximierung der Transparenz durchgeführt wurde, bleibt er hinter dem als Best Practice für systematische Übersichtsarbeiten empfohlenen Volltext-Screening zurück (Page et al., 2021) und bedeutet, dass einige thematisch relevante Studien ausgeschlossen worden sein könnten, oder dass sich einige eingeschlossene Studien bei genauerer Volltextprüfung als weniger zentral für die Forschungsfragen erweisen könnten, als ihre Metadaten nahelegten.

5.3.2 Limitationen der zugrunde liegenden Primärliteratur

Über die methodischen Einschränkungen der Übersichtsarbeit selbst hinaus sind mehrere Limitationen Eigenschaften der Primärliteratur selbst, die als wiederkehrende Forschungslücken im Verlauf der Synthese in Kapitel 3 identifiziert wurden. Die Literatur zur Fairness der Gesundheits-KI weist eine ausgeprägte Unterentwicklung von Fairness-Methoden für unüberwachtes Lernen im Vergleich zu überwachten Settings auf, trotz des wachsenden klinischen Einsatzes unüberwachter Methoden (Abschnitt 3.2.3; Adouani et al., 2026). Subgruppenstratifizierte Ergebnisberichterstattung fehlt häufig selbst in ansonsten methodisch rigorosen Primärstudien, wie im Kontext KI-gestützter Tools für die psychische Gesundheit dokumentiert (Abschnitt 3.2.4; Humayun et al., 2025), was die Fähigkeit sekundärer Analysen wie dieser Untersuchung einschränkt, feste Schlussfolgerungen über die Gerechtigkeit berichteter Vorteile zu ziehen. Bestimmte Bias-Dimensionen – insbesondere altersbezogener Bias – erhalten unverhältnismäßig weniger mitigationsfokussierte Forschungsaufmerksamkeit als Gender- oder Rassen-Bias, unabhängig von ihrer realen Prävalenz (Abschnitt 3.4.4; Chu et al., 2024). Dies sind Limitationen des Feldes als Ganzes, nicht dieser Untersuchung speziell, doch sie beschränken die Stärke der Schlussfolgerungen, die jede noch so methodisch rigorose Synthese aus der derzeit verfügbaren Evidenzbasis ziehen kann.

5.4 Richtungen für künftige Forschung

Aufbauend auf den über Kapitel 3 hinweg identifizierten Forschungslücken und den oben diskutierten Limitationen erscheinen mehrere Richtungen für künftige Forschung besonders vielversprechend. Erstens sollte künftige empirische Arbeit Fairness-Methoden für unüberwachte und selbstüberwachte Lernparadigmen priorisieren, angesichts ihrer wachsenden Rolle in Gesundheitsanwendungen wie Patient:innen-Phänotypisierung und Kohortenentdeckung (Abschnitt 3.2.3) und der vergleichswisen Knappheit etablierter Fairness-Metriken, die auf Settings ohne Ground-Truth-Labels anwendbar sind. Zweitens sollten angesichts der von Jung et al. (2025) nachgewiesenen und in den Abschnitten 1.2.3 und 3.1.4 diskutierten Fragilität von Fairness unter Prompt-Variation künftige LLM-Fairness-Benchmarks Robustheitstests systematisch als Standardkomponente der Evaluation integrieren, statt Robustheit als separate oder optionale Erwägung zu behandeln. Drittens sollte künftige Forschung die LLM-Fairness-Literatur und die Literatur zur Fairness im klinischen maschinellen Lernen expliziter verbinden, aufbauend auf den in Abschnitt 5.1 identifizierten strukturellen Ähnlichkeiten, um zu prüfen, ob in einer Domäne entwickelte Mitigation-

Techniken wirksam auf die andere übertragbar sind. Viertens würde künftige Forschung angesichts der in den Abschnitten 3.2.5 und 4.3 identifizierten institutionellen und Governance-Lücken von stärkerer interdisziplinärer Zusammenarbeit zwischen Informatik und Organisations-/Gesundheitspolitikforschung profitieren, dem Beispiel integrativer Übersichtsarbeiten wie Hickman et al. (2025) folgend, um Governance-Mechanismen zu entwickeln, die in der Lage sind, technische Fairness-Fähigkeit in verlässlich durchgesetzte klinische und organisatorische Praxis zu übersetzen.

5.5 Schlussbemerkungen

Diese Arbeit hatte zum Ziel, die sich rasch entwickelnde und begrifflich fragmentierte Literatur zu Bias und Fairness in der Künstlichen Intelligenz zu synthetisieren, mit besonderem Augenmerk auf große Sprachmodelle und deren Anwendung im Gesundheitswesen. Die hier durchgeführte systematische Untersuchung zeigt, dass diese Fragmentierung nicht bloß eine Funktion der Jugend des Feldes ist, sondern in wichtigen Aspekten echte und dauerhafte Komplexität widerspiegelt: Fairness ist als normatives Konstrukt unauflöslich plural (Hellman, 2020), formal unmöglich, entlang aller gewünschten Dimensionen simultan vollständig zu erreichen (Kleinberg et al., 2017), und nur teilweise durch verbindliche Regulierung adressierbar, die einsetzenden Organisationen notwendigerweise erheblichen interpretativen Spielraum lassen muss (Abschnitt 4.1).

Diese Komplexität sollte jedoch nicht als Argument gegen fortgesetzte Bemühungen missverstanden werden. Die in dieser Arbeit untersuchte Evidenz zeigt konkreten, messbaren Fortschritt: Bias-Erkennungsmethoden sind zunehmend ausgereift und in der Lage, zuvor unsichtbare Formen von Bias auf Datensatz- und Repräsentationsebene zu identifizieren (Abschnitt 3.1, 3.4.1); Mitigation-Techniken reduzieren, wenngleich unvollkommen, nachweislich – auch wenn sie nicht eliminieren – gemessene Ungleichheiten in eingesetzten Systemen (Abschnitt 3.4); und regulatorische Rahmenwerke haben, obwohl unvollständig, begonnen, verbindliche Rechenschaftspflichten dort durchzusetzen, wo zuvor keine bestanden (Kapitel 4). Der zentrale Beitrag dieser Arbeit besteht darin, explizit zu machen, dass diese drei Fortschrittslinien – technisch, methodisch und regulatorisch – in bewusster Koordination statt isoliert verfolgt werden müssen, da die zentrale Lehre dieser Untersuchung lautet, dass keine von ihnen allein, wie erfolgreich auch für sich genommen, ausreicht, um sicherzustellen, dass KI-Systeme die von ihnen betroffenen Menschen fair behandeln.

Hinweis zur Verwendung generativer KI

Im Rahmen der Erstellung dieser Masterarbeit wurde das KI-basierte Sprachmodell Claude (Anbieter: Anthropic, Modell Sonnet 4.6) im Zeitraum März bis Juli 2026 verwendet. Der Einsatz beschränkte sich auf die sprachliche Überarbeitung und stilistische Verbesserung einzelner, vom Autor selbst verfasster Textpassagen.

Die inhaltliche Konzeption, die Erhebung und Auswertung der Daten sowie alle Analysen, Interpretationen und Schlussfolgerungen wurden eigenständig durch den Autor erbracht. Alle durch das Tool überarbeiteten Formulierungen wurden vom Autor geprüft.

Die Verantwortung für den gesamten Inhalt dieser Arbeit liegt vollständig beim Autor.

Literaturverzeichnis

- Adouani, M., Annane, D., & Chelly Dagdia, Z. (2026). Rethinking fairness in unsupervised healthcare AI: A methodological scoping review. *Journal of Biomedical Informatics*. <https://doi.org/10.1016/j.jbi.2026.105040>.
- Alvarez, J. M., Carro, R., & Dignum, V. (2024). Policy advice and best practices on bias and fairness in AI. *Ethics and Information Technology*, 26(3), 19. <https://doi.org/10.1007/s10676-024-09746-w>
- An, J., Huang, D., Lin, C., & Tai, M. (2025). Measuring gender and racial biases in large language models: Intersectional evidence from automated resume evaluation. *PNAS Nexus*, 4(3), pgaf089. <https://doi.org/10.1093/pnasnexus/pgaf089>
- Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press. <https://fairmlbook.org>
- Basu, S., Patel, S., Sheth, P., et al. (2026). Mechanistic interpretability of reinforcement learning in Medicaid care coordination. *BMJ Health & Care Informatics*. <https://doi.org/10.1136/bmjhci-2025-101935>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of FAccT '21, ACM*, pp. 610–623. <https://doi.org/10.1145/3442188.3445922>
- Birdi, S., Rabet, R., Durant, S., et al. (2024). Bias in machine learning applications to address non-communicable diseases at a population-level: A scoping review. *BMC Public Health*. <https://doi.org/10.1186/s12889-024-21081-9>
- Carbunaru, S., Neshatvar, Y., Do, H., et al. (2024). Survival after radical cystectomy for bladder cancer: Development of a fair machine learning model. *JMIR Medical Informatics*. <https://doi.org/10.2196/63289>
- Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7), Article 166. <https://doi.org/10.1145/3616865>

- Chamouni, G., Lococo, F., Sassorossi, C., et al. (2025). Ethical and legal concerns in artificial intelligence applications for the diagnosis and treatment of lung cancer: A scoping review. *Frontiers in Public Health*. <https://doi.org/10.3389/fpubh.2025.1663298>
- Chen, F., Wang, L., Hong, J., et al. (2024). Unmasking bias in artificial intelligence: A systematic review of bias detection and mitigation strategies in electronic health record-based models. *Journal of the American Medical Informatics Association*, 31(5). <https://doi.org/10.1093/jamia/ocae060>
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153–163. <https://doi.org/10.1089/big.2016.0047>
- Chu, Z., Wang, Z., & Zhang, W. (2024). Fairness in large language models: A taxonomic survey. *ACM SIGKDD Explorations Newsletter*, 26(1), 34–48. <https://doi.org/10.1145/3682112.3682117>
- Colacci, M., Huang, Y. Q., Postill, G., et al. (2025). Sociodemographic bias in clinical machine learning models: A scoping review of algorithmic bias instances and mechanisms. *Journal of Clinical Epidemiology*. <https://doi.org/10.1016/j.jclinepi.2024.111606>
- Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., & Goel, S. (2023). The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312), 1–117. <https://doi.org/10.5555/3648699.3649011>
- De, A., & Lohani, A. (2026). Regulatory adoption of AI, ML, computational modeling & simulation in in-silico clinical trials for medical devices: A systematic review. *Therapeutic Innovation & Regulatory Science*. <https://doi.org/10.1007/s43441-025-00871-2>
- Dogan, M. V., et al. (2025). Evaluating and addressing demographic disparities in medical large language models: A systematic review. *International Journal for Equity in Health*. PMC11866893. <https://doi.org/10.1186/s12939-025-02419-0>
- Dominguez-Catena, I., Paternain, D., & Galar, M. (2024). Metrics for dataset demographic bias: A case study on facial expression recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. <https://doi.org/10.1109/TPAMI.2024.3352181>

- Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), eaao5580. <https://doi.org/10.1126/sciadv.aao5580>
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of ITCS 2012*, ACM. <https://doi.org/10.1145/2090236.2090255>
- Europäische Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act). *Official Journal of the European Union*. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>
- Ferrara, E. (2024). Fairness and bias in AI: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1), 3. <https://doi.org/10.3390/sci6010003>
- Gallegos, I. O., Rossi, R. A., Barrow, J., et al. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3), 1097–1179. https://doi.org/10.1162/coli_a_00524
- Ganta, T., Kia, A., Parchure, P., et al. (2024). Fairness in predicting cancer mortality across racial subgroups. *JAMA Network Open*, 7(8). <https://doi.org/10.1001/jamanetworkopen.2024.27791>
- Gray, M., Milanova, M., & Wu, L. (2024). Enhancing bias assessment for complex term groups in language embedding models: Quantitative comparison of methods. *JMIR Medical Informatics*. <https://doi.org/10.2196/60272>
- Gulamali, F., Sawant, A. S., Liharska, L., et al. (2025). Detecting, characterizing, and mitigating implicit and explicit racial biases in health care datasets with subgroup learnability. *Journal of Medical Internet Research*. <https://doi.org/10.2196/71757>
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems (NeurIPS 2016)*. arXiv:1610.02413. <https://doi.org/10.48550/arXiv.1610.02413>
- Hasanzadeh, F., Liu, M., & Wang, J. (2025). Bias recognition and mitigation strategies in artificial intelligence for healthcare. *NPJ Digital Medicine*, 8(1503). <https://doi.org/10.1038/s41746-025-01503-7>

- Hassan, A. M., Coert, J. H., Clemens, M. W., et al. (2026). Governance framework for safe and ethical implementation of artificial intelligence in surgery: A modified Delphi consensus. *Journal of the American College of Surgeons*. <https://doi.org/10.1097/XCS.0000000000001834>
- Hellman, D. (2020). Measuring algorithmic fairness. *Philosophy & Public Affairs*, 48(4), 431–461. <https://doi.org/10.1111/papa.12176>
- Hickman, L., Huynh, C., Gass, J., et al. (2025). Whither bias goes, I will go: An integrative, systematic review of algorithmic bias mitigation. *Journal of Applied Psychology*. <https://doi.org/10.1037/apl0001255>
- Huang, Y., Guo, J., Chen, W. H., et al. (2024). A scoping review of fair machine learning techniques when using real-world data. *Journal of Biomedical Informatics*. <https://doi.org/10.1016/j.jbi.2024.104622>
- Humayun, A., Madawana, A. M., Hassan, A., et al. (2025). Artificial intelligence as a predictive tool for mental health status: Insights from a systematic review and meta-analysis. *PLOS One*. <https://doi.org/10.1371/journal.pone.0332207>
- Hussain, S. A., Bresnahan, M., & Zhuang, J. (2025). The bias algorithm: How AI in healthcare exacerbates ethnic and racial disparities — a scoping review. *Ethnicity & Health*. <https://doi.org/10.1080/13557858.2024.2422848>
- ISO. (2023). ISO/IEC 23894:2023 – Artificial intelligence — Guidance on risk management. International Organization for Standardization. <https://www.iso.org/standard/77304.html>
- Jabbour, S., Fouhey, D., Shepard, S., et al. (2023). Measuring the impact of AI in the diagnosis of hospitalized patients: A randomized clinical vignette survey study. *JAMA*, 330(23). <https://doi.org/10.1001/jama.2023.22295>
- Javed, K., & Li, J. (2025). Bias in adjudication: Investigating the impact of artificial intelligence, media, financial and legal institutions in pursuit of social justice. *PLOS One*. <https://doi.org/10.1371/journal.pone.0315270>

- Jung, D., Lee, J., Kim, Y., & Cho, K. (2025). FLEX: A benchmark for evaluating robustness of fairness in large language models. arXiv:2508.08965. <https://doi.org/10.48550/arXiv.2508.08965>
- Kale, A. U., Dattani, R., Tabansi, A., et al. (2024). AI as a medical device adverse event reporting in regulatory databases: Protocol for a systematic review. *JMIR Research Protocols*. <https://doi.org/10.2196/48156>
- Kheiri, F., Rahnamayan, S., & Makrehchi, M. (2026). Mitigating data center bias in cancer classification. *Artificial Intelligence in Medicine*. <https://doi.org/10.1016/j.artmed.2026.103351>.
- Kim, S., & Lee, S. W. (2025). AmbiBias Contrast: Enhancing debiasing networks via disentangled space from ambiguity-bias clusters. *Neural Networks*. <https://doi.org/10.1016/j.neunet.2024.106857>.
- Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering (EBSE Technical Report, Version 2.3). Keele University and University of Durham.
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *Proceedings of ITCS 2017*. LIPIcs. <https://doi.org/10.4230/LIPIcs.ITCS.2017.43>
- Korukoglu, F. B., Rishan, R. M., Elhamer, Z., et al. (2025). Mapping ethical guidelines for AI in healthcare: A global perspective. *Studies in Health Technology and Informatics*. <https://doi.org/10.3233/SHTI250624>
- Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems (NeurIPS 2017)*. arXiv:1703.06856. <https://doi.org/10.48550/arXiv.1703.06856>
- Li, X., Yan, X., & Lai, H. (2025). The ethical challenges in the integration of artificial intelligence and large language models in medical education: A scoping review. *PLOS One*. <https://doi.org/10.1371/journal.pone.0333411>

- Li, Z., Yang, W., Shu, X., et al. (2026). BiasField: Interactive bias probing of machine learning datasets. *IEEE Transactions on Visualization and Computer Graphics*. <https://doi.org/10.1109/TVCG.2026.3694439>
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), Article 115. <https://doi.org/10.1145/3457607>
- Mohammed, S. A. A. Q., Osman, Y. M. M., Ibrahim, A. M., et al. (2025). Ethical and regulatory considerations in the use of AI and machine learning in nursing: A systematic review. *International Nursing Review*. <https://doi.org/10.1111/inr.70010>
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys*, 55(3), Article 51. <https://doi.org/10.1145/3494672>
- Rahman, M. I. (2026). Algorithmic bias and transparency in artificial intelligence tools for nursing education: A scoping review. *Nurse Education in Practice*. <https://doi.org/10.1016/j.nepr.2026.104756>
- Rickman, S. (2025). Evaluating gender bias in large language models in long-term care. *BMC Medical Informatics and Decision Making*, 25, 274. <https://doi.org/10.1186/s12911-025-03118-0>
- Schnetler, R., van der Vegt, A., Kalke, V. R., et al. (2025). False hope of a single generalisable AI sepsis prediction model: Bias and proposed mitigation strategies. *BMJ Quality & Safety*. <https://doi.org/10.1136/bmjqs-2024-018328>
- Schuch, H. S., Furtado, M., Silva, G. F. D. S., et al. (2023). Fairness of machine learning algorithms for predicting foregone preventive dental care for adults. *JAMA Network Open*, 6(7). <https://doi.org/10.1001/jamanetworkopen.2023.22523>

- Ukanwa, K. (2024). Algorithmic bias: Social science research integration through the 3-D Dependable AI Framework. *Current Opinion in Psychology*. <https://doi.org/10.1016/j.copsyc.2024.101836>.
- Wang, S., Zhao, T., Zhang, W., & Huang, T. (2024). CEB: Compositional evaluation benchmark for fairness in large language models. *arXiv:2407.02408*. <https://doi.org/10.48550/arXiv.2407.02408>
- Weidinger, L., Mellor, J., Rauh, M., et al. (2021). Ethical and social risks of harm from language models. *arXiv:2112.04359*. <https://doi.org/10.48550/arXiv.2112.04359>
- Woo, B. F. Y., Cato, K., Cho, H., et al. (2026). The use of large language models in clinical documentation: A scoping review. *International Journal of Nursing Studies*. <https://doi.org/10.1016/j.ijnurstu.2025.105322>
- Xiang, A., Andrews, J. T. A., Bourke, R. L., et al. (2025). Fair human-centric image dataset for ethical AI benchmarking. *Nature*. <https://doi.org/10.1038/s41586-025-09716-2>
- Yang, Y., Lin, M., & Zhao, H., et al. (2024). A survey of recent methods for addressing AI fairness and bias in biomedicine. *Journal of Biomedical Informatics*. <https://doi.org/10.1016/j.jbi.2024.104646>
- Yuan, C., Linn, K. A., & Hubbard, R. A. (2023). Algorithmic fairness of machine learning models for Alzheimer disease progression. *JAMA Network Open*, 6(11). <https://doi.org/10.1001/jamanetworkopen.2023.44737>