INTERNATIONALE HOCHSCHULE

www.iu.de

IU DISCUSSION PAPERS

Human Resources

Diskriminierungssensible Nutzung von Künstlicher Intelligenz in der Personalentwicklung

SARAH FRIESE
MAJA STÖRMER

IU Internationale Hochschule

Main Campus: Erfurt

Juri-Gagarin-Ring 152

99084 Erfurt

Telefon: +49 421.166985.23

Fax: +49 2224.9605.115

Kontakt/Contact: kerstin.janson@iu.org

Autorenkontakt/Contact to the authors:

Prof. Dr. Maja Störmer

ORCID-ID: 0000-0001-8647-5151

IU Internationale Hochschule - Campus Berlin

E-Mail: maja.stoermer@iu.org

IU Discussion Papers, Reihe: Human Resources, Vol. 6, No. 5 (AUG 2026)

ISSN: 2750-0721

DOI: https://doi.org/10.56250/4144

Website: https://repository.iu.org

DISKRIMINIERUNGSSENSIBLE NUTZUNG VON KÜNSTLICHER INTELLIGENZ IN DER PERSONALENTWICKLUNG

Sarah Friese

Maja Störmer

ABSTRACT

This paper examines the use of artificial intelligence (AI) in human resource development, with a particular focus on discrimination-sensitive design. Drawing on legal frameworks, particularly the EU AI Act, as well as sociological and technical perspectives, the study analyzes both the potential and the risks of AI deployment. In human resource development, AI offers significant added value through personalized learning processes, competence analysis, and the evaluation of development measures. At the same time, there is a risk that socially embedded forms of discrimination may be reproduced through biased training data. The analysis shows that, to date, only a few technically effective fairness approaches exist, particularly for generative AI. Against this background, responsibility lies not only with developers but also with users of AI systems. The paper derives practice-oriented recommendations for the legally compliant, fair, and human-centered use of AI in human resource development and emphasizes the need for a holistic sociotechnical approach.

KEYWORDS

Artificial intelligence, human resource development, fairness, discrimination, EU AI Act

AUTORINNEN

Sarah Friese *hat ihren Masterabschluss im Studiengang Personalentwicklung an der IU Internationale Hochschule absolviert und interessiert sich für Themen des digitalen Lernens und Diversity. Ganz besonders auch für deren Umsetzung in KMUs. Als Personalreferentin übernimmt sie die Betreuung von Auszubildenen und Mitarbeitenden mit der Perspektive auf die Optimierung von Personalprozessen durch den Einsatz Künstlicher Intelligenz.*

Maja Störmer *ist Professorin für Personalentwicklung an der IU Internationale Hochschule und Leiterin des Fachgebiets Human Resources. Sie forscht, lehrt und publiziert zu den Themen Organisations- und Personalentwicklung, technologische Transformationsprozesse, Organisationskommunikation sowie Diversity Management. Auf internationalen Konferenzen hielt sie Fachvorträge, etwa für die Universität Bergen oder die Universität Stellenbosch. Als selbstständige Trainerin leitet und leitete Sie regelmäßig Workshops, u.a. zu dem Thema Interkulturelle Kompetenz und Umgang mit Unsicherheit, wie für den Interculture e.V., den DAAD, die Airbus Operation GmbH oder das African Center for Career Enhancement and Skills Support.*

1. Einleitung: Künstliche Intelligenz in der Personalentwicklung

1.1 THEMATISCHE RELEVANZ

Der Einsatz von Künstlicher Intelligenz (KI) in der Personalarbeit ist längst obligatorisch und kann entlang des gesamten *Employee Lifecycle*, die Phasen, die eine Person in ihrer Beziehung zu einem Unternehmen durchläuft, vielfältig eingesetzt werden (Fink, 2021, S. 57). Auch für die Personalentwicklung ist der Einsatz lohnend, befindet sich aber in den meisten Unternehmen noch am Anfang (Renz, 2025, S. 243).

Die Ziele des Einsatzes von KI in der Personalentwicklung umfassen zum einen die Bedürfnisse angehender und bestehender Mitarbeitenden ressourcensparend zu bedienen und zum anderen in der Organisation eine lernende Kultur zu entwickeln, die das lebenslange Lernen ermöglicht (Fink, 2021, S. 57). Künstliche Intelligenz kann gewinnbringend eingesetzt werden, um die Personalentwicklung „personalisiert, integriert und selbstorganisiert“ (Fink, 2021, S. 104) zu gestalten. Mit sogenannten *Learning Analytics* können die Daten der lernenden Person analysiert und ein adaptives, also an die individuellen Bedürfnisse der Person angepasstes Lernen realisiert werden (Hasenbein, 2023, S. 100). Auf diese Weise können die Bedarfe zunächst identifiziert und anschließend Lernziele, Lerninhalte und Lernformate individuell empfohlen werden (ebd.).

Der Einsatz von KI in Unternehmen führt häufig zu einem Wandel der von Mitarbeitenden benötigten Kompetenzen und erzeugt dadurch neue Entwicklungsbedarfe (Hägerbäumer & Drewes, 2025, S. 64). Daraus ergibt sich die Notwendigkeit, Beschäftigte gezielt für den Umgang mit Künstlicher Intelligenz zu qualifizieren, denn „eine Qualifizierung für Künstliche Intelligenz [wird] eine absolute Notwendigkeit sein“(Hasenbein, 2023, S. 100). Künstliche Intelligenz wird damit selbst zu einem Treiber der Personalentwicklung, da ihr Einsatz neue Anforderungen schafft, auf die Personalentwicklungsmaßnahmen reagieren müssen. Infolgedessen verändern sich nicht nur die Inhalte, sondern auch die Aufgaben und das Vorgehen der Personalentwicklung durch den Einsatz von KI (ebd.).

Folglich gilt es auch die Zusammenarbeit von Menschen und KI kritisch zu betrachten. Mit der Gestaltung einer menschenzentrierten KI wird das Ziel verfolgt, dass Ausgaben einer KI nachvollziehbar und erklärbar sein sollen (Hasenbein, 2023, S. 163). Künstliche Intelligenz soll nichtdiskriminierend und fair sein. Was zunächst trivial klingen mag, stellt jedoch eine sehr komplexe Anforderung dar. Diese beginnt bereits bei den Begrifflichkeiten. Eine gleiche Behandlung aller Menschen erzeugt im individuellen Einzelfall nicht notwendigerweise ein gerechtes Ergebnis für die Einzelperson. Der Anspruch an gerechte und unvoreingenommene Ausgaben durch algorithmische Entscheidungssysteme stellt damit eine grundlegende Herausforderung dar (Hasenbein, 2023, S. 165).

Eine Nichtdiskriminierung anhand bestimmter Merkmale ist gesetzlich geregelt und Kriterien für einen ethischen Einsatz wurden benannt. Die Grundlage bilden das Grundgesetz (GG), das Allgemeine Gleichbehandlungsgesetz (AGG) und die Charta der Grundrechte der Europäischen Union (EU-GRCh). Mit dem EU AI Act ist der rechtliche Rahmen für den Schutz vor Diskriminierung im Einsatz von KI-Systemen geschaffen worden. Ergänzend sind in der Assessment *List for Trustworthy Artificial Intelligence* (ALTAI) Anforderungen an vertrauenswürdige und nichtdiskriminierende KI-Systeme formuliert.

Während eine Nichtdiskriminierung eindeutig geregelt ist, ist einem Verständnis von Fairness eine gewisse Subjektivität inhärent. So bestehen beispielsweise Zweifel an Fairness und Gerechtigkeit von KI-Tools, die zur Eignungsdiagnostik bereits eingesetzt werden (Fink, 2021, S. 101).

Neben den Schwierigkeiten Fairness sicherzustellen ist es wichtig anzuerkennen, dass KI-Systeme nicht grundsätzlich diskriminierend sind. Dahinter steht ein lernendes Modell, dass Verzerrungen reproduzieren und festigen kann. Ein Teil der Begründung, warum KI-Systeme zu diskriminierenden Ausgaben führen können, ist in den Trainingsdaten zu finden auf denen sie trainiert wurden. Diese bilden im Internet verfügbare Daten ab und spiegeln damit wider, wenn beispielsweise marginalisierte Gruppen in Daten generell unterrepräsentiert sind. Die Problematik von diskriminierenden Ausgaben durch KI-Systemen fängt also bereits vor Entwicklung entsprechender Tools an. Gleichzeitig bietet sich hier die Möglichkeit und mittlerweile sogar gesetzliche Verpflichtung zur Sicherstellung von Fairness, damit Menschen nicht diskriminiert werden.

Dafür gibt es im Bereich der Softwareentwicklung eine Reihe von Fairness-Konzepten, die je nach Anwendungsfall ausgewählt werden können und Fairness sicherstellen sollen. In diesem Beitrag werden Fairness-Konzepte für sogenanntes Maschinelles Lernen (ML) betrachtet. Hierauf basieren die meisten aktuell genutzten KI-Anwendungen, wie z.B. die sogenannte Generative KI (GenKI), die wiederum Basis z.B. als Basis von Tools wie beispielsweise Chat-GPT eingesetzt wird. Sie bezeichnet KI-Systeme, die neue Inhalte erzeugen können, statt nur vorhandene Daten zu analysieren, zu klassifizieren oder vorherzusagen.

Der Einsatz von GenKI bringt viele Möglichkeiten mit sich und gleichzeitig ist es hierbei besonders schwierig Fairness sicherzustellen. Aus Sicht der Autorinnen ist es wichtig in diesem Bereich einen interdisziplinären Diskurs zu stärken, da dieser bisher vor allem technisch bleibt (Renz, 2025, S. 240). Letztlich geht es zwar darum im technischen Kontext und im Tool selbst Fairness sicherzustellen, aber auch um die Gestaltung von Menschen und Technologie in einem soziotechnischen System. Dies geht über die Betrachtung technischer Lösungen hinaus, weshalb auch die theoretische Fundierung breit gefächert und interdisziplinär angelegt ist.

1.2 FORSCHUNGSFRAGE UND HYPOTHESEN

Trotz der Dringlichkeit KI in der Personalentwicklung einzusetzen, um den Veränderungen der Arbeitswelt gerecht zu werden, muss bei der Nutzung eine Nichtdiskriminierung sichergestellt sein. Dies kann rechtlich, durch die Einhaltung geltender Gesetze, nur bedingt sichergestellt werden. Zum einen aufgrund der soziologisch beleuchteten Herausforderung komplexer Diskriminierungen und zum anderen aufgrund von technischen Herausforderungen. Ein Ansatzpunkt ist daher die Sicherstellung von Fairness in entsprechenden KI-Tools. Mit Blick auf den effizienten Einsatz eines KI-Systems sollte ein Teil der Lösung ein technischer Ansatz sein, da nur ein solcher von Maschinen ausgeführt konsistent und skalierbarer wirken kann.

Daher wurde im Rahmen dieser Arbeit ein interdisziplinärer Ansatz verfolgt, der sowohl gesetzliche Bestimmungen als auch soziologische und technische Überlegungen heranzieht. Die Ergebnisse dieser Arbeit bedienen damit die Schnittstelle zwischen technischen Gegebenheiten und praktischen Einsatzmöglichkeiten.

Die Forschungsfrage lautet: Wie kann Fairness beim Einsatz von GenKI sichergestellt werden?

Das Ziel dieser Arbeit ist es mittels einer systematischen Literaturanalyse die bisher bestehenden technischen Ansätze zur Sicherstellung von Fairness in GenKI Systemen zu identifizieren, systematisch zu sammeln und schlussendlich für die Entwicklung von Handlungsempfehlungen heranzuziehen. Mittels einer systematischen Literaturanalyse können einerseits Erkenntnisse zu einer Technologie zusammengefasst und andererseits Lücken in der aktuellen Forschung identifiziert werden, die wiederum einen Anknüpfpunkt für weitere Forschung vorschlägt.

2. Theoretische Fundierung: Diskriminierungssensible Nutzung von Künstlicher Intelligenz in der Personalentwicklung

2.1 BEGRIFFLICHE UND THEORETISCHE GRUNDLAGEN VON FAIRNESS UND DISKRIMINEUNRG

Diskriminierung ist soziologisch betrachtet ein Phänomen, dass nicht aus der Handlung oder Einstellung eines Einzelnen hervorgeht, sondern Bestandteil und Ursprung sozialer Strukturen und Prozesse ist (Scherr et al., 2017, S. 40). Dem liegt ein Verständnis von Diskriminierung als „soziale Konstruktion und Verwendung von Unterscheidungen zwischen Personenkategorien und imaginären Gruppen, die mit Vorstellungen über Ähnlichkeit und Fremdheit, Zugehörigkeit und Nicht-Zugehörigkeit sowie über angemessene Positionen im Gefüge der sozialen Ungleichheit verbunden sind“ (Scherr et al., 2017, S. 39) zugrunde.

Das Recht auf Nichtdiskriminierung ergibt sich zunächst aus dem Grundgesetz Art. 3 Abs. 1 mit dem allgemeinen Gleichheitssatz, wonach alle Menschen vor dem Gesetz gleich sind. Aus Art. 3 Abs. 3 GG ergibt sich das Verbot der Benachteiligung wegen des Geschlechts, der Abstammung, der „Rasse“, der Sprache, der Heimat und Herkunft, des Glaubens oder der religiösen oder politischen Anschauung. Diese Merkmale werden im folgende als sensible Attribute zusammengefasst. Mit dem Allgemeinen Gleichbehandlungsgesetz (AGG) ist ebendieser Schutz für den Arbeitskontext konkretisiert. Mit § 1 AGG erweitert das AGG das GG um einen Handlungsauftrag zur Beseitigung von Benachteiligung die durch sensible Attribute entstehen. Nach § 2 AGG ist eine Benachteiligung insbesondere bei der Auswahl und Einstellungen von Mitarbeitenden und in Bezug auf Arbeitsbedingungen sowie in der Aus- und Weiterbildung unzulässig. Die sensiblen Merkmale sind daher auch im Kontext der Personalentwicklung relevant. Im juristischen Kontext stellen die genannten Merkmale zum jetzigen Zeitpunkt eine abschließende Aufzählung dar und werden daher herangezogen, wenn es darum geht die Nichtdiskriminierung eines KI-Systems sicherzustellen.

Für die Verbindung dieser gesetzlich festgeschriebenen Merkmale aus dem arbeitsrechtlichen Kontext zu dem Einsatz algorithmischer Systeme die potenziell diskriminierend sind, kann die Charta der Grundrechte der Europäischen Union (EU-GRCh) herangezogen werden auf die sich wiederum der EU AI Act bezieht. Hier ist festgehalten, dass auch die Diskriminierung in Ausgaben von KI-Systemen verboten sind. KI-Systeme müssen damit so reguliert werden, dass sie nicht gegen das Verbot der Diskriminierung verstoßen. Dem ist inhärent, dass es beim Einsatz von KI-Systemen potenziell zu Diskriminierung kommen kann. Diskriminierung kann neben den genannten Merkmalen auch durch einen nicht barrierefreien Zugang zu KI-Systemen entstehen. Es ist daher in der Gestaltung von KI-Systemen zu gewährleisten, dass Menschen mit Behinderung gleichberechtigt Zugang haben.

Nichtdiskriminierung, auch in Ausgaben durch Ausgaben KI gestützter Tools, ist gesetzlich festgelegt. In der Softwareentwicklung wird dies versucht über das Konzept der Fairness sicherzustellen.

Fairness kann als ein ethisches Prinzip betrachtet werden, welches in KI-Systemen verfolgt wird, damit die Ausgaben nichtdiskriminierend sind (Afreen et al., 2025, S. 4798). Bei der Herstellung von Fairness geht es darum, Diskriminierung zu beseitigen, während technische Verzerrungen betrachtet werden, die entweder positiv oder negativ sein können (Afreen et al., 2025, S. 4798). Aus mathematischer Sicht kann Fairness über Statistik sichergestellt werden. Fair bedeutet im statistischen Sinne zunächst eine Gleichheit in der Verteilung, also keine Ungleichheit oder ungewollte Verzerrung. Aber Gleichheit (Equality) bedeutet nicht zwangsläufig Gerechtigkeit (Equity). Was fair ist, gilt es im jeweiligen Kontext zu definieren und kann auch erst dann sichergestellt sein.

Die Gestaltung eines diskriminierungssensiblen Einsatzes von KI in der Personalentwicklung befindet sich an einer Schnittstelle innerhalb des soziotechnischen Systems mit den Akteuren „Mensch“ und „Technologie“. In der kontextualisierten Betrachtung der Beziehung beider Systeme zueinander entstehen Möglichkeiten der konkreten Ausgestaltung.

2.2 MASCHINELLES LERNEN UND FAIRNESSKONZEPTE

Ein Großteil der heute unter dem Begriff der KI zusammengefassten Anwendungen basiert auf Methoden von ML. Charakteristisch für ML ist, dass Entscheidungsregeln nicht explizit durch Menschen programmiert werden, sondern durch Algorithmen aus Daten erlernt werden. Dabei werden große Datenmengen analysiert, um Muster zu identifizieren und daraus Vorhersagen oder Entscheidungen für neue Eingaben abzuleiten. Die grundlegenden Lernarten beim maschinellen Lernen sind *Supervised Learning*, *Unsupervised Learning* und *Reinforcement Learning*. Klassische Anwendungsfälle des MLs sind Regression und Klassifikation, bei denen Modelle entweder numerische Zielwerte prognostizieren oder Eingaben bestimmten Kategorien zuordnen.

Mit Fortschritten in der KI-Forschung und entsprechenden Weiterentwicklungen der verfügbaren KI-Anwendungen sind jedoch zunehmend komplexe, mehrstufige Modelle entstanden, die unterschiedliche Lernverfahren miteinander kombinieren. Hervorzuheben ist hier die GenKI, die in den letzten Jahren besonders an Popularität gewonnen hat. Generative Künstliche Intelligenz basiert häufig auf einer Kombination der oben genannten Lernarten. GenKI-Modelle werden zunächst auf großen Datenmengen vortrainiert und anschließend für spezifische Aufgaben angepasst (Ertel, 2025, S. 203). Durch diese mehrstufige Struktur entstehen zusätzliche Herausforderungen hinsichtlich der Nachvollziehbarkeit und Bewertung ihrer Ergebnisse.

Für die oben genannten grundlegenden Lernarten des ML gibt es etablierte Fairnessmetriken, die auf Basis statistischer Forderungen an das Verhältnis zwischen Modell Ein- und Ausgabe formuliert sind und so für ein bestehendes Modell überprüft werden können. Auf diese Weise lassen sich verschiedene Fairnessanforderungen für „einfache“ KI-Systeme technisch sicherstellen. Aufgrund der Komplexität lässt sich dies jedoch nicht direkt auf GenKI übertragen.

Eine weitere Herausforderung für die Fairnessbewertung von GenKI liegt in der Struktur der Ein- und Ausgabe. Generative KI-Systeme erzeugen Inhalte wie Texte, Bilder oder Videos und erlauben diese auch als Eingabe. Dies macht Ein- und Ausgaben schwerer vergleichbar, schlechter quantifizierbar und

somit schlechter in statistischen Anforderungen zu nutzen. Die erschwert die Anwendung klassischer Fairnessmetriken. Eine Fairnessbetrachtung ist jedoch unabdingbar, da häufig aus dem Internet stammenden Trainingsdaten verwendet werden. Diese Daten enthalten die Verzerrungen, die sich nachweislich in diskriminierenden oder stereotypen Ausgaben äußern (Kotek et al., 2023; Kruspe & Stillman, 2024; Locke & Hodgdon, 2025; Yang, 2025).

Eine weitere Dimension, die Fairnessbewertung herausfordernd macht ist, dass diese weiterhin vom Anwendungskontext abhängt. Um den Anwendungskontext einbeziehen zu können, sollten die Konzepte *Equality* (Gleichberechtigung) und *Equity* (Chancengleichheit) einbezogen werden. *Equality* beschreibt eine Gleichbehandlung aller Personen, etwa wenn ein System identische Entscheidungen für alle Nutzenden trifft. Equity hingegen zielt auf eine gerechte Behandlung ab, bei der individuelle Unterschiede berücksichtigt werden, um vergleichbare Chancen oder Ergebnisse zu ermöglichen. Insbesondere in vielen Anwendungskontexten der Personalentwicklung kann eine strikte Gleichbehandlung sogar zu unfairen Ergebnissen führen, während eine differenzierte Behandlung unter Berücksichtigung relevanter Merkmale als gerechter bewertet wird. So kann beispielsweise ein KI-System, dass für Personalentwicklungsmaßnahmen eingesetzt wird, bestehende Geschlechterungleichheiten verstärken. Sind die Trainingsdaten verzerrt und Frauen in Führungspositionen unterrepräsentiert, könnte das KI-System Frauen seltener für Führungskräfteentwicklungsmaßnahmen vorschlagen. Die Vorschläge des Systems wären dann zwar für alle Mitarbeitenden gleich, aber nicht fair, da es zu einer geschlechterspezifischen Benachteiligung führen würde.

3. Methodisches Vorgehen

Für die Identifizierung bestehender technischer Lösungsansätze, wie Fairness in GenKI sichergestellt werden kann, wurde eine Systematische Literaturanalyse durchgeführt.

3.1 DATENERHEBUNG UND DURCHFÜHRUNG

Zur Identifizierung passender Schlüsselbegriffe und Entwicklung der Suchstrategie wurde das SPIDER Schema verwendet (Cooke et al., 2012). SPIDER steht als Akronym für *Sample, Phenomenon of Interest, Design, Evaluation und Research type* (Cooke et al., 2012, S. 1437). Entsprechend des Forschungsvorhabens erfolgte eine Operationalisierung der Elemente des Schemas. Das *Sample* ist in dieser technologieorientierten systematischen Literaturanalyse die Funktionalität der Technologie selbst, also GenKI. Das *Phenomenon of Interest* ist Diskriminierung und Fairness in GenKI. Das *Design* bezieht sich auf die in der Studie verwendete Methode der Datenerhebung und soll Robustheit sicherstellen (Cooke et al., 2012, S. 1438). Für das Forschungsvorhaben wurde dieses Element für die Entwicklung der Suchstrategie nicht berücksichtigt, da es hierbei zunächst um die Identifikation aller relevanten Studien geht. Als Evaluation wurden Lösungsansätze zur Herstellung von Fairness und Vermeidung von Diskriminierung erwartet. Der Research Type dient der Suche nach technischen Arbeiten. Eine Übersetzung der S, PI und E Elemente des SPIDER Schemas in Suchbegriffe wurde wie folgt vorgenommen:

S	Generative AI oder GenAI
PI	Fairness oder Bias oder discrimination oder equity
E	Mitigation oder mitigate oder reduction oder reduce

Aus den Schlüsselwörtern wurde dann durch Verbindung mittels boolescher Operatoren zwei Suchstrings gebildet.

Suchstring 1:

„Generative AI“ OR „GenAI“ AND
„fairness“ OR „equity“ AND
„increase“ OR „ensure“ OR „establish“

Suchstring 2:

„Generative AI“ OR „GenAI“ AND
„bias“ OR „discriminiation“ AND
„mitigation“ OR „mitigate“ OR „reduction“ OR „recduce“

Die Literaturrecherche wurde in den wissenschaftlichen Datenbanken ACM Digital Library, IEEE Xplore und ScienceDirect (Elsevier) durchgeführt, da diese insbesondere technikorientierte Forschung abdecken und damit für die Beantwortung der Forschungsfrage geeignet sind. Die Suchstrings wurden jeweils an die spezifische Abfragesyntax der Datenbanken angepasst. Zusätzlich wurde der Publikationszeitraum auf Studien ab dem Jahr 2018 begrenzt, um aktuelle Forschung im Kontext generativer KI zu berücksichtigen. Die Datenbanksuche erfolgte am 9. November 2025 und ergab insgesamt 1.291 Treffer.
Die identifizierten Publikationen wurden anschließend exportiert und in eine strukturierte Excel-Arbeitsmappe überführt, um den weiteren Screening-Prozess zu unterstützen. Ergänzend zur Datenbanksuche wurde eine Suche nach grauer Literatur auf der Preprint-Plattform arXiv durchgeführt, um aktuelle, möglicherweise noch nicht peer-reviewte Forschung einzubeziehen. Die dort identifizierten relevanten Beiträge wurden ebenfalls in das Screening aufgenommen und anschließend anhand definierter Ein- und Ausschlusskriterien geprüft. Diese wurden für die Auswahl und Identifizierung passender Literatur zur Beantwortung der Fragestellung formuliert. Eingeschlossen wurden Studien, die sich auf die Technologie der GenKI beziehen. Weiterhin wurden Studien eingeschlossen, die sich mit Fairness beziehungsweise Verzerrung befassen und konkrete Lösungen vorschlagen diese herzustellen oder zu reduzieren. Gesucht waren technische Publikationen mit konkreten Lösungsansätzen, sodass technische Arbeiten eingeschlossen wurden. Um entsprechende der Fragestellung aktuelle und relevante Studien einzuschließen wurde der Publikationszeitraum auf zwischen 2018 bis 2025 festgelegt. Es wurden englischsprachige, veröffentlichte Studien eingeschlossen und Studien die, neben der grauen Literatur, peer-reviewed sind. Ausgeschlossen wurden folglich Studien in denen es nicht um GenKI geht. Außerdem ebendiese, in denen keine Konzeptionalisierung von Fairness/Bias erfolgte, oder keine Verknüpfung von Modell beziehungsweise Methode zu Fairness/Bias erfolgte. Weiterhin wurden Studien ausgeschlossen, in denen ausschließlich einer theoretischen oder ethischen Diskussion der Problematik erfolgte. Auch quantitative Studien wurden ausgeschlossen, sowie Studien, die vor 2018 oder nicht auf Englisch veröffentlicht wurden. Ebenso wurden Studien ausgeschlossen, deren Volltext nicht zugängig war und die nicht peer-reviewed sind. Dies galt nicht für die graue Literatur.

3.2 DATENSICHTUNG UND ANALYSE

Die Datensichtung erfolgte im Rahmen eines mehrstufigen Screening-Prozesses, um aus den identifizierten Publikationen diejenigen Studien auszuwählen, die für die Beantwortung der Forschungsfrage relevant sind. Insgesamt wurden zunächst 1.291 Publikationen aus den Datenbanken ScienceDirect,

ACM Digital Library und IEEE Xplore identifiziert. Nach der Entfernung von 356 Duplikaten verblieben 945 Datensätze für das Screening. Im ersten Schritt wurden Titel und Abstracts anhand der definierten Ein- und Ausschlusskriterien geprüft, wodurch 914 Publikationen ausgeschlossen wurden. Für die verbleibenden 31 Studien wurden anschließend die Volltexte analysiert. Im Zuge dieser Volltextprüfung wurden weitere 16 Arbeiten ausgeschlossen, sodass schließlich 15 Studien in die systematische Literaturanalyse einbezogen wurden.

Für die Analyse der eingeschlossenen Studien wurde eine qualitative Inhaltsanalyse nach Mayring genutzt. Ziel dieses Vorgehens war eine strukturierte und nachvollziehbare Auswertung der identifizierten Volltexte anhand theoriegeleiteter Kriterien. Hierfür wurde ein deduktives Kategoriensystem entwickelt, das sich sowohl an der Forschungsfrage als auch an den zuvor erarbeiteten theoretischen Grundlagen orientiert. Die Kategorien wurden vorab definiert und mit Kodierungsregeln versehen, sodass eine systematische Zuordnung relevanter Textstellen möglich war. Die genutzten Analysekategorien sind: die Art des Lösungsansatz, der Eingriffszeitpunkt der Maßnahme im ML -Prozess, weitere nichttechnische Lösungsansätze, der jeweilige Typ generativer KI sowie der Stand der Implementierung der vorgeschlagenen Lösungen.

Die einbezogenen 15 Studien (zwischen 2024 und 2025 publiziert) konzentrieren sich vor allem auf die Untersuchung und Reduktion von Bias in generativer KI. Im Mittelpunkt stehen dabei sowohl technische Ansätze zur Bias-Mitigation in Sprach- und Bildmodellen als auch ethische und governance-bezogene Fragestellungen. Die Arbeiten adressieren Anwendungsbereiche wie Bildung, Medizin, Softwareentwicklung und Empfehlungssysteme und betonen die Bedeutung von Fairness, Transparenz und verantwortungsvoller KI-Nutzung.

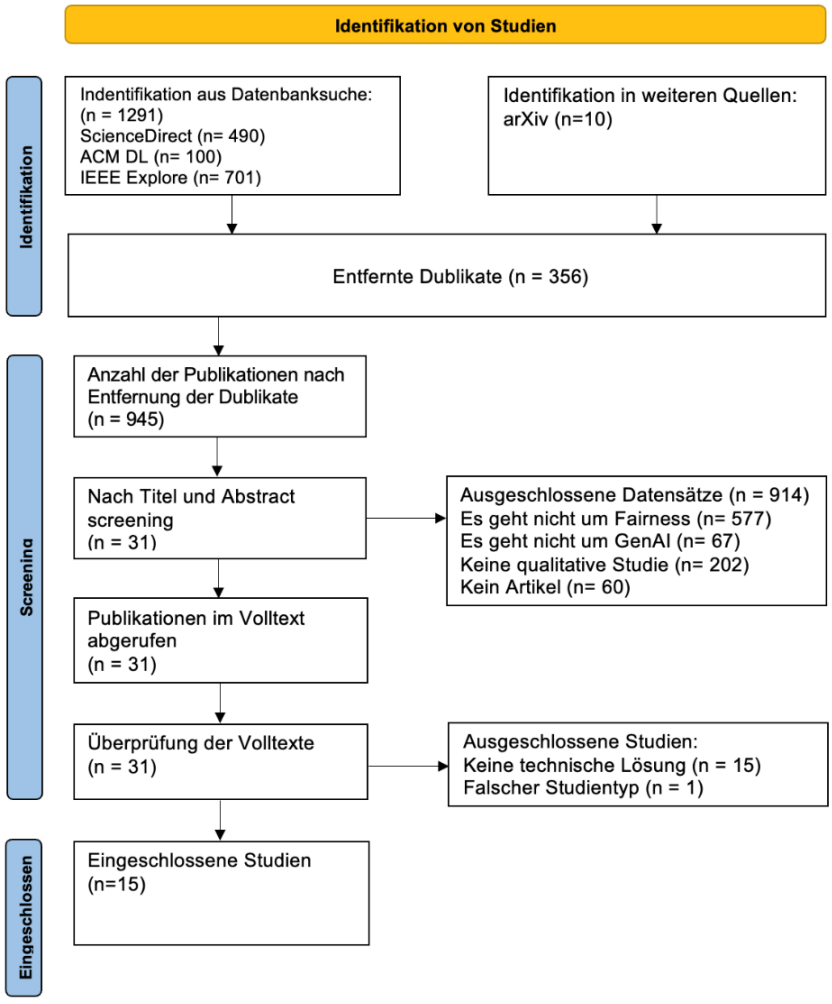
Folgende Studien wurden eingeschlossen:

1. A Vision for Operationalising Diversity and Inclusion in AI (Bano et al., 2024)
2. Mitigating Biased, Brittle and Baroque Generative AI (Maybury, 2025)
3. Towards Inclusive Educational AI: Auditing Frontier LLMs for Cultural Biases through a Multiplexity Lens (Mushtaq et al., 2025)
4. Improving Ethical Considerations in GenAI Responses Using (Sarukkai, 2024)
5. Responsible Generative AI for Software Development Life Cycle (Shah & Bajpai, 2025)
6. Data and AI governance: Promoting equity, ethics, and fairness in large language (Abhishek et al., 2025)
7. Evaluating and Mitigating Bias in AI-Based Medical Text Generation (Chen et al., 2025)
8. Mitigation of Gender and Ethnicity Bias in AI-Generated Stories through Model Explanations (Dimgba et al., 2025)
9. Reasoning Towards Fairness: Mitigating Bias in Language Models through Reasoning-Guided Fine-Tuning (Kabra et al., 2025)
10. DeBiasMe: De-biasing Human-AI Interactions with Metacognitive AIED (AI in Education) Interventions (Lim, 2025)

11. LLM4Rec: Large Language Models for Multimodal Generative Recommendation with Causal Debiasing (Ma et al., 2025)
12. Fine-Grained Bias Detection in LLM: Enhancing detection mechanisms for nuanced biases. (Mohanty, 2025)
13. Fair Generation without Unfair Distortions: Debiasing Text-to-Image Generation with Entanglement-Free Attention (Park et al., 2025)
14. Bias Mitigation Agent: Optimizing Source Selection for Fair and Balanced Knowledge Retrieval (Singh et al., 2025)
15. Generative AI and Power Imbalances in Global Education: Frameworks for Bias Mitigation (Nyaaba et al., 2025)

Die nachstehende Prisma Flow Chart zeigt den Identifikationsprozess und die schlussendlich eingeschlossenen Publikationen.

Abb. 1 Prisma Flow Chart



Quelle: Eigene Darstellung

Zur Qualitätssicherung der eingeschlossenen Studien wurde ein strukturierter Bewertungsansatz auf Basis der Fragen des *Critical Appraisal Skills Programme* (Critical Appraisal Skills Programme, 2024) angewendet. Da die identifizierten Arbeiten überwiegend theoretische und konzeptionelle Beiträge darstellen, wurde der ursprüngliche Fragenkatalog angepasst und die Qualität der eingeschlossenen Studien wie folgt ermittelt: 1. Wurde eine klare und testbare Formulierung der Ziele der Forschung gegeben? 2. War das Forschungsdesign geeignet, um die Ziele der Forschung zu erreichen? 3. Wurden die Daten in einer Weise erhoben, die die Forschungsfrage adressiert? 4. Wurden ethische Fragen berücksichtigt? 5. Gibt es eine klare Darstellung der Ergebnisse? 6. Sind die Ergebnisse selbst reproduzierbar beziehungsweise öffentlich verfügbar?

Einzelne Fragen, die sich beispielsweise auf qualitative Datenerhebungen oder die Beziehung zwischen Forschenden und Teilnehmenden beziehen, wurden ausgeschlossen. Ergänzend wurden Kriterien aufgenommen, die insbesondere den Peer-Review-Status sowie die Transparenz und Nachvollziehbarkeit der Forschung berücksichtigen.

Die strukturierte Bewertung ermöglichte eine systematische Einschätzung der Vertrauenswürdigkeit der Studien und machte zugleich Stärken und Schwächen der jeweiligen Beiträge transparent. Die Qualitätsbewertung der eingeschlossenen Studien zeigt insgesamt ein solides wissenschaftliches Fundament der eingeschlossenen Studien. 12 der 15 Arbeiten (Nummern 1-4, 6-9, 11-14) weisen eine klare und testbare Beschreibung des Forschungsziels auf. Einschränkungen zeigen sich jedoch teilweise hinsichtlich der empirischen Validierung der vorgeschlagenen Methoden sowie der ausführlichen Dokumentation von Datensätzen und Implementierungsdetails. So weisen nur neun der 15 Arbeiten (Nummer 1, 2, 7-9, 11-14) eine klare Darstellung der Ergebnisse auf und nur bei sieben (3, 7-9, 11, 13-14) können sie als teilweise reproduzierbar beziehungsweise teilweise reproduzierbar bezeichnet werden. Aufgrund der hohen Aktualität des Themas waren zum Zeitpunkt der Analyse nur fünf Arbeiten in begutachteten wissenschaftlichen Publikationsformaten veröffentlicht (1-5).

4. Ergebnisse

4.1 AUSWERTUNG DER LITERATURANALYSE

Die analysierten Studien schlagen unterschiedliche technische Ansätze zur Reduktion von Bias in generativen KI-Systemen vor. Ein Teil der Arbeiten entwickelt Evaluationsframeworks und Testmethoden zur systematischen Identifikation von Verzerrungen in generierten Inhalten(Abhishek et al., 2025; Mohanty, 2025; Shah & Bajpai, 2025). Andere Studien konzentrieren sich auf Anpassungen während des Modelltrainings, etwa durch modifizierte Trainingsstrategien oder Fine-Tuning-Verfahren mit kuratierten Datensätzen (Kabra et al., 2025; Ma et al., 2025; Park et al., 2025). Ziel dieser Methoden ist eine Reduktion von Bias bereits während der Modelloptimierung. Weitere Ansätze adressieren die Auswahl und Strukturierung von Informationsquellen, beispielsweise durch mehrstufige Systeme oder agentenbasierte Architekturen, die verschiedene Perspektiven in den Generierungsprozess einbeziehen (Bano et al., 2024; Singh et al., 2025). Darüber hinaus beschreiben einige Arbeiten Strategien zur Steuerung generierter Inhalte durch strukturierte Prompting-Techniken oder zusätzliche Kontrollmechanismen (Dimgba et al., 2025; Sarukkai, 2024). Zusammenfassend zeigen die eingeschlossenen Arbeiten, dass Bias-Reduktion in generativer KI an verschiedenen Stellen des Entwicklungs- und Nutzungsprozesses ansetzen kann.

Die systematischen Literaturanalyse wurde mit der Frage nach technischen Lösungsansätzen, um Fairness beim Einsatz von GenKI sicherzustellen durchgeführt. Alle Studien schlagen einen solchen vor, allerdings unterscheidet sie sich in der Art und Weise wie konkret die Lösung beschrieben und ausgeführt wird. Wie schon im Rahmen der Qualitätsbewertung angeführt, unterscheiden sich diese hinsichtlich des Reifegrades. Ein Teil der Studien formuliert konzeptionelle oder normative Vorschläge zur Reduktion von Bias in generativen KI-Systemen, ohne konkrete technische Implementierungen bereitzustellen. Andere Studien beschreiben spezifische methodische Ansätze, etwa Anpassungen von Trainingsverfahren, Modellarchitekturen oder Evaluationsmethoden. Somit stellen die Studien entweder eine konzeptionelle oder abstrakte Lösung vor, eine konkrete technische Lösung, oder es konnte zusätzlich ein nachweisbarer Impakt der eingesetzten technischen Lösung aufgezeigt werden. Aus der Analyse der Studien lassen sich damit drei Kategorien ausmachen, die hier für die Synthetisierung der Ergebnisse herangezogen werden. In Tab. 1 befindet sich eine Zuordnung der eingeschlossenen Studien zu den aufgeführten Kategorien.

Tab. 1 Kategorisierung der Studien

Kategorie	**Stand der Implementierung**	**Studien**
1	Arbeiten mit konzeptioneller oder abstrakter Lösung	1, 2, 5, 6, 15
2	Arbeiten mit konkreter technischer Lösung	4, 10, 12
3	Arbeiten mit nachweisbarem Impakt	3, 7, 8, 9, 11, 13, 14

Quelle: Eigene Darstellung

Die Unterscheidung von Kategorie 1 und Kategorie 2 ergibt sich aus der praktischen Umsetzbarkeit sowie dem Konkretisierungsgrad der vorgeschlagenen Maßnahmen. Bei Kategorie 3 kommt zur praktischen Umsetzbarkeit ein Nachweis der Wirksamkeit hinzu.

Studien der Kategorie 1 ergänzende nicht-technische Ansätze zur Reduktion von Bias in generativen KI-Systemen. Diese Ansätze adressieren organisatorische und prozessuale Aspekte der Entwicklung und Nutzung von generativer KI. Ein Schwerpunkt liegt auf der Zusammensetzung interdisziplinärer und diverser Entwicklungsteams (2), um unterschiedliche Perspektiven in den Entwicklungsprozess einzubeziehen. Weitere Vorschläge betreffen eine stärkere Dokumentation von Datensätzen, Modellannahmen und Trainingsprozessen, um Transparenz und Nachvollziehbarkeit zu erhöhen (5, 15). Zudem betonen einige Arbeiten die Bedeutung von Richtlinien und Schulungen für Entwicklerinnen und Entwickler sowie für Anwenderinnen und Anwender, um einen bewussten und reflektierten Umgang mit generativer KI zu fördern (2).

Die technischen Lösungen aus Kategorie 2 und 3 setzen an verschiedenen Zeitpunkten von Modellentwicklung bis -einsatz an. In zwei Studien basiert die technische Lösung darauf die Outputs zu gestalten (4, 8). Sechs der Studien präsentieren einen technischen Ansatz, über den versucht wird das Modelldesign anzupassen, um *Biases* zu reduzieren (9, 10, 11, 12, 13, 14). *Pre-process* kann bei der Datenauswahl, in der Planungsphase beziehungsweise des Designprozesses und während der Anforderungsanalyse angesetzt werden, um Fairness sicherzustellen. Acht Ansätze der analysierten Papiere sind dieser Phase zuzuordnen (1,2,3,5,6,13,14,15). Ansätze von vier der Studien ließen sich der Phase *in-process* zuordnen (2,6,7,11). Mit den Ansätzen wird im Training der Software und bei Verarbeitung der Dateneingabe Korrektur angesetzt. *Post-process*, also eine technische Lösung die nach dem Lernen ansetzt, wird in vier

Studien vorgeschlagen (6,8,10,13). Nahezu alle Ansätze sind auf textbasiete GenKI fokusiert, lediglich Ansätze aus drei Arbeiten (7, 11, 13) beschäftigen sich mit konkret mit multimodaler GenKI.

Zusammenfassend lässt sich sagen, dass die identifizierten technischen Ansätze anf verschiedenen Stellen der Entwicklung beziehungsweise des Einsatzes ansetzen. Aktuell sind post-processing Ansätze führend wenn es darum geht, nachweislich wirksam Verzerrung zu reduzieren. Gleichzeitig zeigt sich jedoch, dass die Anzahl konkreter technischer Lösungsansätze gering ist. Viele Ansätze scheinen sich noch in einem frühen Stadium zu befinden und für wenige ist eine Wirksamkeit empirisch nachgewiesen. Hinsichtlich der betrachteten sensiblen Attribute bestehen zudem Einschränkungen, sodass festgehalten werden muss, dass ein Anatz der für unterschiedliche GenKI-Typen und verschiedene sensible Attribute generalisiert werden kann, zum Zeitpunkt der Untersuchung nicht vorgestellt wurde.

4.2 FAIRNESSMETRIKEN FÜR DIE PRAXIS

Im Rahmen der systematischen Literaturanalyse wurde untersucht, inwiefern KI in der Personalentwicklung diskriminierungssensibel eingesetzt werden kann, indem die Potenziale zur Sicherstellung algorithmischer Fairness analysiert wurden. Sie können für den Einsatz entsprechender Tools Orientierung bieten und das Bewusstsein für Fairness und Nichtdiskriminierung schärfen.

Die Ergebnisse der systematischen Literaturanalyse unterstreichen, dass die Verantwortung der Sicherstellung von Fairness und Nichtdiskriminierung sowohl bei den entwickelnden Personen als auch bei den nutzenden Personen eines Tools liegt. Die entwickelten Handlungsempfehlungen setzen daher an zwei Stellen an und sind in Ergänzung zueinander zu denken. Zum einen bei der Auswahl des Tools und zum anderen bei dem Einsatz des Tools. Zudem gilt es, die Komplexität Fairness in KI-Tools sicherzustellen, die von der dahinterstehenden Art des MLs abhängt. Daher bildet die Art des MLs, die als Technologie eingesetzt wird, den Ausgangspunkt zu Überlegungen, wie eine Nichtdiskriminierung sichergestellt werden kann. Es wird daher empfohlen, dass einzusetzende Tool systematisch zu überprüfen und an geeigneten Stellen Maßnahmen zu ergreifen. Im Folgenden werden die Handlungsempfehlungen, basierend auf den Ergebnissen der SLR, beschrieben.

Basiert das eingesetzte Tool auf *Supervised*, *Unsupervised* oder *Reinforcement Learning*, gilt es für den jeweiligen Anwendungsfall festzulegen, welche Ausgabe als gerecht zu bewerten ist. Je nachdem wie *Equity* und *Equality* für den individuellen Anwendungsfall zu bewerten sind, muss eine der benannten Fairness Metriken eingesetzt und technisch sichergestellt sein. Ganz konkret muss hier durch das Unternehmen oder die Abteilung, die das Tool einsetzt, daher im vorliegenden Fall in der Abteilung der Personalentwicklung sichergestellt werden, dass das eingesetzte Tool auf die beschriebene Art Fairness sicherstellt. Dafür kann es notwendig sein mit dem Tool-Hersteller in Kontakt zu treten oder zunächst mit der IT-Abteilung des Unternehmens.

Der Nutzen dieser Überprüfung zu Beginn besteht vor allem darin, dass die Ausgaben in der Folge als fair angesehen werden können. Dies ermöglicht einen diskriminierungssensiblen Einsatz von KI in der Personalentwicklung. Da für diese Art der Systeme keine Prompts genutzt werden, sondern die Ausgabe feststeht, können Nutzende auf diese Weise keine Diskriminierung einbringen und die Trainingsdaten wurden entsprechend einer Fairness-Metrik bereinigt.

Basiert das eingesetzte Tool auf GenKI, braucht es andere Lösungsansätze, da es keine festgelegten Ausgaben gibt. Entsprechend der Erkenntnisse der systematischen Literaturanalyse existieren bisher nur Ansätze, die mit zur Entwicklungszeit festgelegten sensiblen Attributen arbeiten. Daher muss bei der Tool-Auswahl ebenfalls darauf geachtet werden, für welche sensiblen Attribute Fairness sichergestellt wird. Für andere sensible Attribute ist dies dann nicht automatisch auch der Fall.

Gleichwohl ist eine kritische Haltung wichtig und es gilt stets Ausgaben zu prüfen. Ausgaben einer GenKI Systems sollten nicht alleinig für die Entscheidung über zum Beispiel eine Beförderung herangezogene werden. Letztlich ist es eine Art Zusammenarbeit zwischen Menschen und KI, die einen Mehrwert bietet.

Abschließend bleibt festzuhalten, dass der Einsatz von KI die auf *Supervised*, *Unsupervised* oder *Reinforcement Learning* basiert, besonders mittels Sicherstellung der Verwendung einer Fairness-Metrik diskriminierungssensibel erfolgen kann. Gleichwohl löst Fairness in einem Tool nicht die Herausforderungen von Diskriminierung grundsätzlich. Wird ein KI-Systeme im Unternehmen eingesetzt gilt es die Einhaltung der gesetzlichen Bestimmungen, wie zum Beispiel der DSGVO, sicherzustellen. GenKI Systeme sind grundsätzlich potenziell diskriminierend und es gibt bisher keine technische Lösung, die generalisiert für alle sensible Attribute und GenKI-Typen, Fairness und Nichtdiskriminierung sicherstellt. Im besten Fall erfolgt daher zum einen die Sicherstellung von Fairness bei der Tool-Auswahl und zum anderen ein diskriminierungssensibler Tool Einsatz in dem die Mitarbeitenden entsprechend geschult werden.

7. Fazit

7.1 BEANTWORTUNG DER FORSCHUNGSFRAGE

Aus den gesetzlichen Vorgaben ergibt sich die Relevanz, eine Nichtdiskriminierung aufgrund sensibler Merkmale, auch in technischen Systemen sicherzustellen. Diversität, Nichtdiskriminierung und Fairness stellen einen eigenständigen Aspekt in der Auflistung der ALTAI (AI HLEG, 2025) dar und sind damit wichtig für die Gestaltung einer vertrauenswürdigen KI, die einen menschenzentrierten Ansatz verfolgt. Dies geht vor allem aus dem EU AI Act hervor, der über einen risikobasierten Ansatz KI-Systeme mit unterschiedlichen Regulationen belegt. Daraus ergibt sich besonders für die Personalentwicklung, dass ein Einsatz eines KI-Tools, das eine Bewertung der lernenden Person vornimmt und das Lernniveau entsprechend steuert, als risikoreich zu bewerten ist. Auch aus dem EU AI Act ergibt sich sowohl eine entwicklungsseitige als auch nutzungsseitige Verantwortung. Die Herausforderungen für Unternehmen besteht neben der sich fortlaufend verändernden Gesetzeslage in der genauen Überprüfung der jeweiligen Tools.

Die Herausforderung, KI-Systeme fair zu gestalten, ergibt sich unter anderem aus der Komplexität von Diskriminierung in Trainingsdaten. Mit den Erkenntnissen zur Intersektionalität wurde aufgezeigt, wie tief Diskriminierung in verschiedensten Strukturen verankert ist und dass Diskriminierung auch entlang mehrerer „Kategorien“ erfolgen kann. So ist durch eine Sicherstellung von Fairness für das Geschlecht nicht automatisch auch Fairness für Ethnie gewährleistet. Mit der *doing gender* Theorie wurde aufgezeigt, dass die im Gesetz erstmal recht statisch wirkenden „Kategorien“ sensibler Attribute durch die Gesellschaft hergestellt werden und einem Menschen nicht inhärent sind. Dies ist wichtig, um zu verstehen, warum Trainingsdaten, anhand der eine KI trainiert wird, potenziell diskriminierend sind und es in der Folge auch die Ausgaben sein können.

Es konnte aufgezeigt werden, dass es bisher wenig technische Lösungsansätze für Fairness in GenKI gibt und dieses sich hauptsächlich auf Textausgaben beziehen. Die Tatsache, dass bisher kaum technische Lösungsansätze mit bewiesener Wirksamkeit vorliegen, führt dazu, dass das Dilemma der Fairness nicht nur in der Verantwortung der entwickelnden Personen des Tools liegen. Auch Nutzende haben eine Verantwortung und können durch ihr Verhalten den Einsatz von KI diskriminierungssensibel gestalten. Wie sich dies gestalten ließe, beziehungsweise welche Maßnahmen für einen diskriminierungssensiblen Einsatz von KI in der Personalentwicklung ergriffen werden können, ist in den Handlungsempfehlungen formuliert. Dies ist besonders deshalb relevant, weil GenKI und generell KI in der Personalentwicklung eingesetzt wird, auch wenn die Herausforderungen bezüglich Fairness bestehen.

Vor allem mit Hinblick auf die Betrachtung von GenKI und Menschen als eigenständige Systeme innerhalb eines soziotechnischen Systems ist es sinnvoll einen ganzheitlichen Ansatz einer diskriminierungssensiblen Nutzung zu gestalten. Die Analyse technischer Lösungsansätze ist im Rahmen dieser Arbeit daher eingebettet in den sozialen Kontext. Die soziale Verantwortung und die Bedeutung einer technikmündigen und technikreflektierten Gesellschaft ergeben sich aus dem Einsatz von Gen-KI in verschiedensten Bereichen des menschlichen Lebens. Inwiefern algorithmische Automatisierung in für Menschen sensiblen Entscheidungsprozessen stattfinden sollte, gilt es abzuwägen, gleichwohl Diskriminierung eben auch durch Menschen passiert.

Inwiefern die Arbeitsgestaltung gelingt, hängt einerseits von dem Zusammenspiel der Systeme und andererseits von der „vernunftbasierte[n] Technikmündigkeit und Technikreflexion der Gesellschaft ab“ (Liggieri & Müller, 2019, S. 302).

Es bleibt festzuhalten, dass ganz grundsätzlich Anstrengungen unternommen werden sollten, strukturelle Diskriminierung abzubauen. Denn KI-Systeme sind durch die vorhandenen Trainingsdaten potenziell diskriminierend, nicht aus sich selbst heraus.

Gleichzeitig bleibt auch bei fortschreitender technischer Entwicklung eine wichtige Rolle für menschliche Akteurinnen und Akteure bestehen. Durch diskriminierungssensibles Prompten oder die kritische Prüfung generierter Inhalte, kann zumindest die Ausgabe überprüft werden. Letztlich sollte zukünftige Forschung daher sowohl technische Lösungen menschenzentrierte Ansätze berücksichtigen, um das soziotechnische Gefüge zwischen Mensch und KI zu gestalten.

Darüber hinaus stellt die Übertragung der wissenschaftlichen Erkenntnisse in praxisnahe Anwendungen eine wesentliche Herausforderung dar. Unternehmen stehen vor der Aufgabe GenKI Systeme zu nutzen und nutzen zu müssen und gleichzeitig einen rechtssicheren und nichtdiskriminierenden Einsatz sicherzustellen.

7.1 AUSBLICK

Ein diskriminierungssensibler Einsatz von KI ist nicht nur ethisch geboten, sondern auch rechtlich vorgeschrieben, etwa durch Antidiskriminierungsrecht und KI-spezifische Regulierungen. Auf Seiten der KI-Forschung besteht weiterhin die offene Aufgabe, Ansätze zu entwickeln, die nachweislich für alle ge-

setzlich relevanten sensiblen Attribute funktionieren. Während bestehende Verfahren häufig nur einzelne Dimensionen adressieren, bedarf es robuster Methoden, die auch bei komplexen, sich überschneidenden Merkmalsausprägungen verlässliche Aussagen zur Fairness von Modellen erlauben.

Daran anschließend liegt auf Seiten der KI-Unternehmen die Herausforderung, diese wissenschaftlich fundierten Ansätze in konkrete KI-Produkte zu integrieren und unter realistischen Einsatzbedingungen zu erproben. Dies betrifft beispielsweise die Implementierung automatisierter Fairness-Checks oder transparenter Erklärmechanismen in Lernplattformen oder Talentmanagementsystemen, die in der Praxis dauerhaft wartbar und nachvollziehbar sein müssen. Mit Blick auf den diskriminierungssensiblen Einsatz wären skalierbaren Verfahren, um Bias in multimodalen generativen KI-Ausgaben, etwa in Text-Bild-Kombinationen oder erstellen Videos, zuverlässig zu erkennen, besonders lohnend.

Daraus ergibt sich ein klarer offener Punkt für Forschung und Entwicklung. Ein Forschungsbedarf liegt in der Suche nach Fairness-Metriken die auch für GenKI nutzbar gemacht werden können. Weiterhin gilt es Ansätze zu finden, wie Diskriminierung in multimedialen Ausgaben vermieden werden kann. Auch die Herausforderung, dass eine technische Lösung für verschiedene sensible Attribute generalisierbar anwendbar sein muss, sollte Gegenstand der weiteren Entwicklungsarbeit sein. Darüber hinaus muss die Forschung für den Umgang mit indirekter Diskriminierung intensiviert werden

Für aktuell existierende und sicherlich in Einsatz befindlichen GenKI-Systemen fällt hier aktuell umso mehr Verantwortung auf die Anwendenden, die Ausgaben mit menschlichem Einsatz prüfen müssen. Aber auch wenn technische Herausforderungen gemeistert werden, bleiben für die Anwendenden zentrale Aufgaben bestehen. Diskriminierungssensibler Prompten stellt einen wichtigen Baustein dar, etwa bei der Generierung von Trainingsinhalten oder Entwicklungs-empfehlungen. Mit den zuvor erwähnten *Bias*-Erkennungen eröffnet sich hier eventuell die Möglichkeit, gezielte Feedback-Schleifen zum diskriminierungssensiblen Prompten zu etablieren. Werden beispielsweise systematische Verzerrungen in generierten Lernmaterialien identifiziert, können Prompts angepasst und iterativ verbessert werden, sodass strukturelle Defizite in den zugrunde liegenden Daten zumindest teilweise auf Ebene der Anwendung reflektiert und kompensiert werden.

Literaturverzeichnis

Abhishek, A., Erickson, L., & Bandopadhyay, T. (2025). Data and AI governance: Promoting equity, ethics, and fairness in large language models. *MIT Science Policy Review*, *6*, 139–146. https://doi.org/10.38105/spr.1sn574k4lp

Bano, M., Zowghi, D., & Gervasi, V. (2024). A Vision for Operationalising Diversity and Inclusion in AI. *Proceedings of the 2nd International Workshop on Responsible AI Engineering*, 36–45. https://doi.org/10.1145/3643691.3648587

Chen, X., Wang, T., Zhou, J., Song, Z., Gao, X., & Zhang, X. (2025). *Evaluating and Mitigating Bias in AI-Based Medical Text Generation* (arXiv:2504.17279). arXiv. https://doi.org/10.48550/arXiv.2504.17279

Critical Appraisal Skills Programme. (2024). *CASP qualitative checklist* (Critical Appraisal Skills Programme (CASP), Hrsg.). https://casp-uk.net/casp-checklists/CASP-checklist-qualitative-2024.pdf

Dimgba, M. O., Oba, S., Agrawal, A., & Giabbanelli, P. J. (2025). *Mitigation of Gender and Ethnicity Bias in AI-Generated Stories through Model Explanations*.

Ertel, W. (2025). *Introduction to Artificial Intelligence*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-43102-0

Fink, V. (2021). *Künstliche Intelligenz in der Personalarbeit: Potenziale nutzen und verantwortungsbewusst handeln* (1. Auflage). Schäffer-Poeschel. https://doi.org/10.34156/9783791052212

Hägerbäumer, M., & Drewes, S. (2025). Die Bedeutung von Future Skills in HR-Management und Corporate Learning. In M. Hägerbäumer, U. Thelen, & A. Renz (Hrsg.), *Future Skills in Human Resource Management und Corporate Learning: Neue Perspektiven durch Analytics, EdTech und KI*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-46481-3

Hasenbein, M. (2023). *Mensch und KI in Organisationen: Einfluss und Umsetzung Künstlicher Intelligenz in wirtschaftspsychologischen Anwendungsfeldern*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-66375-2

Kabra, S., Jha, A., & Reddy, C. K. (2025). *Reasoning Towards Fairness: Mitigating Bias in Language Models through Reasoning-Guided Fine-Tuning* (arXiv:2504.05632). arXiv. https://doi.org/10.48550/arXiv.2504.05632

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in Large Language Models. *Proceedings of The ACM Collective Intelligence Conference*, 12–24. https://doi.org/10.1145/3582269.3615599

Kruspe, A., & Stillman, M. (2024). Saxony-Anhalt is the Worst: Bias Towards German Federal States in Large Language Models. In A. Hotho & S. Rudolph (Hrsg.), *KI 2024: Advances in Artificial Intelligence* (S. 160–174). Springer Nature Switzerland.

Lim, C. (2025). *DeBiasMe: De-biasing Human-AI Interactions with Metacognitive AIED (AI in Education) Interventions* (arXiv:2504.16770). arXiv. https://doi.org/10.48550/arXiv.2504.16770

Locke, L. G., & Hodgdon, G. (2025). Gender bias in visual generative artificial intelligence systems and the socialization of AI. *AI & SOCIETY*, *40*(4), 2229–2236. https://doi.org/10.1007/s00146-024-02129-1

Ma, B., Li, H., Hu, Z., Gui, X., Liu, L., & Lau, S. (2025). *LLM4Rec: Large Language Models for Multimodal Generative Recommendation with Causal Debiasing* (arXiv:2510.01622). arXiv. https://doi.org/10.48550/arXiv.2510.01622

Maybury, M. (2025). Mitigating Biased, Brittle and Baroque Generative AI. *2025 IEEE International Conference on AI and Data Analytics (ICAD)*, 1–7. https://doi.org/10.1109/ICAD65464.2025.11114038

Mohanty, S. S. (2025). *Fine-Grained Bias Detection in LLM : Enhancing detection mechanisms for nuanced biases.*

Mushtaq, A., Naeem, R., Taj, I., Ghaznavi, I., & Qadir, J. (2025). Towards Inclusive Educational AI: Auditing Frontier LLMs for Cultural Biases through a Multiplexity Lens. *2025 IEEE Global Engineering Education Conference (EDUCON)*, 1–10. https://doi.org/10.1109/EDUCON62633.2025.11016405

Nyaaba, M., Wright, A., & Choi, G. L. (2025). *Generative AI and Power Imbalances in Global Education: Frameworks for Bias Mitigation*.

Park, J., Lee, J., Chung, C., Lee, J., Choo, J., & Gu, J. (2025). *Fair Generation without Unfair Distortions: Debiasing Text-to-Image Generation with Entanglement-Free Attention* (arXiv:2506.13298). arXiv. https://doi.org/10.48550/arXiv.2506.13298

Renz, A. (2025). Educational Technology und Corporate Learning. In M. Hägerbäumer, U. Thelen, & A. Renz (Hrsg.), *Future Skills in Human Resource Management und Corporate Learning: Neue Perspektiven durch Analytics, EdTech und KI* (S. 227–251). Springer Fachmedien Wiesbaden.

Sarukkai, A. R. (2024). Improving Ethical Considerations in GenAI Responses Using Introspection. *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*, 198–199. https://doi.org/10.1109/IRI62200.2024.00049

Shah, J. A., & Bajpai, G. (2025). Responsible Generative AI for Software Development Life Cycle. *2025 IEEE World AI IoT Congress (AIIoT)*, 0056–0061. https://doi.org/10.1109/AIIoT65859.2025.11105309

Singh, K., Muppiri, D., & Ngu, W. (2025). *Bias Mitigation Agent: Optimizing Source Selection for Fair and Balanced Knowledge Retrieval* (arXiv:2508.18724). arXiv. https://doi.org/10.48550/arXiv.2508.18724

Yang, Y. (2025). Racial bias in AI-generated images. *AI & SOCIETY*, *40*(7), 5425–5437. https://doi.org/10.1007/s00146-025-02282-1