Hochschulforum Digitalisierung

DISKUSSIONSPAPIER NR. 39 / JUNI 2026

On how to bring up baby robots –

Ein Plädoyer für einen bias-sensiblen und machtkritischen Umgang mit generativer KI an Hochschulen

Autorinnen

Johanna Leifeld und Sarah Becker

„Feminists could have a lot of advice to offer on bringing up babies – even when they are baby robots."

Alison Adam (1995)

„At the moment, Siri, Alexa [...], they are all baby robots. What we do now sets the scene for how these robots are gonna grow up in the future. My vision is that we all raise these baby robots to support gender equality in our communities, and that they will highlight our own behaviour back to us when we fail to uphold those values. This is about investing in our future, because if these chatbots do grow up to become our robot overlords, we'd better hope they are feminist overlords."

Josie Young (2019)

Warum dieses Thema?

> „AI is perpetuating our unjust systems. It's translating the unjust analog world into the digital one, and it's consequently widening the power imbalances in both."
>
> — Eva Gengler (2023)

Mit diesen Worten beschreibt Eva Gengler ein zentrales, wenn auch häufig unbeachtetes Problem gegenwärtiger Künstlicher Intelligenz. Generative Künstliche Intelligenz ist nicht objektiv und nicht neutral, denn sie wird mit Daten trainiert, die aus bestehenden gesellschaftlichen Strukturen stammen, und sie operiert innerhalb institutioneller Kontexte, die selbst von Machtverhältnissen, Ungleichheiten und Ausschlüssen geprägt sind. Anstatt diese gesellschaftlichen Strukturen automatisch durch Technologien zu überwinden, reproduzieren und verstärken KI-Systeme sie häufig, während sie nach außen hin objektiv wirken.

Auch Hochschulen und ihre Akteur:innen setzen zunehmend KI-Systeme ein: zur Studienberatung, zur Unterstützung beim Schreiben, zur Analyse von Lernverhalten oder zur Organisation administrativer Prozesse. Statt Zukunftsthema ist KI bereits Teil des hochschulischen Alltags und wird dabei häufig als effizientes, neutrales Werkzeug verstanden, das Prozesse optimiert und individuelle Entscheidungen unterstützt. Doch diese Neutralitätsannahme ist trügerisch bis falsch, denn zahlreiche Studien zeigen, dass KI-Systeme systematische Verzerrungen aufweisen können – etwa entlang von Geschlecht, Herkunft, Sprache oder sozialem Status (vgl. Noble, 2018; Kruspe et al., 2024; Gengler, 2024). Solche Bias-Effekte sind dabei kein rein technisches Problem fehlerhafter Algorithmen, sondern Ausdruck gesellschaftlicher Ungleichheiten, die in Daten, Modellen, Entwicklungs- und Nutzungskontexten eingeschrieben sind. Damit ist algorithmischer Bias nicht nur ein technisches Problem, sondern im Kern ein gesellschaftliches. Denn einerseits spiegelt er die bereits bestehenden diskriminierenden Strukturen in der Gesellschaft wider. Andererseits können diese verstärkt werden, wenn Nutzende KI-Outputs unreflektiert übernehmen und dadurch diskriminierende Deutungsmuster weiter verfestigen.

Gerade Hochschulen nehmen in diesem Zusammenhang eine besondere Rolle ein, denn sie sind nicht nur Orte des Lernens und der Forschung, sondern zentrale gesellschaftliche Institutionen der Wissensproduktion und -legitimation. Hochschulen bilden dabei die Entscheidungsträger:innen von morgen aus – Wissenschaftler:innen, Entwickler:innen, Führungskräfte, politische Akteur:innen – und prägen damit, wie KI in Zukunft in Wirtschaft, Verwaltung und Politik eingesetzt wird. Viele der heutigen Studierenden werden in naher Zukunft selbst an der Entwicklung, Implementierung oder Regulierung von KI-Systemen beteiligt sein oder sie in ihrem Berufsalltag nutzen, unabhängig davon, ob sie Informatik, Psychologie, Sozialwissenschaften oder andere Fächer studieren. Problematisch wird es, wenn die in KI-Systemen enthaltenen Verzerrungen unkritisch als objektive Wahrheiten übernommen werden und darauf basierende Entscheidungen unser Verständnis von Wirklichkeit und letztlich auch die gesellschaftliche Entwicklung in eine womöglich unerwünschte Richtung lenken. Die Art und Weise, wie Hochschulen heute mit KI umgehen, prägt daher nicht nur ihre eigenen Strukturen, sondern wirkt weit über sie hinaus. Wenn Hochschulen keine Sensibilität für die gesellschaftlichen und normativen Implikationen von KI vermitteln, stellt sich die Frage: Wo lernen wir es dann?

Auffällig ist jedoch, dass der hochschulische KI-Diskurs derzeit stark verengt ist. Im Zentrum stehen vor allem Fragen der Leistungsbewertung, des Prüfungsbetrugs und der Kontrolle: Wie lassen sich Täuschungsversuche erkennen? Wie kann die „Authentizität" studentischer Leistungen

sichergestellt werden? Technische Lösungen wie KI-Monitoring-Tools oder Detektionssoftware werden als Antwort auf ein vermeintliches Kontrollproblem diskutiert (Budde, Tobor & Friedrich, 2024), über das auch Lea Hildermeier und Inga Gostmann in ihrem Blogartikel sprechen. Andere Problemdefinitionen treten demgegenüber in den Hintergrund: Diskriminierung durch KI, die Reproduktion sozialer Ungleichheit, epistemische Verzerrungen in generierten Inhalten oder Ausschlüsse durch verzerrte Trainingsdaten, sprachliche Normierungen und implizite Normen und Maßstäbe in Daten und Modellen. Dass Bias im hochschulischen KI-Diskurs bislang selten explizit thematisiert wird, ist kein Zufall, denn Problemdefinitionen sind nie neutral. Sie entscheiden darüber, welche Risiken als relevant gelten, welche Maßnahmen ergriffen werden und wessen Perspektiven gehört werden. In der aktuellen Rahmung erscheinen vor allem jene Probleme als dringlich, die institutionell riskant sind, etwa für Prüfungsordnungen, Leistungsnachweise oder die rechtliche Absicherung von Hochschulen. Demgegenüber bleiben jene Effekte unsichtbar, die gesellschaftlich ungerecht wirken, aber weniger unmittelbar als institutionelle Bedrohung wahrgenommen werden.

Vor diesem Hintergrund verfolgt dieses Diskussionspapier drei Ziele: Erstens eine Sensibilisierung für Bias als zentrales, bislang marginalisiertes Thema im hochschulischen KI-Diskurs. Zweitens soll der aktuelle Stand zu KI und Bias an deutschen Hochschulen sichtbar gemacht werden. Drittens richtet es den Blick nach vorn und fragt, was sich künftig ändern muss und wie. Somit versteht sich dieses Diskussionspapier nicht nur als Problembeschreibung, sondern auch als Beitrag zur Frage, wie Hochschulen den Umgang mit KI verantwortungsvoll gestalten können.

Ein wenig Begriffspolitik: Viele Worte, und alles das Gleiche?

Wir bewegen uns in einem Diskurs, in dem unterschiedliche Akteur:innen mit unterschiedlichen Begriffen arbeiten. Die einen sprechen vom ethischen Einsatz von KI-Systemen, andere vom sozialverträglichen Einsatz. In manchen Texten ist von fairer KI die Rede, in anderen von feministischer KI. Doch handelt es sich dabei tatsächlich um unterschiedliche Konzepte, oder beschreiben sie im Kern dasselbe? Und wo verorten wir uns selbst in diesem Spannungsfeld? Welche Begriffe wählen wir bewusst?

Besonders häufig begegnet einem der Ausdruck „ethische KI". Für unseren Kontext ist er jedoch zu weit gefasst. Wir verstehen ihn eher als Oberbegriff, unter dem sich unterschiedliche Perspektiven und Ansätze versammeln. Dazu zählen etwa Datentransparenz, Datenschutz, Inklusion, Fairness oder auch Nachhaltigkeit. Oft bleibt jedoch unklar, welche dieser Dimensionen konkret gemeint ist. Gerade dadurch kann „ethische KI" wie eine Beruhigungsformel wirken: Der Ausdruck signalisiert Verantwortungsbewusstsein, ohne zwingend konkrete Resultate oder Maßnahmen zu benennen. Er überzeugt in Aufzählungen, entfaltet aber erst durch eine präzise Einordnung wirkliche Aussagekraft.

Auch der Begriff feministische KI wird in der Debatte häufig verwendet. In einem Interview beschreibt Eva Gengler darunter Ansätze zur Entwicklung und Nutzung von KI-Systemen, die über die bloße Reduktion von Bias hinausgehen und auf eine aktive Veränderung bestehender Macht- und Ungleichheitsverhältnisse zielen. Der Begriff selbst geht auf Alison Adam zurück, die zwischen einer feministisch gestalteten KI und einem feministischen Einsatz von KI unterscheidet – entscheidend ist in beiden Fällen der verfolgte Zweck. Feministische KI basiert dabei auf einem intersektionalen Verständnis von Diskriminierung und berücksichtigt, dass soziale Kategorien wie Geschlecht, Herkunft oder

Alter zusammenwirken und Ungleichheiten sich überschneiden und verstärken können. Entsprechend rückt in den Fokus, inwiefern KI-Systeme bestehende strukturelle Benachteiligungen reproduzieren und wie sie zugleich genutzt werden können, um diesen entgegenzuwirken.

Diese Definition kommt dem, was wir adressieren möchten, sehr nahe. Warum verwenden wir den Begriff dennoch nicht? Obwohl feministische KI inhaltlich präzise ist und unsere Zielsetzung gut beschreibt, verzichten wir in diesem Kontext bewusst darauf, denn Sprache setzt Schwerpunkte: Sie beeinflusst, welche Themen wahrgenommen und wie relevant sie eingeschätzt werden. Der Begriff „feministisch" ist dabei stark politisch markiert und stößt insbesondere in institutionellen Kontexten wie Hochschulen mitunter auf Vorbehalte oder wird als normativ überladen wahrgenommen. Auch wenn wir unsere Arbeit inhaltlich durchaus in dieser Tradition verorten, sprechen wir stattdessen von bias-sensibler und machtkritischer KI-Nutzung.

Das hat zwei Gründe: Zum einen wollen wir vermeiden, dass potenzielle Adressat:innen sich bereits beim Titel nicht angesprochen fühlen und innerlich aussteigen. Zum anderen machen diese Begriffe unmittelbar deutlich, worum es uns geht: um Sensibilisierung für bestehende Bias in KI-Systemen, um eine kritische Auseinandersetzung mit den Machtverhältnissen, die in ihrer Nutzung wirksam werden (können), und um den Fokus auf die Anwendung von KI. Denn Hochschulen entwickeln KI-Systeme bislang nur selten selbst. Sie haben weder Einfluss auf Datenauswahl noch auf Entwicklungsprozesse. So wichtig ein umfassendes Problemverständnis auch ist – im Hochschulkontext liegt der entscheidende Hebel aus unserer Sicht in einer reflektierten Nutzung dieser Systeme und in der Sensibilisierung und Vorbereitung derjenigen, die KI-Systeme in Zukunft beauftragen, entwickeln und beeinflussen: den Studierenden. Dies ist keine individuelle Aufgabe einzelner Lehrender oder Studierender, sondern muss als organisationsentwicklungsrelevante Handlungsaufgabe verstanden und von Hochschulen entsprechend unterstützt werden.

KI ist nicht neutral: Soziotechnische Praxis, Bias und Macht

Dieses Kapitel folgt der zentralen These, dass Bias in Künstlicher Intelligenz kein technischer Fehler, sondern Ausdruck gesellschaftlicher Machtverhältnisse ist. Verzerrungen in KI-Systemen entstehen nicht zufällig und lassen sich nicht allein durch bessere Daten oder optimierte Algorithmen beheben – auch wenn techno-solutionistische Ansätze genau davon ausgehen.

Der Begriff des *Techno-Solutionism* wurde maßgeblich durch Evgeny Morozov (2014) geprägt. Er beschreibt die Vorstellung, dass komplexe gesellschaftliche Probleme grundsätzlich als technische Probleme verstanden und mit den „richtigen" technologischen Lösungen behoben werden können. Soziale, politische oder institutionelle Herausforderungen erscheinen in dieser Perspektive vor allem als Fragen effizienter Optimierung. Digitale Tools, Algorithmen oder automatisierte Systeme gelten dabei häufig als objektiver, schneller und wirksamer als politische Aushandlungsprozesse, strukturelle Reformen oder soziale Interventionen. Gesellschaftliche Probleme werden dadurch häufig aus ihrem sozialen und historischen Kontext gelöst, auf messbare Variablen reduziert und als technisch optimierbare Aufgaben neu gerahmt.

Im Kontext von KI zeigt sich diese Denkweise besonders deutlich in der Annahme, dass Bias primär ein Datenproblem darstellt und sich folglich durch bereinigte, ausgewogenere oder „fairere" Trainingsdaten beheben lässt. Diese Perspektive wird der Komplexität des Problems jedoch nicht

gerecht. Zwar können verzerrte oder unrepräsentative Datensätze diskriminierende Effekte verstärken, doch Bias in KI-Systemen lässt sich nicht auf die Ebene der Trainingsdaten reduzieren, wie in einem späteren Kapitel gezeigt wird. Verzerrungen sind vielmehr das Ergebnis soziotechnischer Bedingungen, in denen technologische Entwicklung, institutionelle Zielsetzungen, Wissensordnungen und gesellschaftliche Machtstrukturen untrennbar miteinander verwoben sind.

Ausgangspunkt unserer Überlegungen ist daher ein Verständnis von KI als soziotechnische Praxis sowie eine feministische KI-Ethik, wie sie insbesondere von Eva Gengler entwickelt wird. Bias wird dabei nicht als isoliertes Fairnessproblem verstanden, sondern als strukturelles Problem, das aus historischen Ungleichheiten, epistemischen Ausschlüssen und politisch wirksamen Entscheidungen hervorgeht. Ergänzend beziehen wir uns auf die Arbeiten von Miranda Fricker zu epistemischer Ungerechtigkeit.

KI als soziotechnische Praxis – warum Neutralität ein Mythos ist

Künstliche Intelligenz wird im öffentlichen Diskurs häufig als neutrales und objektives Werkzeug verstanden, das unabhängig von gesellschaftlichen Kontexten funktioniert. Dieses Verständnis greift jedoch zu kurz. KI ist keine bloße Technologie, sondern eine soziotechnische Praxis – also ein Geflecht aus menschlichen Akteur:innen, technischen Systemen und institutionell-normativen Rahmenbedingungen. Technik wirkt dabei nicht als neutrales Werkzeug, sondern beeinflusst soziale Prozesse aktiv mit: Sie beeinflusst, welche Handlungen möglich erscheinen, welche Entscheidungen getroffen werden und welche Routinen sich etablieren, während sie zugleich selbst durch soziale Praktiken entwickelt, interpretiert und verändert wird.

Ein Beispiel dafür ist KI-gestützte Bewerber:innenvorauswahl in Personalabteilungen. Problematisch wird dies dort, wo bestehende gesellschaftliche Bewertungsmuster in technische Entscheidungsstrukturen übersetzt werden. Historisch gewachsene Vorstellungen von „geeigneten" Bildungs- und Erwerbsbiografien – lineare Karriereverläufe, kontinuierliche Vollzeitbeschäftigung oder Abschlüsse bestimmter Institutionen – werden dabei als relevante Merkmale in Trainingsdaten, Feature-Auswahl und Zielgrößen definiert. Bewerber:innen, deren Lebensläufe von diesen Normen abweichen, werden dadurch systematisch benachteiligt. Mit dem Einsatz solcher Systeme werden diese Bewertungsmaßstäbe zugleich stabilisiert und legitimiert. Vorauswahl, Rankings und Scores erscheinen als objektive Entscheidungshilfen, obwohl sie auf normativen Vorannahmen beruhen. In der Folge verändern sich dann organisationale Routinen: Profile außerhalb der algorithmisch erzeugten Rangordnung gelangen seltener in menschliche Entscheidungsprozesse, Verantwortung wird auf das System verlagert, und Ausschlüsse werden entpersonalisiert. Gleichzeitig passt sich die soziale Praxis an die technische Logik an. Bewerber:innen optimieren ihre Unterlagen für maschinelle Lesbarkeit, Beratungsangebote richten sich an algorithmischen Selektionskriterien aus, und Organisationen justieren Anforderungsprofile entlang dessen, was technisch erfass- und vergleichbar ist. KI wirkt hier als soziotechnische Praxis, in der gesellschaftliche Normen und technische Verfahren sich wechselseitig hervorbringen, verstärken und in neue Routinen überführen. Auf diese Weise können KI-Systeme bestehende Machtverhältnisse stabilisieren, indem sie bestehende soziale Ordnungen formalisieren, skalieren und in automatisierten Entscheidungsprozessen reproduzieren.

Der Begriff der soziotechnischen Praxis macht deutlich, dass KI-Systeme nicht außerhalb sozialer Ordnungen entstehen, sondern diese aktiv mit hervorbringen. Entscheidungen darüber, welche Daten gesammelt werden, welche Probleme als relevant gelten, welche Zielgrößen optimiert werden und

welche Anwendungen als legitim erscheinen, sind nie rein technisch, sondern immer auch normativ und politisch geprägt. KI-Systeme spiegeln deshalb bestehende Machtverhältnisse nicht nur wider, sondern können sie stabilisieren und verstärken. Dabei spielen auch institutionelle Zielsetzungen eine Rolle – etwa Effizienz, Vergleichbarkeit, Standardisierung oder Automatisierung. Solche Zielsetzungen prägen mit, wie KI-Systeme entwickelt und eingesetzt werden. Technik ist damit nicht neutral, sondern machtförmig strukturiert.

An dieses Verständnis von KI als soziotechnischer Praxis knüpft eine feministische KI-Ethik an. Gengler, Hagerer und Gales verstehen Diversität und feministische Perspektiven als intersektionalen und inklusiven Ansatz, der bestehende Machtstrukturen, Vorurteile und Stereotype in KI-Systemen sichtbar machen und transformieren will (Gengler et al., 2024a, S. 229–231). Ausgangspunkt ist die Beobachtung, dass viele gegenwärtige KI-Systeme gesellschaftliche Ungleichheiten nicht nur abbilden, sondern aktiv reproduzieren. Dadurch können sich bestehende Ungerechtigkeiten weiter verschärfen und somit die Welt noch ungerechter machen (Gengler et al., 2024a, S. 230). Ein feministischer Ansatz zur KI zielt dabei nicht nur auf einzelne Diskriminierungsformen, sondern auf das Zusammenspiel verschiedener Marginalisierungsdimensionen im Sinne einer intersektionalen Perspektive (Gengler et al., 2024a, S. 229–230). Fragen von Diversität, Inklusion und Gerechtigkeit sollen dabei nicht erst nachträglich berücksichtigt werden, sondern bereits bei Design, Entwicklung und Einsatz von KI-Systemen eine zentrale Rolle spielen. Die Entwicklung und Nutzung von KI erscheinen somit nicht als neutraler Optimierungsmechanismus, sondern als gestaltbare soziale Praxis. Entscheidend ist, ob sie bestehende Machtverhältnisse stabilisiert oder hinterfragt. Zugleich kritisieren Gengler et al., dass ein Großteil der bisherigen Forschung und Debatten zu KI und Ungleichheit auf der Ebene abstrakter ethischer Leitlinien verbleiben. Obwohl empirische Studien zeigen, dass KI bestehende soziale Ungleichheiten verschärft, werde die Bedeutung der sozialen Praxis und ihrer Prägung durch patriarchale und machtvolle Strukturen systematisch vernachlässigt (Gengler et al., 2025, S. 95). Ohne diese Perspektive drohen ethische Ansätze wirkungslos zu bleiben und auf der Ebene normativer Ideale zu verharren.

Der feministische KI-Ansatz macht deshalb deutlich, dass Fragen von Bias und Gerechtigkeit nicht erst auf der Ebene technischer Artefakte beginnen. Sie sind bereits in den sozialen, institutionellen und politischen Bedingungen verankert, unter denen KI entwickelt und eingesetzt wird. KI als soziotechnische Praxis zu verstehen, bedeutet daher, Machtverhältnisse und Ausschlussmechanismen von Beginn an mitzudenken und Neutralität nicht als Ausgangspunkt, sondern als kritische Fragestellung zu behandeln (vgl. Gengler et al., 2024a, S. 231–233).

Takeaway:

- KI ist keine neutrale Technologie, sondern eine soziotechnische Praxis.
- Sie entsteht im Zusammenspiel von technischen Entscheidungen, institutionellen Zielsetzungen und gesellschaftlichen Machtverhältnissen.
- Wer KI als objektives Werkzeug begreift, übersieht, dass soziale Normen und gesellschaftliche Praktiken, Bewertungsmaßstäbe und Ausschlussmechanismen aktiv in technische Systeme eingeschrieben werden und dass sich soziale Praktiken zugleich an die technischen Systeme anpassen. Es kommt somit zu einem Reproduktions- und Rückkopplungseffekt.

Was ist Bias?

Bias in KI bezeichnet systematische Verzerrungen in datengetriebenen Entscheidungs- und Klassifikationssystemen sowie generativen Anwendungen, durch die bestimmte Gruppen, Perspektiven oder Handlungsoptionen strukturell bevorzugt oder benachteiligt werden. Entscheidend ist dabei, dass diese Ungleichbehandlung nicht sachlich gerechtfertigt ist, sondern auf sozialen Zuschreibungen, historischen Ungleichheiten und normativen Vorannahmen beruht. Dadurch materialisieren und verstetigen sich bestehende Macht- und Bewertungsordnungen in diesen Systemen. In diesem Sinne definieren Gengler et al. Bias als

> „a systematic disadvantage, which is described as the unequal treatment of individuals from a particular group who do not differ from individuals in other groups in a way that justifies such disadvantages"
>
> — Gengler et al. (2024a), S. 230.

Bias ist damit nicht primär ein individuelles Fehlurteil, sondern eine Form systematischer Benachteiligung. Empirische Forschung zeigt, dass KI-Systeme insbesondere rassistische und geschlechtsspezifische Verzerrungen reproduzieren. Andere Diskriminierungsdimensionen – etwa sexuelle Identität, Alter, soziale Klasse oder Religion – sind ebenfalls relevant, werden bislang jedoch deutlich seltener systematisch untersucht. Studien weisen zudem darauf hin, dass Diskriminierung besonders stark dort ausfallen kann, wo mehrere Marginalisierungsdimensionen zusammenwirken. Das unterstreicht die zentrale Bedeutung einer intersektionalen Perspektive (vgl. Buolamwini & Gebru, 2018).

Diese Verzerrungen entstehen im Zusammenspiel sozialer Praktiken, technischer Entscheidungen und institutioneller Rahmenbedingungen. Um Bias analytisch greifbar zu machen, unterscheiden wir in Anschluss an Gengler et al. (2024a; 2025) vier zentrale Ebenen: Bias in *Trainingsdaten*, in der algorithmischen *Modellierung*, in *Entwicklungs- und Entscheidungsprozessen* sowie in der konkreten *Nutzung* von KI-Systemen.

Bias in Trainingsdaten

Bias kann bereits in den Trainingsdaten eines KI-Systems entstehen. Datensätze bilden die gesellschaftliche Wirklichkeit nicht neutral ab, sondern enthalten bestehende Ungleichheiten, Ausschlüsse und historische Verzerrungen. Wenn KI-Systeme auf solchen Daten trainiert werden, übernehmen sie diese Muster und führen sie automatisiert fort. Das zeigt sich etwa dort, wo bestimmte Gruppen in Datensätzen systematisch unterrepräsentiert sind oder verzerrt dargestellt werden (vgl. Guibeault et al., 2024; Shankar et al., 2017). Gengler et al. (2024a, S. 230f.) betonen, dass Trainingsdaten häufig Resultat historisch gewachsener sozialer Ordnungen sind, in denen marginalisierte Gruppen systematisch weniger sichtbar sind oder schlechter bewertet wurden. Diese Verzerrungen werden durch KI nicht nur übernommen, sondern skaliert und automatisiert weitergeführt.

Dies betrifft beispielsweise häufig die Unterrepräsentation von Frauen, People of Color und intersektional marginalisierten Gruppen in Datensätzen oder die Reproduktion historischer Diskriminierung z. B. in Hiring- oder Healthcare-Daten. In vielen Datensätzen sind Männer deutlich stärker vertreten als Frauen, Personen aus dem Globalen Norden dominieren, während marginalisierte Gruppen entweder verzerrt erscheinen oder gar nicht. Große Sprachmodelle basieren zudem überwiegend auf

englischsprachigen, frei zugänglichen Quellen aus dem Globalen Norden – Inhalte hinter Paywalls, aus nicht-westlichen Kontexten oder aus marginalisierten Wissensbeständen fehlen dort systematisch.

Bias in Trainingsdaten bedeutet damit nicht nur, dass Informationen fehlen oder unausgewogen verteilt sind. Entscheidend ist, dass bestehende gesellschaftliche Ungleichheiten als scheinbar neutrale Datengrundlage in technische Systeme eingehen.

Bias in Algorithmen und Modellierung

Bias entsteht nicht nur durch Trainingsdaten, sondern auch in der Art und Weise, wie KI-Systeme Informationen verarbeiten und Entscheidungen treffen. Auf dieser Ebene geht es um die technische Modelllogik selbst: also darum, welche Merkmale ein System als relevant einstuft, welche Zielgrößen optimiert werden und welche Fehler als akzeptabel gelten. Denn auch formal korrekt funktionierende Algorithmen können diskriminierende Effekte erzeugen. Das liegt nicht daran, dass sie „falsch" rechnen, sondern daran, dass technische Systeme immer entlang bestimmter Zielsetzungen entwickelt werden. Wird ein System beispielsweise vor allem auf Effizienz, Genauigkeit oder Profit optimiert, bleiben soziale Auswirkungen häufig unberücksichtigt.

Gengler et al. (2024a, S. 233; 2025, S. 92) machen deutlich, dass Algorithmen deshalb nicht neutral sind. In ihnen werden normative Annahmen technisch formalisiert: etwa darüber, welche Merkmale als relevant gelten, welche Kategorien gebildet werden oder welche Abweichungen als problematisch erscheinen. Welche Merkmale ein System berücksichtigt und welche Fehler als akzeptabel gelten, ist damit immer auch Ergebnis sozialer und institutioneller Prioritätensetzungen.

Im Hochschulkontext wird das besonders dort relevant, wo KI zur Bewertung, Sortierung oder Empfehlung eingesetzt wird, beispielsweise bei KI-Detektionssoftware. Solche Systeme müssen modellieren, welche sprachlichen Muster als „typisch menschlich" und welche als Hinweis auf KI-generierte Texte gelten. Damit ist immer auch eine implizite Vorstellung davon verbunden, was als erwartbares oder „normales" wissenschaftliches Schreiben gilt. Studien zeigen jedoch, dass nicht-muttersprachliche englische Texte deutlich häufiger fälschlicherweise als KI-generiert klassifiziert werden als Texte von Muttersprachler:innen (Liang et al., 2023). Das Problem liegt hier nicht primär in fehlerhaften Daten, sondern in den Bewertungsmaßstäben und Modellannahmen des Systems selbst. Ein System, das studentische Texte primär nach sprachlicher Kohärenz, formaler Struktur oder Ähnlichkeit zu bestehenden Mustern bewertet, könnte bestimmte Ausdrucksformen systematisch benachteiligen – etwa nicht-muttersprachliche Schreibweisen, alternative Argumentationsstile oder sprachliche Varianz. Bias entsteht hier nicht zwingend durch fehlerhafte Daten, sondern durch das Bewertungsprinzip des Systems selbst, indem bestimmte sprachliche Normen technisch als Standard gesetzt werden und Abweichungen davon als verdächtig erscheinen.

Bias auf der Entwicklungs- und Gestaltungsebene

Bias entsteht auch auf der Ebene derjenigen, die KI-Systeme entwerfen, entwickeln und über ihren Einsatz entscheiden. Während es bei der algorithmischen Modellierung um die technische Entscheidungsstruktur eines Systems geht, rückt hier die soziale Ebene in den Blick: Wer entwickelt KI? Wer setzt Prioritäten? Und wessen Perspektiven fehlen dabei?

Die Entwicklungs- und Gestaltungsebene bildet somit eine zentrale Schnittstelle zwischen gesellschaftlichen Machtverhältnissen und technischen Artefakten. Gengler et al. (2024a, S. 231) identifizieren insbesondere die mangelnde Diversität in Entwickler:innen-Teams und in strategischen Entscheidungspositionen als eine zentrale strukturelle Ursache für diskriminierende KI-Systeme. Insbesondere die Überrepräsentation weißer Männer in technischen und organisatorischen Schlüsselpositionen führt dazu, dass bestimmte Perspektiven, Problemlagen und Erfahrungswelten systematisch unberücksichtigt bleiben. Diese Homogenität betrifft dabei nicht nur die technische Entwicklung, sondern ebenso jene Akteur:innen, die über Budgets, Einsatzfelder und Personalentscheidungen bestimmen und somit darüber, wessen Interessen Priorität haben. Problematisch ist dabei nicht primär das individuelle Vorurteil einzelner Entwickler:innen, sondern die strukturelle Homogenität der Kontexte, in denen KI-Systeme entstehen. Denn diese Homogenität und im Umkehrschluss die institutionalisierten Ausschlüsse entscheiden darüber, welche Fragen überhaupt gestellt werden, welche potenzielle Fehlfunktionen nicht erkannt oder diskriminierende Effekte übersehen werden. KI-Systeme bauen so auf „data, logic, and power relations from the past" auf und reproduzieren diese in automatisierter Form (Gengler et al., 2024a, S. 231).

Empirisch spiegelt sich diese Problematik in der Zusammensetzung der KI-Industrie wider. Nach aktuellen Zahlen aus dem Jahr 2024 (Pal et al., 2024) sind nur etwa 22 % der Beschäftigten in der KI-Branche Frauen und marginalisierte Perspektiven fehlen darüber hinaus weitgehend. Entwickler:innen-Teams sind damit überwiegend männlich, weiß und im Globalen Norden angesiedelt. Gengler et al. (2024a) argumentieren, dass diese Homogenität dazu führt, dass patriarchale und koloniale Machtverhältnisse in technische Systeme einfließen.

Hier zeigt sich eine strukturelle Parallele zur Datenebene: Marginalisierte Gruppen sind häufig nicht nur in Trainingsdaten unterrepräsentiert oder verzerrt dargestellt, sondern auch in den Teams und Entscheidungskontexten, in denen KI-Systeme entwickelt und legitimiert werden. Bias entsteht damit nicht nur dadurch, *was* ein System lernt, sondern auch dadurch, *wer* darüber entscheidet, was es lernen, leisten und optimieren soll.

Diese Ebene ist auch für Hochschulen relevant, obwohl sie KI-Systeme meist nicht selbst entwickeln. Denn auch die Entscheidung darüber, welche Systeme institutionell eingesetzt, empfohlen oder legitimiert werden, ist nicht neutral. Die Zusammensetzung solcher Entscheidungsprozesse beeinflusst mit, welche Risiken wahrgenommen, welche Fragen gestellt und welche Perspektiven berücksichtigt werden.

Bias in Nutzung und Anwendung

Bias entsteht nicht nur bei der Entwicklung von KI-Systemen, sondern auch in ihrer konkreten Nutzung und der Einbettung in institutionelle Kontexte, deren Ziele, Leistungsverständnisse und Machtverhältnisse sie mittragen. Diskriminierende Effekte entstehen insbesondere dort, wo KI-generierte Ergebnisse als objektiv oder neutral interpretiert werden und kritische Einordnung, Kontextualisierung oder menschliche Kontrolle ausbleiben. Ohne eine solche Reflexion können bestehende Diskriminierungen unbemerkt reproduziert werden.

Gerade im Hochschulkontext ist diese Ebene besonders relevant. Generative KI-Systeme werden hier bereits für Recherche, Textproduktion, Studienorganisation oder perspektivisch auch für Beratung, Bewertung und administrative Entscheidungen genutzt. Problematisch wird dies dort, wo KI-Ergebnisse nicht als kontextabhängige technische Outputs, sondern als scheinbar neutrale Entscheidungshilfen verstanden werden. Studien zeigen etwa, dass Sprachmodelle in Empfehlungsschreiben Personen mit weiblichen Namen („Kelly") als warmherzig und emotional beschreiben, während Personen mit männlichen Namen („Joseph") automatisch als durchsetzungsstark und führungsfähig beschrieben werden (Wan et al., 2023). Solche Zuschreibungen beeinflussen nicht nur sprachliche Darstellungen, sondern potenziell auch reale Bewertungen, etwa bei Bewerbungen, Leistungsbewertungen oder Karriereverläufen. KI reproduziert hier also nicht einfach gesellschaftliche Vorurteile, sondern kann ihnen durch ihre scheinbare Objektivität zusätzliche Legitimität verleihen und bestehende Ungleichheiten verstärken.

Wie subtil solche Verzerrungen in großen Sprachmodellen wirken können, zeigen aktuelle Studien. Sprachmodelle schlagen etwa für Frauen oder Personen mit Migrationshintergrund systematisch niedrigere Gehälter vor (Sorokovikova et al., 2025). Problematisch ist dabei nicht nur die Verzerrung selbst, sondern ihre geringe Sichtbarkeit: Nutzer:innen erhalten in der Regel nur einen einzelnen plausiblen Output und haben kaum Vergleichsmöglichkeiten, um diskriminierende Muster zu erkennen. Wer etwa nicht parallel testet, welche Antwort ein männlich konnotierter Name erhalten hätte, nimmt den Output leicht als neutrale Empfehlung wahr. Auch geografische Biases in großen Sprachmodellen verdeutlichen diese Problematik. Eine aktuelle Studie zeigt, dass Modelle wie ChatGPT oder LeoLM ostdeutsche Bundesländer systematisch negativer bewerten als andere Regionen – etwa im Hinblick auf positive Eigenschaften wie Attraktivität, Intelligenz oder Arbeitsmoral, aber teilweise auch bei objektiv messbaren Merkmalen wie Bildungsniveau oder Säuglingssterblichkeit (Kruspe & Stillman, 2024). Besonders problematisch ist, dass sich solche Verzerrungen in generativen KI-Systemen häufig weniger offensichtlich zeigen als in klassischen datengetriebenen Entscheidungssystemen. Während im Fall des 2018 bekannt gewordenen Amazon-Recruiting-Tools systematische Benachteiligungen von Frauen als strukturelles Problem identifiziert werden konnten (Dastin, 2018), erscheinen Verzerrungen in Sprachmodellen oft als einzelne, sprachlich plausible Antworten. Bias verschwindet hier nicht, sondern wird subtiler – und gerade deshalb schwerer erkennbar. Das erhöht das Risiko, dass diskriminierende Annahmen in LLM-Outputs unkritisch übernommen werden.

Gengler et al. (2024a, S. 234–236) zeigen, dass KI-Systeme in organisationalen Routinen häufig Entscheidungen vorstrukturieren oder legitimieren. Das kann dazu führen, dass Verantwortung schrittweise auf technische Systeme verlagert wird und bestehende Machtverhältnisse stabilisiert werden. Im Hochschulkontext könnte sich dies etwa dort zeigen, wo KI-gestützte Bewertungssysteme bestimmte sprachliche Ausdrucksformen systematisch bevorzugen, KI-gestützte Beratung stereotype Annahmen reproduziert oder algorithmisch erzeugte Empfehlungen unkritisch als objektive Orientierung übernommen werden. Bias entsteht hier nicht nur durch das System selbst, sondern durch die Art und Weise, wie mit seinen Ergebnissen umgegangen wird.

Konklusion

Diese vier Ebenen zeigen, dass Bias in KI kein isolierter technischer Fehler ist, sondern aus dem Zusammenspiel soziotechnischer Faktoren entsteht. Die Konsequenz daraus ist, dass KI-Systeme gesellschaftliche Ungleichheiten nicht nur übernehmen, sondern sie durch Automatisierung, Skalierung und institutionelle Anwendung häufig verstärken können. Große Datenmengen oder technologische Komplexität kompensieren diese strukturellen Verzerrungen nicht automatisch. Vielmehr reproduzieren KI-Modelle bestehende Macht- und Bewertungsordnungen häufig in großem Maßstab und verleihen ihnen durch ihre technische Form zusätzliche Legitimität. Werden solche Systeme anschließend in sensiblen gesellschaftlichen Kontexten eingesetzt, können diskriminierende Effekte verstärkt werden. Bias verschiebt sich damit von einer sozialen Ungleichheit hin zu einer technisch vermittelten, scheinbar objektiven Entscheidung.

Dabei handelt es sich nicht um voneinander getrennte Problemfelder, sondern um gegenseitig verstärkende Mechanismen. Verzerrte Daten prägen algorithmische Modelle, diese Modelle entstehen in spezifischen Entwicklungs- und Entscheidungsstrukturen, und ihre Ergebnisse entfalten Wirkung in konkreten Nutzungskontexten. Diese Nutzung beeinflusst wiederum Entscheidungen, institutionelle Routinen und soziale Erwartungen und erzeugt damit neue Datengrundlagen, die in zukünftige Systeme zurückfließen. Bias wirkt damit nicht linear, sondern in Form von Rückkopplungsschleifen, in denen bestehende Ungleichheiten, Macht- und Bewertungsordnungen technisch reproduziert, stabilisiert und potenziell weiter verschärft werden.

Für Hochschulen ist diese Dynamik besonders relevant. Generative KI greift bereits zunehmend in zentrale akademische Praktiken ein: Studierende, Lehrende und Forschende nutzen KI für Recherche, wissenschaftliches Schreiben, Ideenfindung oder die Strukturierung wissenschaftlicher Inhalte. Darüber hinaus werden Anwendungen in Bewertungs-, Beratungs- oder administrativen Entscheidungsprozessen bereits diskutiert. Bias in KI betrifft im Hochschulkontext daher nicht nur technische Zuverlässigkeit oder die korrekte Anwendung einzelner Systeme. Wenn KI in solche akademischen Kernpraktiken eingreift, berührt dies grundlegende Fragen danach, welche Ausdrucksformen als wissenschaftlich legitim gelten, welche Perspektiven sichtbar bleiben und welche Formen von Wissen im akademischen Alltag marginalisiert werden.

Takeaway:

- Bias in KI ist kein einzelner Fehler, sondern ein strukturelles Phänomen.
- Er entsteht im Zusammenspiel von Trainingsdaten, algorithmischer Modellierung, Entwicklungs- und Gestaltungskontexten sowie konkreten Nutzungspraktiken. KI-Systeme reproduzieren dabei nicht nur bestehende gesellschaftliche Ungleichheiten, sondern stabilisieren und skalieren sie in institutionellen Entscheidungsprozessen.
- Eine wirksame Auseinandersetzung mit Bias erfordert daher nicht nur technische Korrekturen, sondern auch eine Reflexion der sozialen, epistemischen und institutionellen Bedingungen, unter denen KI entwickelt und eingesetzt wird.

Bias als epistemische Ungerechtigkeit

Die bisherige Analyse zeigt, dass Bias in KI im Hochschulkontext nicht nur Fragen technischer Fairness oder Zuverlässigkeit berührt. Eine weitergehende Frage lautet vielmehr: Was bedeutet es, wenn KI nicht nur Entscheidungen unterstützt oder Outputs produziert, sondern zunehmend in Prozesse wissenschaftlicher Wissensproduktion eingreift? Wenn KI diese Prozesse mitprägt, geht es nicht mehr allein um Fairness oder technische Zuverlässigkeit, sondern explizit auch um epistemische Gerechtigkeit. Um diese Perspektive zu schärfen, greifen wir im Folgenden auf Miranda Frickers Konzept epistemischer Ungerechtigkeit zurück.

Frickers Ansatz verschiebt den Fokus von der Frage, ob ein System „korrekt" funktioniert, hin zu der Frage, wie Wissen, Glaubwürdigkeit und gesellschaftliche Deutungsmacht verteilt sind. Er ermöglicht es, Verzerrungen in Wissensprozessen nicht lediglich als Fehler oder Effizienzprobleme zu beschreiben, sondern als Formen struktureller Ungleichbehandlung im Raum des Wissens (Fricker, 2023, S. 24f.), sowohl analog als auch digital. Epistemische Ungerechtigkeit bezeichnet Situationen, in denen Personen oder Gruppen als Wissenssubjekte benachteiligt werden, also darin, Wissen zu äußern, Erfahrungen verständlich zu machen oder als glaubwürdig anerkannt zu werden (Fricker, 2023, S. 24). Dafür wird zwischen zwei zentralen Formen epistemischer Ungerechtigkeit unterschieden: *Zeugnisungerechtigkeit* und *hermeneutische Ungerechtigkeit* (Fricker, 2023, S. 24f.).

Zeugnisungerechtigkeit liegt vor, wenn Personen aufgrund sozialer Vorurteile weniger Glaubwürdigkeit zugeschrieben wird, als ihnen eigentlich zukäme (Fricker, 2023, S. 25–27). Solche Zuschreibungen sind nicht zufällig, sondern knüpfen systematisch an soziale Machtverhältnisse und identitätsbezogene Vorurteile an, etwa im Hinblick auf Geschlecht, soziale Herkunft oder Zugehörigkeit zu marginalisierten Gruppen (Fricker, 2023, S. 26f.). Auch im Hochschulkontext lassen sich solche Mechanismen beobachten. Studierende können etwa in partizipativen oder strategischen Entscheidungsprozessen weniger ernst genommen werden, weil ihnen pauschal mangelnde Erfahrung oder fehlendes institutionelles Verständnis zugeschrieben wird. Ihre Beiträge werden dann nicht primär nach ihrem Inhalt bewertet, sondern durch Vorannahmen über ihre soziale Position als „nur Studierende" epistemisch abgewertet.

Übertragen auf KI-Systeme wird diese Form epistemischer Ungerechtigkeit dort relevant, wo algorithmische Bewertungen oder Empfehlungen bestehende Vorannahmen über Kompetenz, Autorität oder Glaubwürdigkeit reproduzieren und sie zugleich durch ihre scheinbare Objektivität stabilisieren und legitimieren können (vgl. Fricker, 2023, S. 27–28). Eine Studie von He (2025) zeigt etwa, dass große Sprachmodelle bei der Auswahl wissenschaftlicher Referenzen sowohl einen Citation Bias aufweisen, indem sie männlich verfasste Arbeiten bevorzugen, als auch einen Majority Bias reproduzieren, bei dem jene Geschlechtsgruppe bevorzugt wird, die im jeweiligen Datensatz oder Fachgebiet bereits numerisch dominiert. Wenn beispielsweise in einem Forschungsfeld überwiegend Männer publizieren, verstärkt das Modell diese Dominanz nochmals, indem es bevorzugt auf diese Arbeiten zurückgreift.

Während Zeugnisungerechtigkeit die ungleiche Anerkennung von Sprecher:innen betrifft, beschreibt die hermeneutische Ungerechtigkeit eine tiefere strukturelle Ebene. Gemeint sind Situationen, in denen Menschen und Gruppen nicht über die gesellschaftlichen Begriffe, Perspektiven oder Deutungsmuster verfügen, um ihre Erfahrungen angemessen zu verstehen oder verständlich zu machen (Fricker, 2023, S. 24f.; S. 30f.). Solche Lücken betreffen insbesondere Gruppen, deren Perspektiven in gesellschaftlichen Wissens- und Bedeutungsordnungen ohnehin marginalisiert sind und die nicht

im gleichen Maße an Praktiken gesellschaftlicher Bedeutungsproduktion beteiligt sind (Fricker, 2023, S. 30f.). Ein Beispiel im Hochschulkontext zeigt sich dort, wo Studierende problematische Erfahrungen mit KI-gestützten Systemen machen, diese aber nicht ohne Weiteres als strukturelles Problem benennen können. Wenn etwa KI-Detektionssysteme bestimmte sprachliche Ausdrucksweisen häufiger als verdächtig markieren oder generative KI vor allem dominante wissenschaftliche Perspektiven reproduziert, kann das von Betroffenen als Benachteiligung wahrgenommen werden. Fehlen jedoch die Begriffe, das technische Verständnis oder die institutionellen Deutungsressourcen, um diese Erfahrung als Bias, epistemische Benachteiligung oder strukturelles Problem zu artikulieren, bleibt die Benachteiligung schwer adressierbar.

Somit können KI-Systeme beide beschriebene Mechanismen epistemischer Ungerechtigkeit verstärken. Wenn Modelle auf dominanten Wissensbeständen oder gesellschaftlichen Stereotypen beruhen, können die Akteur:innen, die marginalisiertes Wissen einbringen, epistemisch abgewertet werden und es werden vor allem auch nicht-dominante Perspektiven systematisch ausgeblendet. Im Hochschulkontext ist das besonders relevant, weil generative KI-Systeme zunehmend auch Wissens- und Forschungsprozesse mit strukturieren. Das kann einerseits dazu führen, dass bestimmte wissenschaftliche Ausdrucksformen bevorzugt, nicht-standardisierte Argumentationsweisen als Defizit markiert oder dominante akademische Normen technisch reproduziert werden. Dadurch entsteht Druck zur epistemischen Standardisierung.

Andererseits greift generative KI subtil in die Produktion von Wissen selbst ein. Forschende, Lehrende und Studierende nutzen generative KI bereits in frühen Phasen wissenschaftlicher Wissensproduktion, etwa zur Ideenfindung, Literaturstrukturierung oder Formulierung von Forschungsfragen. Da diese Systeme in der Regel zunächst auf dominante, stark repräsentierte Wissensbestände zurückgreifen, orientieren sich die ersten Vorschläge häufig an etablierten, insbesondere westlich geprägten Mainstream-Diskursen, während marginalisierte, intersektionale oder nicht-dominante Perspektiven seltener sichtbar werden.

KI beeinflusst damit nicht nur, wie Wissen dargestellt oder bewertet wird, sondern auch, welches Wissen überhaupt als denkbar, relevant oder wissenschaftlich anschlussfähig erscheint. Generative KI bildet epistemische Ausschlüsse damit nicht nur ab, sondern kann aktiv in Prozesse akademischer Wissensproduktion eingreifen und bestehende epistemische Ungleichheiten verstärken.

Für Hochschulen ist die Perspektive epistemischer Ungerechtigkeit von besonderer Bedeutung, da sie nicht nur Orte der Wissensvermittlung sind, sondern epistemische Institutionen im starken Sinne. Sie produzieren, selektieren, legitimieren und sanktionieren Wissen und Wissenssubjekte zugleich. Damit strukturieren sie nicht nur, was als Wissen gilt, sondern auch, wer als glaubwürdig, kompetent und epistemisch relevant anerkannt wird (Fricker, 2023, S. 28–29).

Werden dominante Wissensbestände, bestehende Ausschlüsse und epistemische Machtverhältnisse in KI-Systemen unreflektiert übernommen, besteht die Gefahr, dass KI nicht nur bestehende Ungleichheiten abbildet, sondern diese in akademischen Wissenspraktiken verstärkt. Was zunächst als technische Unterstützung erscheint, kann so epistemische Ordnungen stabilisieren, in denen bestimmte Perspektiven systematisch sichtbar und legitim bleiben, während andere marginalisiert oder unsichtbar gemacht werden. KI verschiebt epistemische Ungerechtigkeit damit von einer oft impliziten sozialen Praxis hin zu einer technisch vermittelten und institutionell verfestigten Struktur. Wenn etwa bestimmte sprachliche Ausdrucksformen oder Forschungsansätze systematisch als wissenschaftlich legitimer markiert werden als andere, passen sich Nutzer:innen und institutionelle Routinen an diese Maßstäbe an. Bereits dominante Perspektiven werden dadurch weiter gestärkt,

während alternative Wissensformen zunehmend an Sichtbarkeit verlieren. Diese veränderten Wissenspraktiken fließen wiederum in die Daten- und Nutzungskontexte ein, auf deren Grundlage zukünftige KI-Systeme trainiert, angepasst und eingesetzt werden. Epistemische Ungleichheiten werden so nicht nur technisch abgebildet, sondern in Rückkopplungsschleifen sozial und institutionell verstärkt.

Bias in KI-Systemen verweist vor diesem Hintergrund auf grundlegende Fragen epistemischer Macht und sozialer Ungleichheit. Hochschulen tragen daher eine besondere Verantwortung, KI nicht nur fair oder effizient, sondern epistemisch gerecht zu gestalten und ihre Mitglieder für Bias und epistemische Ungerechtigkeit zu sensibilisieren.

Takeaway:

- Bias in KI ist nicht nur unfair, sondern epistemisch ungerecht.
- Er benachteiligt Personen und Gruppen in ihrer Rolle als Wissenssubjekte – etwa darin, gehört, verstanden oder als glaubwürdig anerkannt zu werden. Hochschulen tragen hierfür eine besondere Verantwortung, da sie als epistemische Institutionen darüber entscheiden, welches Wissen zählt und wessen Perspektiven legitim sind.
- Wird KI in diesen Kontexten unreflektiert eingesetzt, kann epistemische Ungerechtigkeit technisch verfestigt und institutionell legitimiert werden.

Aktueller Stand des Diskurses rund um KI an deutschen Hochschulen

Der Diskurs um den Einsatz Künstlicher Intelligenz an deutschen Hochschulen hat in den vergangenen Jahren deutlich an Dynamik gewonnen. Insbesondere seit 2023 setzen sich viele Hochschulen intensiver mit der Frage auseinander, ob, wann und in welchem Umfang KI-Tools von Studierenden und Lehrenden genutzt werden dürfen. Inzwischen existiert eine Vielzahl institutioneller Leitlinien, die Orientierung im Umgang mit generativer KI bieten und den Einsatz entsprechender Systeme strukturieren sollen. Damit ist das Thema sichtbar im Hochschulalltag angekommen. Fraglich ist jedoch, wie tief diese Auseinandersetzung tatsächlich reicht und ob KI bislang eher punktuell reguliert als strukturell in der Hochschule verankert wird.

Im aktuellen HFD-Blickpunkt zu KI-Leitlinien an deutschen Hochschulen (Becker et al., 2026) wurde untersucht, inwiefern Problematiken rund um Verzerrungen in KI-Systemen in bestehenden Leitlinien thematisiert werden. Dabei zeigte sich, dass Bias zwar vereinzelt erwähnt wird, meist jedoch nur in knapper Form als allgemeiner Risikohinweis. Lediglich etwa ein Viertel der untersuchten Dokumente greift das Thema überhaupt auf. Eine Einordnung als organisationsrelevantes Governance-Thema bleibt weitgehend aus. Ebenso fehlen konkrete Hinweise auf Verfahren, Zuständigkeiten oder institutionalisierte Maßnahmen, um mit solchen Verzerrungen systematisch umzugehen.

Auffällig ist dabei vor allem die Verschiebung von Verantwortung auf die individuelle Ebene. Obwohl Bias häufig als strukturelles Phänomen beschrieben wird, richten sich Handlungsempfehlungen primär an Studierende und Lehrende, die KI-generierte Inhalte kritisch prüfen sollen. Damit wird ein systemisches Problem vor allem über individuelles Verhalten adressiert. Diese Perspektive ist zwar anschlussfähig an wissenschaftliche Praxis und Eigenverantwortung, erzeugt jedoch eine deutliche Spannung: Strukturelle Ursachen werden kaum mit entsprechenden institutionellen Maßnahmen beantwortet. Begleitende Unterstützungsangebote, Schulungen oder langfristige Kompetenzstrukturen finden in den Leitlinien bislang nur selten Erwähnung.

Hinzu kommt, dass Bias häufig vor allem als technisches Qualitätsproblem erscheint, während seine gesellschaftliche Dimension unterbelichtet bleibt. Fragen danach, wie KI bestehende soziale Ungleichheiten reproduzieren oder verstärken kann, werden nur selten explizit gemacht. Würde Bias jedoch als sozial wirksamer Mechanismus verstanden, ergibt sich daraus zwangsläufig auch eine institutionelle Verantwortung für Hochschulen: nämlich Räume für kritische Reflexion zu schaffen und geeignete Rahmenbedingungen für einen verantwortungsvollen Umgang mit KI zu entwickeln.

Dabei existieren bereits zahlreiche Initiativen, Forschungsprojekte und Lehrangebote, die genau diese Perspektive aufgreifen. Forschungsprojekte zu fairer und verantwortungsvoller KI finden sich inzwischen an vielen Hochschulen. Beispielhaft lässt sich etwa ein Projekt der Universität Graz nennen, in dessen Rahmen ein Orientierungsleitfaden für KI-Nutzung in der Sozialen Arbeit entwickelt wurde. Auch an einzelnen Hochschulen entstehen Lehrangebote, die KI aus kritisch-reflektierten Perspektiven betrachten. An der Freien Universität Berlin werden beispielsweise Seminare angeboten, die KI aus einer kritisch-feministischen Perspektive analysieren. Die Universität Ulm wiederum entwickelte eine interaktive E-Learning-Einheit zu „Bias und Fairness algorithmischer Entscheidungssysteme“, die ein anwendungsorientiertes Verständnis von Fairness und Bias in KI-Systemen vermittelt. Ergänzend dazu bietet auch der KI-Campus frei zugängliche Online-Kurse an, die sich mit sozialverantwortlicher und diversitätsbewusster KI-Entwicklung beschäftigen.

Gerade diese Vielzahl an Einzelinitiativen macht ein grundlegendes Problem sichtbar: Angebote zu Bias und einer verantwortungsvollen KI-Nutzung existieren zwar, sie erscheinen jedoch häufig projektförmig und stark abhängig vom Engagement einzelner Lehrender oder Forschender. Eine systematische und curriculare Verankerung des Themas bleibt bislang weitgehend aus. Während Unternehmen durch den EU AI Act inzwischen verpflichtet sind, für angemessene KI-Kompetenzen ihrer Mitarbeitenden zu sorgen, einschließlich ethischer und gesellschaftlicher Fragestellungen, existieren vergleichbare verbindliche Anforderungen im Hochschulkontext bislang nicht. Dabei wäre gerade hier eine strukturelle Verankerung naheliegend. Fragen einer bias-sensiblen KI-Nutzung ließen sich beispielsweise gut in bestehende Formate wie wissenschaftliches Schreiben, Methodenlehre oder digitale Kompetenzvermittlung integrieren, also dort, wo ohnehin Fragen wissenschaftlicher Praxis und Reflexion verhandelt werden.

Insgesamt zeigt sich damit ein widersprüchliches Bild: Einerseits ist das Thema KI längst an Hochschulen angekommen, es existieren zahlreiche Forschungsprojekte und engagierte Einzelinitiativen. Andererseits fehlt bislang vielerorts eine strategische, institutionelle und curriculare Verankerung. Genau daraus ergibt sich die zentrale Frage:

Wie können Hochschulen Verantwortung übernehmen, um Bias-Sensibilität im Umgang mit KI nicht nur punktuell, sondern strukturell, wirksam und nachhaltig zu verankern – und welche konkreten Maßnahmen braucht es dafür?

Bias-Sensibilität institutionell gestalten

Wie sehen Hochschulen im Jahr 2035 aus, wenn KI-Systeme selbstverständlicher Teil von Lehre und Studium geworden sind? Die Antwort darauf wird unter anderem davon abhängen, welche Entscheidungen heute getroffen werden. Denkbar sind dabei zwei sehr unterschiedliche Szenarien.

Szenario 1: KI ist an Hochschulen vor allem ein Werkzeug zur Entlastung. Administrative Prozesse laufen effizienter, Lehrende gewinnen Zeit für Betreuung und Forschung, Studierende erhalten individualisierte Unterstützung, Barrieren werden abgebaut. KI hilft dabei, Bildungsangebote zugänglicher zu machen und kann Ungleichheiten sichtbar machen, statt sie zu verschärfen. Hochschulen haben klare Regeln, transparente Prüfmechanismen und eine ausgeprägte Bias-Sensibilität etabliert. Technologie wird hier nicht unkritisch eingesetzt, sondern reflektiert gestaltet.

Szenario 2: KI ist tief in hochschulische Prozesse integriert worden, ohne ihre Machtstrukturen und Verzerrungen ausreichend zu hinterfragen. Unter dem Druck knapper Ressourcen wurden Systeme eingeführt, die Zeit sparen und Entscheidungen automatisieren sollen, beispielsweise Prüfungsbewertungen und Zulassungsverfahren. Bias wird dabei nicht beseitigt, sondern skaliert. Lehrende verlassen sich auf Systeme, deren Entscheidungsstrukturen sie kaum nachvollziehen können. Studierende erleben Bewertungen als intransparent und potenziell diskriminierend. Hochschulen, die eigentlich Orte kritischer Reflexion und Chancengleichheit sein sollten, reproduzieren bestehende Ungleichheiten mithilfe technologischer Systeme.

Die vorherigen Kapitel haben gezeigt, dass KI-Systeme keineswegs neutral sind. Sprachmodelle reproduzieren gesellschaftliche Stereotype, verstärken bestehende Machtverhältnisse und erstellen Ergebnisse entlang verzerrter Datengrundlagen. Gerade weil KI zunehmend als hilfreiches Werkzeug zur Effizienzsteigerung erscheint, liegt darin eine besondere Gefahr für Hochschulen. In Zeiten von Ressourcenknappheit und steigender Arbeitsbelastung wirkt die Vorstellung attraktiv, unliebsame oder repetitive Aufgaben an Maschinen auszulagern. Wer würde sich nicht wünschen, Prüfungen schneller korrigieren, Texte automatisch bewerten oder Verwaltungsprozesse vereinfachen zu können?

Noch gibt es kaum belastbare Zahlen darüber, in welchem Umfang Lehrende bereits KI-Systeme zur Bewertung von Prüfungsleistungen einsetzen. Gleichzeitig deutet vieles darauf hin, dass solche Praktiken längst stattfinden. Ein Bericht von Times Higher Education im April 2026 warnt davor, denn insbesondere die Nutzung generativer KI-Systeme wie ChatGPT zur Bewertung studentischer Arbeiten könnte problematisch werden und unter den EU AI Act als „High-Risk"-Anwendung fallen. Die Anforderungen an Transparenz, Nachvollziehbarkeit und Trainingsdaten wären hier erheblich – Anforderungen, die viele aktuelle Systeme kaum erfüllen.

Noch werden Zulassungsverfahren in Deutschland nicht durch KI-Systeme unterstützt. Doch angesichts des wachsenden finanziellen und organisatorischen Drucks auf Hochschulen stellt sich die Frage, wie lange das so bleibt. Denkbar wäre etwa ein Szenario, in dem Bewerbungen automatisiert vorsortiert werden: Eine KI bewertet Motivationsschreiben, Lebensläufe und schulische Leistungen anhand historischer Daten früherer Zulassungen. Wenn diese Daten bereits soziale Ungleichheiten oder implizite Vorurteile enthalten, könnte das System Bewerber:innen aus bestimmten Regionen, mit nicht-akademischem Familienhintergrund oder mit sprachlich „abweichenden" Schreibstilen systematisch schlechter bewerten. Was folgt, wäre die Reproduktion diskriminierender Muster und die Ungleichbehandlung von Bewerber:innen.

Hochschulen sind nicht nur Ausbildungsorte, sondern Räume der Wissensproduktion und kritischen Reflexion. Sie prägen, wie zukünftige Entscheidungsträger:innen Technologie verstehen und einsetzen. Wenn Bias in KI-Systemen dort nicht thematisiert wird, besteht die Gefahr, dass Studierende algorithmische Entscheidungen später in Unternehmen, Verwaltung oder Politik als selbstverständlich akzeptieren, ohne ihre Voraussetzungen oder Konsequenzen zu hinterfragen. Hochschulen lehren eben nicht nur Fachwissen, sondern auch einen Umgang mit neuen Technologien und das im besten Fall kritisch und verantwortungsvoll. Im Folgenden werden anhand der bereits beschriebenen Vier-Ebenen-Struktur von Bias von Gengler et al. (2024, 2025) Zielbilder formuliert, um einen reflektierten Umgang mit KI zu fördern

Ebene 1: Trainingsdaten – Wessen Wissen zählt?

Auf der Ebene der Trainingsdaten hatten wir bereits folgende zwei Formen von Bias herausgestellt:

1. **Bias innerhalb der Trainingsdaten**
   Trainingsdaten bilden gesellschaftliche Ungleichheiten, Ausschlüsse und Machtverhältnisse ab. KI-Systeme können diese Verzerrungen übernehmen und verstärken.
2. **Bias durch Auswahl und Repräsentation von Wissensbeständen**
   KI-Systeme repräsentieren nicht Wissen „an sich", sondern bestimmte Wissensbestände. Viele genutzte KI-Anwendungen beruhen vor allem auf englischsprachigen, frei zugänglichen Textkorpora aus dem Globalen Norden. Dadurch sind etwa wissenschaftliche Literatur hinter Paywalls, nicht-westliche Wissensformen sowie feministische, postkoloniale oder disability-theoretische Perspektiven systematisch unterrepräsentiert.

Diese Problematik verschärft sich im Hochschulkontext, weil Hochschulen nicht nur Wissen vermitteln, sondern auch daran beteiligt sind, welches Wissen als legitim gilt. Werden KI-Systeme in Studium, Lehre oder Forschung eingesetzt, reproduzieren sie bestehende epistemische Hierarchien: Kanonisiertes Wissen wird weiter sichtbar gemacht, während marginalisierte Perspektiven an den Rand gedrängt bleiben. Studierende, die KI zur Recherche, Strukturierung oder zum Schreiben nutzen, erhalten so potenziell ein verzerrtes Bild dessen, was als relevantes oder legitimes Wissen gilt.

Gleichzeitig ist klar, dass Hochschulen auf dieser Ebene bislang nur begrenzten direkten Einfluss haben. Die meisten Hochschulen entwickeln keine eigenen KI-Systeme und können daher nur eingeschränkt auf Trainingsdaten einwirken. Umso wichtiger ist es, Studierenden Wissen über diese Zusammenhänge zu vermitteln und Fragen bias-sensibler KI-Nutzung in bestehende Lehrformate zu integrieren. Dies betrifft insbesondere Bereiche wie wissenschaftliches Schreiben, Methodenlehre oder digitale Kompetenzvermittlung, also Kontexte, in denen ohnehin Fragen wissenschaftlicher Praxis und Reflexion verhandelt werden.

Dabei geht es auch um grundlegende Fragen qualitativen wissenschaftlichen Arbeitens: Woher stammen die Daten, auf denen KI-Systeme basieren? Warum kann eine KI-gestützte Recherche wissenschaftliche Literaturrecherche nicht ersetzen? Welche Perspektiven fehlen möglicherweise in den generierten Antworten? Solche Fragen sollten gemeinsam mit Studierenden reflektiert werden. Ansatzpunkte für Hochschulen liegen hier etwa in der Reflexion über Datenquellen, sowie in einer expliziten Wissens- und Kanonkritik im Umgang mit KI.

Zugleich gilt: Die Studierenden von heute sind die Entwickler:innen und Entscheidungsträger:innen von morgen. Vermitteln Hochschulen frühzeitig einen reflektierten und verantwortungsvollen Umgang mit KI-Systemen, besteht die Hoffnung, dass Absolvent:innen später in technischen, politischen oder organisatorischen Positionen Einfluss auf gerechtere Datenpraktiken und inklusivere Wissensbestände nehmen können.

Zielbilder

Bias-sensible KI-Kompetenz in die Lehre integrieren

- Hochschulen vermitteln Studierenden, dass KI-Systeme auf begrenzten und häufig unausgewogenen Trainingsdaten basieren. Diese Reflexion wird vor allem in bereits bestehenden Formaten zu wissenschaftlichem Schreiben, Methodenlehre und digitaler Kompetenz verankert.

KI-gestützte Recherche kritisch einordnen

- Studierende lernen, dass KI wissenschaftliche Literaturrecherche nicht ersetzen kann. Hochschulen machen deutlich, dass KI-generierte Antworten immer auf Datenquellen, fehlende Perspektiven und mögliche Verzerrungen geprüft werden müssen.

Wissens- und Kanonkritik im Umgang mit KI stärken

- Hochschulen nehmen KI-Nutzung zum Anlass, sichtbar zu machen, welche Wissensbestände privilegiert werden und welche Perspektiven fehlen. Dazu gehört beispielsweise, marginalisierte, nicht-westliche, feministische und postkoloniale Perspektiven bewusst einzubeziehen.

Ebene 2: Algorithmische Modellierung: Implizite Normen und Bewertungsmaßstäbe

Bias wirkt an Hochschulen auch auf der Ebene der algorithmischen Modellierung. Algorithmen übersetzen komplexe soziale, sprachliche oder fachliche Inhalte in berechenbare Merkmale. Sie operieren mit impliziten Annahmen darüber, was als „gut", „normal", „effizient" oder „erfolgreich" gilt. Diese Annahmen sind nicht neutral, sondern spiegeln dominante akademische Bewertungslogiken wider – etwa Produktivität, sprachliche Standardisierung oder formale Kohärenz.

Ein gutes Beispiel hierfür ist der Einsatz von KI-Detektoren. Studien zeigen, dass solche Systeme Texte von Nicht-Muttersprachler:innen überdurchschnittlich häufig fälschlicherweise als KI-generiert klassifizieren (Liang et al., 2023). In einer Untersuchung lag die Falsch-Positiv-Rate für diese Gruppe bei über 60 %; nachdem die Texte jedoch sprachlich durch ein KI-Tool überarbeitet worden waren, sank diese Rate deutlich. Dies verdeutlicht, dass Detektionssysteme nicht nur KI-Nutzung erfassen, sondern zugleich normative Vorstellungen von sprachlicher Korrektheit und akademischem Ausdruck reproduzieren.

Wird KI also zur Bewertung studentischer Arbeiten eingesetzt, besteht die Gefahr, dass abweichende Schreibstile, nicht-normative Argumentationsformen oder sprachliche Varianz als Defizite interpretiert werden. Studierende mit nicht-akademischem Hintergrund, Mehrsprachigkeit oder Behinderungen können dadurch strukturell benachteiligt werden.

Das Problem liegt also nicht nur in einzelnen fehlerhaften Ergebnissen, sondern in den Bewertungsmaßstäben, die durch algorithmische Modelle stabilisiert werden. Hochschulen müssen deshalb prüfen, welche Normen in KI-gestützten Verfahren wirksam werden, wer durch diese Normen bevorzugt oder benachteiligt wird und an welchen Stellen menschliche, kontextsensible und diskriminierungskritische Entscheidungen unverzichtbar bleiben.

Zielbilder

Algorithmische Bewertung kritisch reflektieren

- Algorithmische Bewertungen werden an Hochschulen nicht als objektive Entscheidungen verstanden, sondern als normativ vorgeprägte Verfahren kritisch reflektiert.

Kontextsensible Entscheidungen sichern

- Automatisierte Verfahren werden grundsätzlich durch menschliche, kontextsensible und diskriminierungskritische Entscheidungen ergänzt.

Vielfalt wissenschaftlicher Ausdrucksformen anerkennen

- KI-Systeme sollten an Hochschulen nicht zur Standardisierung wissenschaftlichen Ausdrucks eingesetzt, sondern so genutzt werden, dass unterschiedliche sprachliche, kulturelle und fachliche Ausdrucksformen als legitime Formen wissenschaftlicher Kommunikation anerkannt bleiben.

Ebene 3: Die Entwicklungsebene als Gestaltungskontext: Wer entscheidet über KI an Hochschulen?

Eine weitere relevante Problemstelle liegt, wie weiter oben beschrieben, in der häufig wenig diversen Zusammensetzung von Entwickler:innenteams. Ihre Homogenität kann dazu führen, dass bestimmte Sichtweisen, Erfahrungen und Lebensrealitäten ausgeklammert werden und entsprechend nicht in die Entwicklung von KI-Systemen einfließen.

Bildungseinrichtungen entwickeln KI-Systeme in der Regel nicht selbst. Um diese Ebene dennoch auf den akademischen Raum zu übertragen, richtet sich der Blick auf jene Entscheidungs- und Gestaltungskontexte, in denen der Einsatz von KI ausgewählt, ermöglicht oder reguliert wird. Genau hier liegt ein wichtiger Hebel: Institutionen entscheiden mit darüber, welche Tools angeschafft oder genutzt werden dürfen, für welche Zwecke sie eingesetzt werden und welche Systeme im Studien- und Lehralltag empfohlen oder legitimiert werden. Damit verbunden ist auch die Verantwortung, zu prüfen, ob diese Systeme bestehende Bias reproduzieren, verstärken oder bestimmte Gruppen benachteiligen können. Entscheidend ist jedoch, wer diese Entscheidungen trifft – und wessen Perspektiven dabei berücksichtigt werden.

Der Technikphilosoph Andrew Feenberg (2002) beschreibt Technik als doppelte Gestaltungsaufgabe: Zum einen geht es um die technischen Artefakte selbst, zum anderen um ihre soziale Nutzung und institutionelle Einbettung. Technik ist demnach nie nur ein fertiges Werkzeug, sondern wird auch durch die Kontexte geprägt, in denen sie eingesetzt wird. Nicht nur technische Systeme selbst,

sondern auch ihre gesellschaftlichen Anwendungs- und Entscheidungszusammenhänge müssen deshalb partizipativ gestaltet werden.

Auch wenn Hochschulen nicht primär Entwicklungsakteurinnen sind, haben sie die Möglichkeit, Entscheidungen zur Nutzung partizipativ zu gestalten, statt sie hinter „geschlossenen Türen" auf Leitungs- oder IT-Ebene zu treffen und ohne die Perspektiven derjenigen einzubeziehen, die durch KI-Anwendungen potenziell benachteiligt werden könnten.

Zielbilder

Bias-Risiken bei KI-Systemen systematisch prüfen

- Empfohlene oder erlaubte KI-Systeme werden im Hochschulalltag systematisch daraufhin geprüft, ob sie Bias reproduzieren, verstärken oder bestimmte Gruppen benachteiligen können.

KI-Entscheidungen partizipativ gestalten

- Entscheidungen über die Einführung und Nutzung von KI werden partizipativ gestaltet und nicht ausschließlich auf Leitungs- oder IT-Ebene getroffen. Studierende, Gleichstellungsbeauftragte und Diversity-Akteur:innen werden frühzeitig in Auswahl- und Bewertungsprozesse von KI-Systemen einbezogen.

Ebene 4: Nutzung und Anwendung: Institutionalisierte Effekte im Alltag

„Durch die Art und Weise wie wir prompten, können wir einen großen Einfluss auf die Ergebnisse der KI haben. Diesen Einfluss sollten wir nutzen, um unsere Welt vielfältiger, gerechter und inklusiver darzustellen und somit einen Beitrag dazu zu leisten, dass sie es im Laufe der Zeit auch wird. Aktuell schaffen wir die Datengrundlagen für neue Generationen von KI und wir gestalten ein Abbild unserer Welt, das unsere eigene Wahrnehmung der Welt beeinflusst. Wir haben die Macht, unsere Welt gerechter, bunter und vielfältiger zu gestalten – auch mit fairem KI-Prompting."

— Gengler et al. (2024b), S. 26

Während die zuvor genannten Ebenen oft nur begrenzt im direkten Einflussbereich von Hochschulen liegen, ist die Ebene der Nutzung und Anwendung besonders relevant. Bias manifestiert sich hier in der alltäglichen Verwendung von KI-Systemen in Lehre und Lernen. Beispiele reichen von KI-gestützter Text- oder Bilderstellung, die stereotype Rollenbilder reproduzieren, bis hin zu automatisierten Empfehlungssystemen, die bestimmte Gruppen systematisch unterschiedlich adressieren.

Hochschulen gestalten aktiv mit, unter welchen Bedingungen KI diskriminierende Wirkungen entfaltet oder kritisch reflektiert eingesetzt wird. Ein bias-sensibler Umgang mit KI ist folglich nicht nur eine individuelle Kompetenzfrage, sondern eine institutionelle Aufgabe. Hochschulen müssen Rahmenbedingungen für Studierende und Lehrende schaffen, die eine fundierte, kritische und verantwortungsvolle Auseinandersetzung mit diesem Thema ermöglichen und fördern.

Zielbilder

Kritische Reflexionsräume in Studium und Lehre verankern

- Hochschulen schaffen Räume, in denen Studierende, Lehrende und Beschäftigte die sozialen, politischen und diskriminierungsrelevanten Effekte von KI-Anwendungen systematisch reflektieren.

Integration von Bias-Sensibilität in bestehende KI-Leitlinien

- Hochschulen erweitern bestehende Leitlinien zum Einsatz von KI gezielt um bias-sensible Perspektiven, damit Fragen von Diskriminierung, Repräsentation und sozialer Ungleichheit systematisch berücksichtigt werden.

KI-Nutzung mit Gleichstellungs- und Diversitätszielen verknüpfen

- Hochschulen behandeln den Einsatz von KI nicht nur als technische oder effizienzbezogene Frage, sondern verbinden ihn aktiv mit bestehenden Strategien zu Gleichstellung, Diversität und Antidiskriminierung.

Handlungsperspektiven für Hochschulen

Die Zielbilder machen deutlich: Wenn Bias auf mehreren Ebenen entsteht, braucht es auch Interventionen auf mehreren Ebenen. Im Folgenden bündeln wir daher noch einmal, in welchen Handlungsfeldern konkrete Gestaltungsmöglichkeiten liegen.

Governance

Strukturelle Probleme lassen sich nicht allein auf individueller Nutzungsebene lösen. Wenn Hochschulen Bias-Sensibilität und machtkritische Perspektiven im Umgang mit KI stärken wollen, müssen sie dafür institutionelle Rahmenbedingungen schaffen. Dazu braucht es verbindliche Governance-Strukturen, denn allgemeine Empfehlungen zur KI-Nutzung reichen nicht aus. Notwendig sind klare Zuständigkeiten, transparente Entscheidungswege und Verfahren zur Prüfung von KI-Systemen, bevor sie in Studium und Lehre eingesetzt werden. Hochschulen benötigen zudem Leitlinien, die nicht nur technische und datenschutzrechtliche Fragen behandeln, sondern Bias, Diskriminierung, Transparenz und Verantwortung ausdrücklich adressieren.

Fragen Sie sich:

- ☐ Wird Bias in bestehenden KI-Leitlinien substanziell behandelt, oder nur als ein Risikofaktor unter vielen genannt?
- ☐ Nach welchen Kriterien werden KI-Systeme für Studium und Lehre ausgewählt oder empfohlen?
- ☐ Wie werden Perspektiven potenziell betroffener Gruppen in Entscheidungsprozesse einbezogen?
- ☐ Gibt es Möglichkeiten, problematische KI-Ergebnisse oder diskriminierende Effekte zu melden?
- ☐ Sind Fragen von Bias und Verantwortung institutionell verankert oder vom Engagement einzelner Personen abhängig?

Lehre & Qualifizierung

Bias-Sensibilität muss Teil akademischer Bildung werden. Studierende und Lehrende sollten verstehen, wie Daten, Entwicklungsprozesse und Nutzungskontexte KI-Ergebnisse beeinflussen und welche gesellschaftlichen Folgen daraus entstehen können. Hochschulen sollten KI-Kompetenz deshalb nicht auf die reine Anwendung von Tools reduzieren, sondern um Fragen nach Fairness, Machtkritik und epistemischer Verantwortung erweitern. Dazu gehört auch die Reflexion darüber, wie wir mit KI-Systemen interagieren. Die Art und Weise, wie Prompts formuliert werden, beeinflusst wesentlich, welche Antworten entstehen, welche Perspektiven sichtbar werden und welche Darstellungen sich verfestigen. Hochschulen sollten daher Räume schaffen, in denen Bias-Sensibilität als grundlegende KI-Kompetenz vermittelt, eingeübt und kritisch reflektiert wird.

Fragen Sie sich:

- ☐ Gibt es regelmäßige Fortbildungen für Hochschulpersonal zu KI-Bias?
- ☐ Wird KI-Kompetenz nur als Anwendungskompetenz verstanden, – oder auch als Reflexions- und Urteilskompetenz?
- ☐ Werden Studierende systematisch für Verzerrungen in KI-Systemen sensibilisiert?
- ☐ Ist dies curricular verankert, oder abhängig von einzelnen Lehrenden?
- ☐ Wird KI als technisches Thema behandelt oder auch als gesellschaftliche Gestaltungsfrage?
- ☐ Haben Studierende Zugang zu offenen Diskussionen, Vorträgen oder Veranstaltungen zu KI und Bias?

Partizipation

Entscheidungen über KI sollten nicht ausschließlich von Hochschulleitungen oder IT-Abteilungen getroffen werden. Studierende, Gleichstellungsbeauftragte, Diversity-Akteur:innen und marginalisierte Gruppen müssen systematisch in die Gestaltung von KI-Strategien eingebunden werden. Die Diversität, die in Entwickler:innenteams häufig fehlt, können Hochschulen zumindest auf der Ebene ihrer eigenen Entscheidungs- und Beteiligungsprozesse stärker berücksichtigen durch frühzeitige Einbindung bei der Auswahl von Tools, bei der Entwicklung von Leitlinien und bei der Bewertung möglicher Risiken.

Fragen Sie sich:

- ☐ Wer entscheidet an unserer Hochschule über die Bereitstellung und Nutzung von KI-Systemen?
- ☐ Sind Studierende, Lehrende, Gleichstellungsbeauftragte und Diversity-Akteur:innen in diese Entscheidungen eingebunden?
- ☐ Gibt es ein zentrales Gremium, das sich mit KI-gestützten Systemen und potenziellen Verzerrungen befasst? Wer sitzt in diesem Gremium?
- ☐ Gibt es ein offenes Feedback-System, über das Studierende und Mitarbeitende Bedenken zu KI-Systemen melden können?
- ☐ Werden Hochschulangehörige regelmäßig über neue Erkenntnisse und Debatten zu KI-Bias informiert?

Eine kritische und reflektierte Nutzung von KI an Hochschulen stellt eine zentrale hochschulische Verantwortung dar. Nur wenn KI konsequent als soziotechnische Praxis begriffen und machtkritisch gerahmt wird, können Hochschulen ihrer Rolle als epistemische Institution gerecht werden. Andernfalls laufen sie Gefahr, selbst zu Akteur:innen der Reproduktion epistemischer und sozialer Ungleichheiten zu werden, anstatt Räume für deren Reflexion und Transformation zu eröffnen. Welche der beiden zu Beginn des Kapitels genannten Hochschulen im Jahr 2035 Realität wird, entscheidet sich daher nicht erst in der Zukunft, sondern durch die institutionellen Entscheidungen, die heute getroffen werden.

Fazit: Warum Hochschulen JETZT handeln müssen

> „Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy"
>
> — Prabhu & Birhane (2020), S. 6

KI-Systeme entstehen nicht außerhalb gesellschaftlicher Verhältnisse. Sie werden mit Daten, Kategorien und Bewertungslogiken entwickelt, die bestehende Machtverhältnisse, Ausschlüsse und Ungleichheiten enthalten. Zu erwarten, dass solche Systeme in hochschulischen Kontexten neutral oder diskriminierungsfrei wirken, ohne ihre Voraussetzungen kritisch zu prüfen, wäre daher naiv.

Hochschulen prägen, wie zukünftige Generationen Wissen erzeugen, bewerten und nutzen. Wenn KI zunehmend Teil wissenschaftlicher Praxis wird, entscheidet sich hier zugleich, ob bestehende epistemische Ungleichheiten fortgeschrieben oder bewusst hinterfragt werden. Die Frage ist daher nicht mehr, ob KI Teil der Hochschule wird, sondern unter welchen Bedingungen.

Dafür reicht es nicht aus, einzelne Leitlinien oder Nutzungsempfehlungen zu formulieren. Notwendig sind Strukturen, die kritische KI-Kompetenz dauerhaft in Studium, Lehre, Forschung und Verwaltung verankern. Dazu gehören Räume für Reflexion, interdisziplinäre Perspektiven, partizipative Entscheidungsprozesse und die Bereitschaft, technologische Entwicklungen nicht nur effizient, sondern auch normativ zu bewerten.

Hochschulen stehen damit an einem entscheidenden Punkt: Sie können KI lediglich als Instrument zur Optimierung bestehender Prozesse begreifen oder sie zum Anlass nehmen, grundlegende Fragen neu zu verhandeln. Wer diese Entwicklung nicht aktiv mitgestaltet, überlässt ihre Richtung anderen. Dieses Paper plädiert daher für eine Hochschulpraxis, die KI nicht nur nutzt, sondern ihre Bedingungen und Folgen kritisch mitgestaltet.

Literaturverzeichnis

Asaf, S. (2026, 27. April). Looming EU AI act could force universities to 'change everything'. Times Higher Education. Abgerufen am 4. Mai 2026 von https://www.timeshighereducation.com/news/looming-eu-ai-act-could-force-universities-change-everything

Adam, A. (1995). Artificial intelligence and women's knowledge: What can feminist epistemologies tell us? Women's Studies International Forum, 18(4), 407–415. https://doi.org/10.1016/0277-5395(95)80032-K

Becker, S., Leifeld, J., Lüthi, R., Tobor, J., & Westermann, A. (2026). Leitlinien-Check 2026: Ein Update zu generativer KI an Hochschulen. Berlin: Hochschulforum Digitalisierung.

Budde, J., Tobor, J., & Friedrich, J. (2024). Künstliche Intelligenz. Wo stehen die deutschen Hochschulen? Berlin: Hochschulforum Digitalisierung.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html.

Dastin, J. (2018, 10. Oktober). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Feenberg, A. (2002). Transforming technology: A critical theory revisited. Oxford University Press.

Fricker, M. (2023). Epistemische Ungerechtigkeit: Macht und die Ethik des Wissens. Suhrkamp.

Gengler, E. J. (2023). Feminism – For more equity in AI [Video]. YouTube. https://www.youtube.com/watch?v=CxcCwvut50A

Gengler, E. J. (2024). Sexism, racism, and classism: Social biases in text-to-image generative AI in the context of power, success, and beauty. Wirtschaftsinformatik 2024 Proceedings, 48. https://aisel.aisnet.org/wi2024/48

Gengler, E. J., Hagerer, I., & Gales, A. (2024a). Diversity bias in artificial intelligence. In: The Routledge Handbook of AI Ethics and Society (S. 229–240). Routledge. https://doi.org/10.4324/9781003383741-23

Gengler, E. J., Kraus, A., & Bodrožić-Brnić, K. (2024b). Faires KI-Prompting – Ein Leitfaden für Unternehmen. BSP Business and Law School – Hochschule für Management und Recht. (1 – 28). https://www.digitalzentrum-zukunftskultur.de/material/faires-ki-prompting-13136/

Gengler, E. J., Wedel, M., Wudel, A., & Laumer, S. (2025). Power imbalances in society and AI: On the need to expand the feminist approach. In: D. Beverungen, C. Lehrer, & M. Trier (Hrsg.), Conceptualizing digital responsibility for the information age. Lecture Notes in Information Systems and Organisation, Bd. 74. Springer. https://doi.org/10.1007/978-3-031-80119-8_6

Guilbeault, D. R., Delecourt, S., Hull, T., Desikan, B. S., Chu, M. & Nadler, E. (2024) Online images amplify gender bias. Nature, 626(8001), 1049–1055.

He, J. (2025). Who gets cited? Gender- and majority-bias in LLM-driven reference selection. ResearchGate. https://www.researchgate.net/publication/394322251_Who_Gets_Cited_Gender-_and_Majority-Bias_in_LLM-Driven_Reference_Selection

Kruspe, A., & Stillman, M. (2024). Saxony-Anhalt is the worst: Bias towards German federal states in Large Language Models. https://doi.org/10.1007/978-3-031-70893-0_12

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, *4*(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779

Morozov, E. (2014). To save everything, click here: The folly of technological solutionism. PublicAffairs.

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.

Pal, S., Lazzaroni, R. M., & Mendoza, P. (2024, 10. Oktober). AI's missing link: The gender gap in the talent pool [Data brief]. Interface. https://www.interface-eu.org/publications/ai-gender-gap

Prabhu, V. U., & Birhane, A. (2020). Large datasets: A Pyrrhic win for computer vision? arXiv. https://arxiv.org/abs/2006.16923

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). No classification without representation: Assessing geodiversity issues in open data sets for the developing world. arXiv. https://arxiv.org/abs/1711.08536

Sorokovikova, A., Chizhov, P., Eremenko, I., & Yamshchikov, I. P. (2025). Surface fairness, deep bias: A comparative study of bias in language models (arXiv Paper No. 2506.10491). arXiv. https://doi.org/10.48550/arXiv.2506.10491

Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., & Peng, N. (2023). "Kelly is a warm person, Joseph is a role model": Gender biases in LLM-generated reference letters. In: Findings of the Association for Computational Linguistics: EMNLP 2023, (S. 3730–3748). Singapore. Association for Computational Linguistics. https://aclanthology.org/2023.findings-emnlp.243/

Young, J. (2019). Why we need to design feminist AI [Video]. YouTube. https://www.youtube.com/watch?v=E-O3LaSEcVw

Impressum

Diskussionspapiere des HFD spiegeln die Meinung der jeweiligen Autor:innen wider.
Das HFD macht sich die in diesem Papier getätigten Aussagen daher nicht zu Eigen.

Dieses Werk ist unter einer Creative Commons Lizenz vom Typ Namensnennung - Weitergabe unter gleichen Bedingungen 4.0 International zugänglich. Um eine Kopie dieser Lizenz einzusehen, konsultieren Sie http://creativecommons.org/licenses/by-sa/4.0/. Von dieser Lizenz ausgenommen sind Organisationslogos sowie falls gekennzeichnet einzelne Bilder und Visualisierungen.

ISSN (Online) 2365-7081; 12. Jahrgang • DOI: 10.5281/zenodo.20606315

Zitierhinweis
Becker, S., Leifeld, J. (2026). On how to bring up baby robots – Ein Plädoyer für einen bias-sensiblen und machtkritischen Umgang mit generativer KI an Hochschulen. Diskussionspapier Nr. 39. Berlin: Hochschulforum Digitalisierung.

Herausgeber
Geschäftsstelle Hochschulforum Digitalisierung beim Stifterverband für die Deutsche Wissenschaft e.V.
Hauptstadtbüro • Pariser Platz 6 • 10117 Berlin • T 030 322982-520
info@hochschulforumdigitalisierung.de

Layout
Satz: Katja Engelhaus, Marieke Einheuser
Vorlage: TAU GmbH • Köpenicker Straße 154a • 10997 Berlin

Das Hochschulforum Digitalisierung ist eine gemeinsame Initiative des Stifterverbandes, des CHE Centrums für Hochschulentwicklung und der Hochschulrektorenkonferenz. Gefördert wird es vom Bundesministerium für Forschung, Technologie und Raumfahrt.

www.hochschulforumdigitalisierung.de