



KOMPASS

KÜNSTLICHE INTELLIGENZ UND VOREINGENOMMENHEIT (BIAS)

Themenblatt

KI und Voreingenommenheit (Bias)

In diesem Themenblatt wird Künstliche Intelligenz und ihre mögliche Voreingenommenheit (Bias) erläutert. Das Themenblatt richtet sich an Lehrkräfte aus dem *Enseignement Fondamental* und *Secondaire* und liefert Hintergrundwissen, um das Thema im eigenen Unterricht aufzugreifen.

Was ist Voreingenommenheit?

Voreingenommenheit bedeutet an sich, eine voreingenommene oder vorgefasste (positive oder negative) Meinung über Menschen, Gruppen oder Situationen zu haben, ohne dass man diese wirklich kennt. Man spricht auch von Vorurteil, da man urteilt, bevor man etwas kennt.

Voreingenommenheit (Bias) in der KI bezeichnet verzerrte oder unfaire Präferenzen oder Vorurteile in KI-Outputs, die unter anderem durch fehlerhafte Daten, Algorithmen oder menschliche Vorurteile verursacht werden können.

Welche Verzerrungen beeinflussen KI-Modelle?

Verzerrung bedeutet, dass etwas von seinem ursprünglichen Zustand verdreht oder verändert wurde bzw. es Abweichungen zur Realität geben kann.

Da KI-Systeme von Menschen entwickelt werden, fließen deren unbewusste (*unconscious bias*) oder bewusste Vorurteile oft in die Modelle¹ ein. Dies geschieht etwa durch Entscheidungen bei der Auswahl und Aufbereitung von Daten, der Definition relevanter Merkmale oder der Modellanpassung. Selbst gut gemeinte Ansätze können unbeabsichtigt bestehende Perspektiven und Annahmen reproduzieren und so die Voreingenommenheit verstärken. Das führt dazu, dass sowohl in Daten als auch in Algorithmen menschliche Vorteile vorzufinden sind. Neben dem **menschlichen Vorurteil** (*human bias*) gibt es weitere Verzerrungen, die KI-Modelle und ihre Outputs beeinflussen können. Im Folgenden eine kurze Erläuterung der wichtigsten:

Datenverzerrung (*data bias*)

Eine Datenverzerrung, auch Verzerrung von Trainingsdaten genannt, kann bereits bei der Erfassung und Aufbereitung der Daten auftreten. Sind Datensätze nicht repräsentativ, liefern darauf trainierte Modelle verzerrte Ergebnisse.

Beispiel: Ein Spracherkennungsmodell, das nur mit männlichen Stimmen trainiert wurde, erkennt weibliche Stimmen schlechter.

Algorithmische Verzerrung (*algorithmic bias*)

¹ In der KI bezeichnen Modelle konstruierte mathematisch-statistische Strukturen, die auf Basis großer Datenmengen trainiert werden, um bestimmte Aufgaben wie beispielsweise Texterkennung, Bilderkennung oder Vorhersagen zu erfüllen. Sie bilden das Zentralelement eines KI-Systems und beeinflussen dessen Verhalten und Ergebnisse maßgeblich.

Eine algorithmische Verzerrung entsteht, wenn das System bestimmte Gruppen systematisch bevorzugt oder benachteiligt. Dies passiert beispielsweise, wenn ein System aus alten Mustern oder Ungleichheiten lernt, diese wiederholt und manchmal auch verstärkt. So kann es sein, dass einzelne Faktoren stärker gewichtet werden als andere, wodurch bestehende Ungleichheiten verstärkt werden.

Besonders kritisch ist dies in sensiblen Bereichen wie Personalwesen, Kreditvergabe oder Strafverfolgung, in denen verzerrte Vorhersagen Diskriminierung begünstigen können.

Beispiel: Ein Algorithmus zur Bewertung der Kreditwürdigkeit stuft Personen aus bestimmten Postleitzahlgebieten systematisch schlechter ein, weil diese Gebiete in den Trainingsdaten häufiger mit Zahlungsausfällen verbunden waren – unabhängig von der tatsächlichen Kreditwürdigkeit der jeweiligen Person.

Warum ist die Voreingenommenheit in KI-Systemen ein Problem?

Voreingenommenheit in KI-Systemen zeigt, dass Technik nicht losgelöst von der Gesellschaft funktioniert. Denn auch KI-Entwickler legen – bewusst oder unbewusst – Kriterien fest, etwa auf der Grundlage politischer oder wirtschaftlicher Interessen. Dies zeigt, dass KI-Systeme, wenn sie nicht angemessen kontrolliert und reflektiert eingesetzt werden, Diskriminierung fördern, Vertrauen untergraben und das Potenzial der Technologie erheblich begrenzen können.

Ein voreingenommenes KI-System kann in zahlreichen gesellschaftlichen Bereichen diskriminierend wirken, etwa bei der Bewertung der Kreditwürdigkeit, in der Spracherkennung (von Dialekten), im Gesundheitswesen, in der Strafverfolgung sowie im Bildungsbereich.

Beispiel aus der Bildung: Eine Universität nutzt Bewertungs- und Zulassungsalgorithmen, um über die Auswahl der angehenden Studierenden zu entscheiden. Die KI prognostiziert den Kandidatinnen und Kandidaten mehr Erfolg, wenn sie aus finanzstarken und gut ausgestatteten Schulen kommen. Schülerinnen und Schüler aus eher ressourcenarmen Schulen werden hingegen in diesem Prozess benachteiligt. Das kann zu weiteren sozialen Ungleichheiten zwischen den sozialen Schichten führen, da marginalisierte Gruppen eher von der Voreingenommenheit des KI-Systems betroffen sind.

Beispiel: Ein KI-System soll Bewerbungen vorsortieren und wurde mit älteren Daten trainiert, in denen vor allem Männer für technische Berufe eingestellt wurden. Deshalb bevorzugt die KI nun automatisch männlich klingende Namen und stuft Bewerbungen von Frauen schlechter ein – unabhängig von ihrer tatsächlichen Qualifikation. Dadurch wird bestehende Diskriminierung nicht abgebaut, sondern verstärkt, und Betroffene werden ungerecht behandelt.

Faire, ausgewogene KI-Systeme haben das Potenzial, Bereiche wie Justiz, Gesundheit und Bildung grundlegend zu verbessern – vorausgesetzt, der Mensch setzt sich kritisch mit ihren Schwächen auseinander. Wichtig ist, dass der Mensch vor der Technologie steht.

Können KI-Systeme neutral sein?

Kann eine KI jemals frei von Voreingenommenheit sein, oder wird sie stets menschliche Schwächen spiegeln? Absolute Neutralität ist in der KI wahrscheinlich ein unrealistisches Ziel, da sie ein menschliches Produkt ist und von menschlichen Zielsetzungen und Denkweisen geprägt wird. Durch technische Korrekturen und ethische Aufsicht kann KI jedoch weniger voreingenommen gestaltet werden. Im Folgenden einige Beispiele dieser Methoden zur Korrektur:

- Breite und repräsentative Datensätze: Entwicklerinnen und Entwickler einer KI sollten immer darauf achten, dass die gesammelten Datensätze möglichst vielfältig und repräsentativ sind, um ein ausgewogeneres Bild einer Situation zu erhalten.
- Testverfahren für Fairness: Es gibt Tools, mit denen sich KI-Systeme vor ihrem Einsatz auf Fairness prüfen lassen, z. B. [AI Sandbox](#) (Luxembourg Institute for Science and Technology, LIST).
- *Explainable AI* (XAI): „Erklärbare KI“-Techniken sorgen dafür, dass jede Entscheidung im *Machine-Learning*-Prozess² nachvollzogen und erklärt werden kann.

Ohne diese Methoden agieren viele KI-Modelle als „Blackbox“, deren interne Abläufe selbst für Entwicklerinnen und Entwickler oft undurchsichtig sind. Dadurch wird die Überprüfung der Ergebnisse erschwert und es gehen Kontrolle sowie Verantwortbarkeit verloren. Mit XAI erhalten Unternehmen Einblick in die Entscheidungsgrundlagen der KI und können Fehler schneller erkennen und beheben.

Ressourcen für den Unterricht

Es gibt einige Möglichkeiten das Thema *KI und Bias* im Unterricht anzusprechen. Ein paar Beispiele finden Sie hier:

- KI Kompass – didaktische Materialien: <https://ki-kompass.lu/praxis-materialien/>
- Die Unterrichtsidee von *Coding for tomorrow* widmet sich dem Thema Bias in der KI aufzudecken. Sie richtet sich eher an Schüler und Schülerinnen ab der 7. Klasse: <https://edulink.lu/ke2a>
- Die *Académie de Nice* gibt in ihrer Unterrichtsidee dem Bias in KI-generierten Bildern. Sie richtet sich an Schülerinnen und Schüler ab ca. 11 Jahren: <https://edulink.lu/jal6>
- Der *KI Campus* gibt in dieser Unterrichtsidee Anreize, wie Vorurteile und Stereotypen in KI-Systemen aufgedeckt werden können: <https://edulink.lu/y8gu>
- Auf *School AI* werden Ideen aufgezeigt, wie KI im Unterricht des *Enseignement Fondamental* nähergebracht werden kann: <https://edulink.lu/6pyj>

² *Machine Learning* (ML, „maschinelles Lernen“) ist ein Teilbereich der Künstlichen Intelligenz. Dabei lernen Computersysteme aus Beispieldaten, erkennen Muster und treffen auf dieser Grundlage Entscheidungen, ohne dafür explizit programmiert worden zu sein.

- Das französische Kulturministerium hat das Vergleichstool *Compar:IA* erstellt, in dem Tools und Datensätze verglichen werden können, um so Bewusstsein über KI zu schaffen: <https://edulink.lu/s0i5>

Hinweise: Lehrkräfte sollten die Ressourcen daraufhin prüfen, ob sie zur jeweiligen Klasse und Kontext passen und die Nutzung gegebenenfalls anpassen.

Zur Umsetzung der dargestellten Unterrichtsideen sollten bevorzugt die Tools genutzt werden, die über die Webseite [KI Kompass](#) zur Verfügung gestellt werden.

Referenzen

BEE SECURE (2025): *Algorithmische Voreingenommenheit – Wenn Technik unsere Vorurteile lernt*. Aufruf über: <https://www.bee-secure.lu/de/news/algorithmische-voreingenommenheit-wenn-technik-unsere-vorurteile-lernt/> (letzter Zugriff: 27.01.2026).

Chapman University: *Bias in AI*. Aufrufbar über: <https://www.chapman.edu/ai/bias-in-ai.aspx> (letzter Zugriff 28.01.2026)

IBM: *KI-Verzerrungen anhand von Beispielen aus der realen Welt beleuchten*. Aufrufbar über: <https://www.ibm.com/de-de/think/topics/shedding-light-on-ai-bias-with-real-world-examples> (letzter Zugriff 17.01.2026)

SCRIPT (2026): *Strategischer Rahmen zum Einsatz von Künstlicher Intelligenz in der Schule* (Version vom 3. Februar 2026). Aufrufbar über: <https://ki-kompass.lu> (letzter Zugriff: 27.01.2026)

Luxembourg Institute of Science and Technology (LIST): *LIST AI SANDBOX – LLM Observatory*. Aufrufbar über: <https://ai-sandbox.list.lu/llm-leaderboard/> (letzter Zugriff: 27.01.2026)

University of Kansas – Center for Teacher Excellence: *Helping students understand the biases in generative AI*. Aufruf über: <https://cte.ku.edu/addressing-bias-ai> (letzter Zugriff: 28.01.2026)