

## M5-S Verzerrungen



Bei der Herstellung dieses Bildes wurde am 03.04.2024 DALL.E eingesetzt.

2016 fand Beauty.AI statt, der erste internationale Schönheitswettbewerb, bei dem ein KI-System die Jury bildete. Das Ergebnis löste Kontroversen aus, da fast alle Gewinnerinnen und Gewinner weiß waren. Lag es daran, dass die Teilnehmenden helle Hautfarben hatten? Nein, es gab auch eine große Anzahl von Teilnehmenden, die nicht als weiß gelesen wurden. Die Erklärung dafür, dass der Algorithmus helle Hautfarben bevorzugte, ist, dass er mit großen Fotodatenbanken trainiert worden war, die Minderheiten nur unzureichend repräsentierten. Dieser Vorfall zeigt sehr deutlich, dass Algorithmen existierende Vorurteile reproduzieren und verstärken können, was zu verzerrten und häufig anstößigen Ergebnissen sowie der Diskriminierung von Minderheiten führen kann (ZEIT ONLINE, 2016).

### Arten von Verzerrungen

Verzerrungen treten bei allen möglichen Datenerhebungen auf, wie bei physikalischen Experimenten, bei Umfragen, bei Wettervorhersagen und vielem mehr. Die Entstehung unzuverlässiger Daten kann dabei auf verschiedene Gründe zurückzuführen sein (Schmid et al, 2024, Abschnitt 4.1).

#### Zufällige Verzerrungen

Zufällige Verzerrungen sind Abweichungen der Messwerte vom wahren Wert, die zufällig auftreten und alle Arten von Messungen betreffen, z. B. Temperaturmessungen bei Wettervorhersagen oder Gewichtsmessungen. So kann es beispielsweise sein, dass es regnet, obwohl sonniges Wetter vorhergesagt wurde.

Die zufälligen Verzerrungen können die Verlässlichkeit von Daten wie Wettervorhersagen oder Wahlprognosen einschränken, aber ihr Einfluss kann durch das Sammeln und Vergleichen mehrerer Datenquellen und das Ermitteln von Mittelwerten begrenzt werden.



Bei der Herstellung dieses Bildes wurde am 11.04.2024 DALL.E eingesetzt.

### Bewusste Verzerrungen

Bewusste Verzerrungen sind absichtliche Manipulationen von Informationen oder Daten, die häufig zur Beeinflussung von Meinungen oder zur Gewinnmaximierung eingesetzt werden, wie z. B. manipulierte Bewertungen von Produkten oder Dienstleistungen. So könnte beispielsweise eine heruntergekommene Imbissbude eine 5-Sterne-Bewertung haben.

Diese bewussten Verzerrungen können der Gesellschaft schaden, indem sie Fehlinformationen verbreiten und Entscheidungen beeinflussen. Schutz davor bieten die Überprüfung von Quellen, der Vergleich mit vertrauenswürdigen Informationen und das kritische Hinterfragen der Intention und Emotionalität von Informationen.



Bei der Herstellung dieses Bildes wurde am 11.04.2024 DALL.E eingesetzt.

### Systematische Verzerrungen (Bias)

Systematische Verzerrungen, auch Bias genannt, treten auf, wenn Daten, die für Entscheidungs- oder Trainingsprozesse verwendet werden, eine unausgewogene oder einseitige Perspektive aufweisen.

Dies führt oft zu verzerrten Ergebnissen oder Urteilen. Solche Verzerrungen beruhen häufig auf menschlichen Vorurteilen oder unvollständigen Datensätzen und können sowohl im menschlichen Denken als auch in KI-Anwendungen auftreten.



Bei der Herstellung dieses Bildes wurde am 11.04.2024 DALL.E eingesetzt.

### **Diskriminierung durch KI-Systeme**

Diskriminierung durch KI-Systeme entsteht häufig unbeabsichtigt durch eine Kombination aus gesellschaftlichen Stereotypen und technischen Entscheidungen, wie der Auswahl von Zielvariablen, Trainingsdaten und Analysemethoden (Deutscher Ethikrat, 2023, Abschnitt 10.9).

Beim maschinellen Lernen werden zum Beispiel Analogien aus Texten gezogen. Aus Wortpaaren wie „Deutschland - Berlin“ oder „Rom - Italien“ findet die Maschine eine sehr harmlose Analogie wie Berlin ist zu Deutschland wie Rom zu Italien, ohne die Beziehung Hauptstadt zu kennen. Aus männlichen und weiblichen Berufsbezeichnungen könnte sich dann folgende Analogie ergeben: Mann zum Arzt wie Frau zur Ärztin. Falls der Trainingsdatensatz veraltet ist, könnte es auch heißen: Mann zum Arzt wie Frau zur Krankenschwester. Das wäre dann ein Stereotyp (Zweig, 2019, S. 207f.).

Wenn ein Algorithmus diskriminiert, bedeutet das in der Regel, dass er Unterschiede zwischen Gruppen erkennt und diese dann anhand bestimmter Merkmale voneinander trennt. Dies wird insbesondere dann problematisch, wenn es sich um grundgesetzlich geschützte Merkmale wie Geschlecht, Religion, Alter oder Gesundheitsinformationen handelt. Kritisch wird es, wenn eine solche Unterscheidung zu Benachteiligungen in wichtigen Bereichen wie Arbeit oder Bildung führt.

## Ursachen für Verzerrungen

Ein anschauliches Beispiel für die Bedeutung der *Qualität von Input-Daten* ist der Fall von Tay, dem KI-Chatbot von Microsoft. Tay wurde entwickelt, um aus Gesprächen auf Twitter zu lernen. Doch innerhalb eines Tages wurde Tay durch negative und extremistische Tweets so stark beeinflusst, dass Microsoft ihn vom Netz nehmen musste. Konfrontiert mit frauenfeindlichen, rassistischen und politisch extremen Äußerungen, begann der Bot, diese Inhalte zu wiederholen. In der Informatik nennt man dies „garbage in, garbage out“ - das bedeutet, dass ein Computerprogramm schlechte Informationen reproduziert, wenn es sie erhält (Vincent, 2016).

Wie bei dem oben erwähnten Beispiel zum Schönheitswettbewerb Beauty.AI zu sehen ist, ist die hinreichende *Vollständigkeit der Daten* in der Trainingsphase des KI-Systems relevant. In dem Beispiel gab es nicht genügend Daten von nicht weißen Menschen, um eine umfassende und faire Attraktivitätsskala zu erstellen. Aus diesem Grund hat das KI-System nur Gewinnerinnen und Gewinner mit weißer Hautfarbe ausgewählt. Ein ähnlicher Effekt zeigte sich bei Bilderkennungssystemen zur Unterscheidung von Melanomen und Leberflecken. Diese funktionierten bei heller Haut zuverlässiger, da es mehr Bilder von Melanomen auf heller als auf dunklerer Haut gibt (Zweig, 2019, S. 190ff.).

Eine weitere mögliche Ursache für Bias ist die *Aktualität der Input-Daten*, mit denen die Maschine trainiert wird.

Darüber hinaus ist die *Vollständigkeit der erhobenen Merkmale* relevant: Das Fehlen sensibler Daten, wie beispielsweise das Merkmal Geschlecht, kann mitunter zu Benachteiligungen führen. So unterstützen in den USA algorithmische Systeme die Entscheidungsfindung, ob Straftäterinnen oder -täter rückfällig werden, oder nicht. Was wäre, wenn Frauen ein anderes Rückfallverhalten hätten als Männer, aber genau dieses sensible Merkmal nicht berücksichtigt würde? Dann würden einige Männer oder Frauen zu Unrecht länger im Gefängnis bleiben (Zweig, 2019, S. 215ff.).

In der Anwendungsphase eines KI-Systems, wie beispielsweise eines neuronalen Netzes, kann dieses durch *manipulierte Daten* gezielt dazu gebracht werden, falsche Entscheidungen zu treffen. Dies ist bei sogenannten Adversarial Attacks der Fall. Ein Beispiel für eine Adversarial Attack ist eine Brille, die mit einem speziellen Adversarial Muster bedruckt ist. In einem Experiment gelang es Forschenden, Gesichtserkennungssysteme mithilfe solcher modifizierter Accessoires derart zu manipulieren, dass ein Mann mit dieser speziellen Brille als Marilyn Monroe identifiziert wurde (Sharif et al, 2016).



Bei der Herstellung dieses Bildes wurde am 27.01.2025 DALL.E eingesetzt.

Bedeutsam sind darüber hinaus *menschliche Voreingenommenheiten* im Kontext erforderlicher menschlicher Entscheidungen, z. B. bei der Auswahl der Trainingsdaten.

## Nachweis von Diskriminierung



Bei der Herstellung dieses Bildes wurde am 18.04.2024 ideogram eingesetzt.

Manchmal ist für Betroffene gar nicht erkennbar, dass sie diskriminiert werden. In einer Untersuchung dazu schaltete AlgorithmWatch mehrere Stellenanzeigen auf Facebook, unter anderem eine Anzeige für eine Stelle als LKW-Fahrerin. Diese Anzeige wurde deutlich häufiger Männern als Frauen angezeigt. Eine Frau, die gerne LKW-Fahrerin werden möchte, hätte diese Anzeige möglicherweise nicht gesehen, da die Werbeanzeigen automatisiert an die vermeintlichen Zielgruppen ausgespielt werden. Sie ist sich damit der Diskriminierung gar nicht bewusst (Kayser-Bril, 2020).

Heute werden in immer mehr Bereichen Maschinen in Entscheidungen einbezogen. Eine zentrale Herausforderung ist dabei zu erkennen, ob die Maschine Menschen diskriminiert.

Dazu kann beispielsweise zum einen der zugrundeliegende Code untersucht werden - vorausgesetzt dieser ist zugänglich, was aufgrund von Patentrechten nicht immer der Fall ist. Diese Untersuchung ist allerdings sehr aufwändig und nur von Expertinnen oder Experten durchführbar. Handelt es sich um selbstlernende Maschinen, dann ist der Prozess der Entscheidungsfindung überhaupt nicht nachvollziehbar. Hier kann das System nur mit der sogenannten Black-Box-Methode analysiert werden. Dabei testet man das zu prüfende System mit vielen verschiedenen Testaccounts und Testanfragen, ohne auf den zugrundeliegenden Algorithmus zugreifen zu müssen. (GI, 2018).

In folgendem Studienbeispiel hat eine externe Bewertungsagentur die Übersetzungsqualität von zwei Übersetzungssystemen untersucht und bewertet, ob diese KI-Dienste unvoreingenommen sind:

Der Satz „He is a nurse. She is an optician“ wurde in mehrere Zielsprachen übersetzt und dann wieder ins Englische zurückübersetzt. Dabei wurde die Rückübersetzung mit dem ursprünglichen Satz verglichen. In zwei Sprachen ging die ursprüngliche Geschlechtsverteilung verloren - die Ausgabe war dann beispielsweise „That nurse. It is an optician.“ In einer anderen Sprache wurde die Geschlechtsverteilung sogar vertauscht: „She is a nurse. He is an optician.“ Der kulturelle Umstand, dass in den betreffenden Ländern Optiker ein typisch männlicher Beruf ist, findet in der Sprache seinen Ausdruck und dadurch auch in der Rückübersetzung (Srivastava, Rossi, 2018).



Bei der Herstellung dieses Bildes wurde am 03.04.2024 DALL.E eingesetzt.

## Quellen:

- Deutscher Ethikrat. (20.3.2023). *Stellungnahme: Mensch und Maschine - Herausforderungen durch Künstliche Intelligenz*. Deutscher Ethikrat. Verfügbar unter: <https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf>, abgerufen am 23.03.2026
- Gesellschaft für Informatik (GI). (10/2018). *Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren*. Studien und Gutachten im Auftrag des Sachverständigenrats für Verbraucherfragen. Berlin: Sachverständigenrat für Verbraucherfragen. Verfügbar unter: [https://gi.de/fileadmin/GI/Allgemein/PDF/GI\\_Studie\\_Algorithmenregulierung.pdf](https://gi.de/fileadmin/GI/Allgemein/PDF/GI_Studie_Algorithmenregulierung.pdf), abgerufen am 23.03.2026
- Kayser-Bril, N. (18.10.2020). *Automatisierte Diskriminierung: Facebook verwendet grobe Stereotypen, um die Anzeigenschaltung zu optimieren*. AlgorithmWatch. Verfügbar unter: <https://algorithmwatch.org/de/automatisierte-diskriminierung-facebook-verwendet-grobe-stereotypen-um-die-anzeigenschaltung-zu-optimieren/>, abgerufen am 23.03.2026
- Schmid, U., Szczepaniak, R. & Gärtig-Daugs, A. (2024). Data Literacy für die Grundschule. KI-Campus. Verfügbar unter: <https://ki-campus.org/courses/dlgrundschule>
- Sharif, M., Bhagavatula, S., Reiter, M. & Bauer, L. (2016). Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. *CS '16: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 1528-1540.
- Srivastava, B. & Rossi, F. (3.2.2018). *Towards Composable Bias Rating of AI Services*. AIES '18. Verfügbar unter: <https://dl.acm.org/doi/pdf/10.1145/3278721.3278744>, abgerufen am 23.03.2026
- Vincent, J. (24.3.2016). *Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day*. The Verge. Verfügbar unter: <https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>, abgerufen am 23.03.2026
- ZEIT ONLINE. (9.9.2016). *Objektiv rassistisch*. ZEIT ONLINE. Verfügbar unter: <https://www.zeit.de/digital/2016-09/beauty-ai-kuenstliche-intelligenz-schoenheitswettbewerb-rassismus-algorithmen>, abgerufen am 23.03.2026
- Zweig, K. (2019). Ein Algorithmus hat kein Taktgefühl. Wo künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können, Heyne.

### Weiterführende und vertiefende Informationen:

- AK Bildung in der digital vernetzten Welt, in: klickpunkt.schule (2024). KI | Verzerrungen. Verfügbar unter: <https://klickpunktschule.bycs.de/beitrag/ki-verzerrungen>
- AK Bildung in der digital vernetzten Welt, in: klickpunkt.schule (2025). KI | Adversarial Attacks. Verfügbar unter: <https://klickpunktschule.bycs.de/beitrag/ki-adversarial-attacks>