

Wie KI-Bilder entstehen — und warum sie nie neutral sind

Informatik · Sekundarstufe I/II (14–18 Jahre) · 90 Minuten, auf 45 kürzbar

Information für Lehrer

Steckbrief

Kategorie	Angabe
Fach	Informatik (fächerübergreifend mit Ethik/Politik einsetzbar)
Klassenstufe	8.–12. Klasse (ca. 14–18 Jahre)
Zeitbedarf	90 Minuten (Kürzung auf 45 Min. möglich – siehe Hinweis S. 2)
Thema	Wie entsteht Bias in KI-Produkten?
Material	Beamer/Whiteboard, Zugang zu einem Bildgenerator für die Live-Erzeugung, Arbeitsblätter 1–3, Zitatkarten (Anlage)

Lernziele

- erklären, wie Trainingsdaten die Ausgaben eines generativen KI-Systems prägen (Ebene 1),
- beschreiben, dass Prompt-Rewriting, Diversitäts-Regeln (Ebene 2) zusätzliche technische Eingriffe von Herstellern sind, die diese nach dem Training vornehmen,
- Original-Quellentexte von Unternehmen (OpenAI, Google) auswerten und fachlich einordnen,
- ein einfaches Pipeline-Modell eines Text-zu-Bild-Systems skizzieren und eine eigene Systemregel entwerfen,
- begründet Stellung nehmen, an welcher Stelle im System sich Bias reduzieren ließe und welche Konsequenzen das haben könnte,
- im Partnerdialog eigene Rechercheergebnisse strukturiert vortragen, der Partnerin/dem Partner zuhören und gemeinsam eine Antwort erarbeiten.

KI-spezifische Lernziele

- Verstehen, dass scheinbar objektive KI-Ausgaben (Zahl, Wahrscheinlichkeit, Textantwort) zwei getrennte Bias-Quellen haben können: Trainingsdatenverzerrung (Ebene 1) und unsichtbare, anbieterseitige Nachbearbeitung (Ebene 2).
- Erkennen, dass eine KI-Ausgabe, die „neutral“ oder „wissenschaftlich“ wirkt, trotzdem eine gesellschaftlich geprägte Verzerrung transportieren kann.
- Methodenkompetenz: Verzerrung durch systematische Variation sichtbar machen (gleiche Frage/gleicher Prompt, unterschiedliche Gruppenmerkmale oder Rahmung) statt durch Einzelbeobachtung.

Verlaufsplan

Phase	Zeit	Geschehen	Sozialform / Material	Sozialform / Material
Einstieg	15'	Live-Bildexperiment in zwei Runden: Runde 1 – je 10 Bilder zu „ein Arzt“/ „eine Krankenschwester“ live mit einem Bildgenerator erzeugen (Beamer) und auszählen (Geschlecht, Alter, Hautfarbe). Runde 2 – Prompt „eine kompetente Person“ (gezielt bias-anfällig, s. Arbeitsblatt 1) live erzeugen und mit Runde 1 vergleichen. (Falls kein Bildgenerator für die Live-Präsentation zur Verfügung steht, können die Bilder auch mitgebracht werden).	Plenum / Bildgenerator + Beamer, Arbeitsblatt 1	Plenum / Bildgenerator + Beamer, Arbeitsblatt 1
Hinführung	5'	Impulsfrage: „Woher kommen die Verteilung? Entscheidet jemand im Hintergrund, was wir zu sehen bekommen? – An dieser Stelle noch nicht die Fragen beantworten.“	Plenum	Plenum

Erarbeitung I	15'	Kurzinput/Lesetext: Trainingsdaten als „Spiegel“ der Gesellschaft. Frühere Studien zu Midjourney (fast nur weiße Männer als Ärzte) und Adobe Firefly (sichtbare, unvollständige Diversifizierung) (Ebene 1).	Einzelarbeit + Plenum / Arbeitsblatt 1	Einzelarbeit + Plenum / Arbeitsblatt 1
Erarbeitung II	20'	Think – Pair – Share: Partner A und B bearbeiten zunächst je allein eine eigene Zitatkarte (Prompt-Rewriting bzw. Diversitätstechnik), erklären sich die Ergebnisse gegenseitig und entwickeln gemeinsam eine technische Vergleichsthese.	Partnerarbeit (Tandem) / Zitatkarten A-B	Partnerarbeit (Tandem) / Zitatkarten A-B
Vertiefung	22'	Kurzer Input-Text zur Pipeline lesen, danach Pipeline-Schema ergänzen (Prompt → Umschreiben → Bildmodell → Diversitäts-Regel → Ausgabe). Eigene Systemregel entwerfen; am Gemini-Beispiel diskutieren, wann eine Regel „danebengeht“.	Partnerarbeit / Arbeitsblatt 3	Partnerarbeit / Arbeitsblatt 3

Sicherung	8'	Merksatz gemeinsam formulieren. Rückbezug zum Einstiegsbild: Auf welcher Ebene könnte hier etwas gewirkt haben? (Die Pointe ist, dass wir es nicht wissen.) Kurzer Ausblick auf Studienergebnis.	Plenum / Tafelbild	Plenum / Tafelbild
Transfer / HA	5'	Gegebenenfalls als Hausaufgabe ein weiteres öffentlich dokumentiertes Bias-/Diversitäts- Update eines Herstellers recherchieren.	Einzelarbeit	Einzelarbeit
Hinweis zur Kürzung auf 45 Minuten Einstieg nur Runde 1, ohne Vergleichsprompt (8'). Erarbeitung I als Kurzinput ohne Lesetext (8'). Erarbeitung II nur „Ich“- und „Du“- Phase mit Zitatkarten A und B, ohne gemeinsame These (12'). Vertiefung nur als Plenumsdiskussion ohne eigenes Pipeline-Schema (10'). Sicherung (5'), Transfer entfällt.	Hinweis zur Kürzung auf 45 Minuten Einstieg nur Runde 1, ohne Vergleichsprompt (8'). Erarbeitung I als Kurzinput ohne Lesetext (8'). Erarbeitung II nur „Ich“- und „Du“- Phase mit Zitatkarten A und B, ohne gemeinsame These (12'). Vertiefung nur als Plenumsdiskussion ohne eigenes Pipeline-Schema (10'). Sicherung (5'), Transfer entfällt.	Hinweis zur Kürzung auf 45 Minuten Einstieg nur Runde 1, ohne Vergleichsprompt (8'). Erarbeitung I als Kurzinput ohne Lesetext (8'). Erarbeitung II nur „Ich“- und „Du“- Phase mit Zitatkarten A und B, ohne gemeinsame These (12'). Vertiefung nur als Plenumsdiskussion ohne eigenes Pipeline-Schema (10'). Sicherung (5'), Transfer entfällt.	Hinweis zur Kürzung auf 45 Minuten Einstieg nur Runde 1, ohne Vergleichsprompt (8'). Erarbeitung I als Kurzinput ohne Lesetext (8'). Erarbeitung II nur „Ich“- und „Du“- Phase mit Zitatkarten A und B, ohne gemeinsame These (12'). Vertiefung nur als Plenumsdiskussion ohne eigenes Pipeline-Schema (10'). Sicherung (5'), Transfer entfällt.	

Hintergrundinformationen für die Lehrkraft

Die klassische Bias-Demonstration, die noch in früheren Versionen von Bild-Generatoren funktioniert hat („Frag nach einem Arzt, es erscheint ein Mann“), führt in aktuellen Modellen zu anderen Ergebnissen. Das bedeutet nicht, dass Verzerrung verschwunden ist, sondern dass sie sich auf eine zweite, unsichtbare Ebene verlagert hat, die für den Informatikunterricht besonders instruktiv ist.

Ebene 1 – Trainingsdaten. Bildgeneratoren lernen aus Millionen von Bildern und den dazugehörigen Texten, die Menschen erstellt und beschriftet haben. Sie spiegeln damit bestehende gesellschaftliche Muster wider. Diese Ebene erklärt die klassische Bias-Lektion und ist auch Schülerinnen und Schülern intuitiv zugänglich.

Wie hartnäckig dieses Muster ist, zeigt eine recht aktuelle Analyse von 1.344 Bildern aus drei verbreiteten Bildgeneratoren (Stable Diffusion, DALL-E, Midjourney). Gab man diesen den neutralen Prompt, „a competent person“ zu zeichnen, so war in der Summe die Wahrscheinlichkeit 9mal größer, dass die KI einen Mann zeichnete als bei der Aufforderung, eine „person“ zu zeichnen. Die Systeme unterschieden sich bei dieser Aufgabe erheblich. Stable Diffusion generierte ausschließlich Männer, bei DALL-E waren es 71,9% Männer und 28,1 % Frauen, bei Midjourney war kein signifikanter Unterschied messbar (Weinmann et al., 2026). Obwohl der Prompt weder Beruf noch Geschlecht noch Herkunft nennt, war das Ergebnis teilweise einseitig. Deshalb wird er im Einstieg als zweite, verschärfte Testreihe genutzt (Arbeitsblatt 1, Runde 2), weil „ein Arzt“/„eine Krankenschwester“ mit aktuellen Modellen den Bias nicht mehr zuverlässig sichtbar macht.

Ebene 2 – nachträgliche technische Eingriffe. Zusätzlich verändern Hersteller, was ein System nach dem Training ausgibt. Dieser Schritt ist jederzeit änderbar, beruht ausschließlich auf Entscheidungen eines Unternehmens und ist im fertigen Bild nicht erkennbar. Für die Unterrichtseinheit relevant sind drei Mechanismen:

- **Prompt-Rewriting:** OpenAI beschreibt in der eigenen Entwicklerdokumentation, dass GPT-4 alle Prompts umschreibt, bevor sie an DALL-E gehen. Es gibt eine Möglichkeit, dieses Verhalten der KI abzuschalten.
- **Diversitätstechnik:** Eine System-Ebene ergänzt Herkunft/Geschlecht, wenn ein Prompt eine Person ohne diese Angaben beschreibt (z. B. „Feuerwehrfrau/-mann“). Nutzerinnen und Nutzer empfanden die Ergebnisse laut OpenAI danach zwölfmal häufiger als divers.
- **Tuning-Fehler:** Google musste die Bildgenerierung von Gemini 2024 pausieren, nachdem das Diversitäts-Tuning auch in Fällen griff, in denen es historisch nicht angemessen war. Google hatte in Gemini eine unsichtbare Regel eingebaut, die bei Personenbildern automatisch für Vielfalt sorgen sollte. Google selbst erklärte es am 23. Februar 2024 in einem offiziellen Blogbeitrag ("Gemini image generation got it wrong. We'll do better.", Autor: Prabhakar Raghavan, Senior VP) so: "Our tuning to ensure that Gemini showed a range of people failed to account for cases that should clearly not show a range." Die Regel unterschied also nicht zwischen Prompts wie "eine Führungskraft" (wo Vielfalt sinnvoll ist) und Prompts mit festgelegtem historischem Kontext (wo es nur eine faktisch richtige Antwort gibt). Konkret sorgte das für drei Tage lang viral gehende Beispiele: die US-amerikanischen Gründerväter wurden als Schwarze Personen dargestellt, ein Papst als Frau, und deutsche

Wehrmachtssoldaten von 1943 mit dunkelhäutigen und asiatischen Gesichtern. Zusätzlich kam ein zweites, separates Problem hinzu, das Google im selben Statement einräumte: Das Modell wurde "way more cautious than we intended and refused to answer certain prompts entirely – wrongly interpreting some very anodyne prompts as sensitive" – es verweigerte also teils auch harmlose Anfragen, weil die Sicherheits-Schwelle zu niedrig eingestellt war. Google pausierte die Personen-Bildgenerierung von Gemini bereits am 21./22. Februar 2024, also nur wenige Tage nach Launch, und stellte sie erst rund sechs Monate später (29. August 2024) mit einem verbesserten Modell (Imagen 3) und zusätzlichen Einschränkungen wieder her.

Für die fachliche Bewertung bedeutsam ist, dass sich aus dem fertigen Bild nicht mehr ablesen lässt, ob eine Veränderung auf Ebene 1 oder Ebene 2 stattgefunden hat. Aus Sicht des Informatikunterrichts ist dieser Umstand relevant, weil Schüler lernen können, dass ein System als mehrschichtige Pipeline mit menschlichen Entscheidungspunkten zu verstehen ist

Für Sie als Lehrer ist vielleicht eine Erfolgsmeldung von Interesse: In einer finnischen Studie mit 209 Schülerinnen und Schülern (Klassen 4 und 7) konnte gezeigt werden, dass Schüler diese Verzerrungen besser erkennen konnten, wenn sie gezielt dazu unterrichtet wurden. Der Anteil von rund 7 % auf 44 % (Vartiainen et al., 2025). Das Thema ist also auch für jüngere Jahrgänge gut vermittelbar.

Quellen

Vartiainen, H. et al. (2025). Enhancing children's understanding of algorithmic biases in and with text-to-image generative AI. *New Media & Society*, 27(9). DOI: 10.1177/14614448241252820

Weinmann, H., Messingschlager, T., & Appel, M. (2026). Gender bias in text-to-image generative artificial intelligence: Neglect and stereotypical presentations across three popular platforms. *New Media & Society*, Online First. <https://doi.org/10.1177/14614448261435197>

Zitate von OpenAI (Entwicklerdokumentation zu DALL·E, Beschreibung der Diversitätstechnik) und Google (Stellungnahme zu Gemini, 2024) wie in der Unterrichtsvorlage zitiert.

Arbeitsblatt 1: Die KI erzeugt die Wahrheit?

Was bedeutet „Bias“?

Bias (englisch, spricht „beiess“) bedeutet Schlagseite oder Voreingenommenheit. Ein KI-Bildgenerator hat Bias, wenn er bestimmte Menschen deutlich häufiger zeigt als andere, obwohl im Text gar nicht danach gefragt wurde – z. B. immer Männer bei „ein Arzt“. Genau das untersucht ihr gleich.

Runde 1: Klassischer Test

Prompts: „ein Arzt“ und „eine Pflegekraft“ – je 10 Bilder live erzeugen.

Bild-Nr.	Prompt	Wahrgenommenes Geschlecht	Geschätztes Alter	Sonstige Auffälligkeiten
1-10	ein Arzt			

1-10	eine Pflegerkraft			
------	----------------------	--	--	--

Runde 2: Verschärfter Bias-Test

Warum dieser Prompt?

In einer Analyse von 1.344 Bildern aus drei verbreiteten Bildgeneratoren zeigte eines der Systeme beim Prompt „Erzeuge das Bild einer kompetenten Person“ keine einzige Frau. Der Prompt nennt weder Beruf noch Geschlecht noch Herkunft und dennoch entstand ein einseitiges Ergebnis. Das macht ihn zu einem zuverlässigeren Bias-Test als „ein Arzt“, weil KI Bildgeneratoren inzwischen gezielt auf klassische Berufs-Prompts reagieren.

Bild-Nr.	Prompt	Wahrgenommenes Geschlecht	Geschätztes Alter	Sonstige Auffälligkeiten
1-10	eine kompetente Person			

Lesetext: Wie ein KI-System überhaupt lernt, wie „ein Arzt“ aussieht

Ein Bildgenerator kennt die Welt nicht aus eigener Erfahrung. Er wurde stattdessen mit Millionen Bildern aus dem Internet trainiert. Zu den Bildern hinzugefügt werden kurze Beschreibungen, die von Menschen erstellt werden. Aus diesen Beispielen „lernt“ das System ein Muster: Wenn auf den meisten Trainingsbildern zum Wort „Arzt“ Männer zu sehen waren, übernimmt das System dieses Muster und gibt es bei neuen Bildern wieder. Hinter dem Antwortverhalten steckt dabei kein System oder eine Absicht, sondern schlicht eine Reproduktion der Regeln, die aus den Bildern „gelernt“ wurden.

Das ist keine bloße Vermutung, sondern durch Untersuchungen belegt: In einem systematischen Test zeigte der Bildgenerator Midjourney beim Prompt „ein Arzt“ fast ausschließlich weiße Männer. Ein anderer Generator, Adobe Firefly, zeigte sichtbar mehr unterschiedliche Menschen. Die zwei Systeme liefern unterschiedliche Ergebnisse, zunächst einmal, weil sich ihre Trainingsdaten unterscheiden. Das nennt man Ebene 1 des Bias: Verzerrung, die aus den Trainingsdaten stammt.

Reflexionsfragen

- Vergleicht Runde 1 und Runde 2: In welcher Runde waren die Unterschiede deutlicher zu erkennen?
- Wieso ist es wahrscheinlicher, dass ein Prompt ganz ohne Berufs-, Geschlechts- oder Herkunftsangabe ein einseitiges Ergebnis liefert?
- Wie könnten die Ergebnisse aus Runde 1 bzw. 2 durch die Trainingsdaten entstanden sein?
- Was müsstet ihr über ein konkretes System wissen, um zu prüfen, ob wirklich die Trainingsdaten und nicht etwas anderes für das Ergebnis verantwortlich sind?

Arbeitsblatt 2 - Prompt Rewriting und Diversitätstechnik

Bildet Zweierteams. Verteilt die Rollen: Partner:in A bearbeitet Zitatkarte A, Partner:in B bearbeitet Zitatkarte B – jede:r zunächst für sich.

Methode: Partnerdialog „Ich - Du - Wir“

Ich (8'): Jede Person liest allein ihre eigene Karte (A bzw. B) und beantwortet schriftlich die Fragen.

Du (3'+3'): Ihr erklärt euch gegenseitig eure Karte in eigenen Worten – nicht vorlesen! Nachfragen sind ausdrücklich erwünscht.

Wir (3'): Vergleicht die beiden Mechanismen und entwickelt gemeinsam eine kurze These zur Frage der Abschaltbarkeit der Mechanismen.

Bevor ihr startet: Was bedeuten diese Begriffe?

Prompt-Rewriting: Bevor euer Text das Bildmodell erreicht, liest ihn zuerst ein anderes KI-System (ein Sprachmodell) und schreibt ihn automatisch um. In der Regel wird der umgeschriebene Prompt ausführlicher sein und Ergänzungen enthalten, die man gar nicht eingegeben hat.

Diversitätstechnik: Die Diversitätstechnik beschreibt eine zusätzliche Regel, die eingreift, wenn euer Prompt keine Angabe zu Geschlecht oder Herkunft der abgebildeten Person macht. Sie sorgt entsprechend der Standards der KI-Firma automatisch für unterschiedlichere Ergebnisse.

GPT-4: bezeichnet ein (älteres) Sprachmodell von OpenAI.

DALL·E: ein Bildgenerator von OpenAI, also eine KI, die aus einem Text ein Bild erzeugt.

Zitatkarte A: Prompt-Rewriting (OpenAI) — Aufgabe für Partner:in A

Eine Information von OpenAI. „... we use GPT-4 to optimize all of your prompts before they're passed to DALL·E“ – und: „this feature isn't able to be disabled at the moment.“

Auf Deutsch: „Wir nutzen GPT-4, um alle eure Prompts zu verbessern, bevor sie an DALL·E weitergegeben werden“ – und: „Diese Funktion lässt sich aktuell nicht abschalten.“

Quelle: OpenAI, Entwicklerdokumentation zu DALL·E

1. Was genau passiert hier, bevor ein Bild entsteht?

2. Warum bedeutet das: „Was ihr eingetippt habt, ist nicht das, was das Bildmodell erhalten hat“?

Vorstellungsauftrag: Erklärt eurem Partner/eurer Partnerin in eigenen Worten, was Prompt-Rewriting technisch macht – und warum es problematisch ist, dass sich diese Funktion nicht abschalten lässt.

Zitatkarte B: Diversitätstechnik (OpenAI) — Aufgabe für Partner:in B

„This technique is applied at the system level when DALL·E is given a prompt describing a person that does not specify race or gender, like 'firefighter.'“

Auf Deutsch: „Diese Technik wird automatisch angewendet, wenn DALL·E einen Prompt bekommt, der eine Person beschreibt, ohne Herkunft oder Geschlecht zu nennen – zum Beispiel ‚Feuerwehrkraft‘.“

Quelle: OpenAI, Beschreibung der Diversitäts-Technik; danach 12× häufigere Wahrnehmung diverser Ergebnisse

1. Unter welcher Bedingung wird diese Technik überhaupt aktiv?

2. Ist „ein Arzt“ ein Prompt, bei dem diese Technik greifen würde? Begründet.

Vorstellungsauftrag: Erklärt eurem Partner/eurer Partnerin in eigenen Worten, wann diese Technik aktiv wird – und was vermutlich passieren würde, wenn im Prompt bereits „ein junger männlicher Arzt“ stünde.

Wir-Phase: Gemeinsame technische These

Es gibt zahlreiche Sprachmodelle, die nicht wie ChatGPT von OpenAI oder Claude von Anthropic auf Servern laufen, sondern die man auf seinem eigenen Computer installieren kann. Überlegt euch eine Hypothese zur Frage, welche der Techniken zur Kontrolle der Antworten ihr auf lokalen Modellen finden könntet. Begründet eure Hypothese.

Arbeitsblatt 3: Wie mach ein Bildgenerator aus Text ein Bild?

Ihr benötigt die Informationen aus dem nachfolgenden Text zur Beantwortung der Aufgabe.

1. Lest ihn daher in Ruhe durch. Er beschreibt vereinfacht, was zwischen eurer Eingabe und dem fertigen Bild passiert.

① **Nutzereingabe (Prompt):** Ihr tippt eure Aufgabe an die KI ein, den Prompt, z. B. „Erzeuge ein Bild von einem Arzt“.

② Prompt-Umschreibung: Bevor daraus ein Bild wird, liest ein Sprachmodell (eine KI, die mit Text arbeitet) euren Prompt und schreibt ihn automatisch um. Dabei werden Informationen ergänzt, z.B. Details oder auch Ergänzungen, die ihr gar nicht mitgedacht habt oder auch nicht mitgedacht haben wollt. Dieser Schritt lässt sich bei den Modellen nicht immer abschalten.

③ Bildmodell erzeugt Bild: Ein zweites KI-System, das Diffusionsmodell genannt wird, erzeugt aus dem umgeschriebenen Text das eigentliche Bild. Es startet mit einem Bild aus reinem Zufallsrauschen und verändert es in vielen kleinen Schritten, bis ein Bild entsteht, das zum Text passt. Es „weiß“, wie z. B. ein Arzt typischerweise aussieht, weil es beim Training Millionen Bilder mit den zugehörigen Beschreibungen gesehen hat.

④ Diversitäts-Regel (optional): Eine der zusätzlichen Informationen, die das Modell ergänzt, sind Angaben z.B. zu Geschlecht oder Herkunft. Das Unternehmen, das die KI vertreibt, greift damit direkt in die Bilderzeugung ein, die für unterschiedlichere Ergebnisse sorgt. Das geschieht automatisch, ohne dass man das im fertigen Bild erkennt

⑤ Ausgabe an Nutzer:in: Ihr bekommt das fertige Bild ohne jeden Hinweis darauf, was in den Schritten ② bis ④ verändert oder ergänzt wurde.

Wichtige Begriffe

Prompt: die Texteingabe, mit der man ein KI-System zu einer Ausgabe auffordert.

Sprachmodell: ein KI-System, das Text „versteht“ und selbst Text erzeugen kann (z. B. GPT-4).

Diffusionsmodell: ein KI-System, das aus zufälligem Bildrauschen schrittweise ein Bild erzeugt.

Trainingsdaten: die Bilder und Texte, aus denen ein System vorab „gelernt“ hat.

2- Übertrag nun die wichtigsten Informationen aus dem Text in eigenen Worten in die Tabelle:

3. Aufgabe: eigene Systemregel entwerfen

Hintergrund: der Gemini-Fall (2024)

Google musste die Bildgenerierung seines Systems Gemini 2024 vorübergehend abschalten. Der Grund dafür war, dass die von Google eingebaute Diversitätsregel auch dort die Bilderzeugung beeinflusste, wo sie keinen Sinn ergab. So wurden bei Bildanfragen zu bestimmten historischen Situationen Personen-Zusammensetzungen erzeugt, die es historisch so nicht gegeben hatte. Google selbst erklärte dazu: Die Regel sei nicht darauf vorbereitet gewesen, Fälle zu erkennen, in denen sie eindeutig nicht passte. Das zeigt: Auch eine gut gemeinte Regel kann danebengehen, wenn sie zu pauschal angewendet wird.

- Entwerft in Partnerarbeit eine eigene Regel, die ein Bildgenerator anwenden könnte, wenn ein Prompt eine Person ohne Geschlecht/Herkunft beschreibt (z. B. „eine Führungskraft“).
- Sucht anschließend nach potenziellen Fällen, in denen eure Regel zu einem unpassenden Ergebnis führen würde.