

How AI images are made — and why they are never neutral

Computer Science · Secondary level (ages 14–18) · 90 minutes, can be shortened to 45

Information for Teachers

Fact sheet

Category	Details
Subject	Computer science (can also be used across subjects in ethics/politics)
Grade level	Grades 8–12 (approx. ages 14–18)
Time required	90 minutes (can be shortened to 45 minutes – see note, p. 2)
Topic	How does bias arise in AI products?
Materials	Projector/whiteboard, access to an image generator for live generation, worksheets 1–3, quote cards (appendix)

Learning objectives

- explain how training data shapes the output of a generative AI system (Layer 1),
- describe that prompt rewriting and diversity rules (Layer 2) are additional technical interventions that manufacturers make after training,
- analyze original source texts from companies (OpenAI, Google) and place them in context,
- sketch a simple pipeline model of a text-to-image system and design a system rule of their own,
- take a reasoned position on where in the system bias could be reduced and what consequences that could have,
- present their own research findings in a structured way during partner dialogue, listen to their partner, and work out an answer together.

AI-specific learning objectives

- Understand that seemingly objective AI outputs (a number, a probability, a text answer) can have two separate sources of bias: distortion from the training data (Layer 1) and invisible post-processing by the provider (Layer 2).
- Recognize that an AI output that seems “neutral” or “scientific” can still carry a socially shaped distortion.
- Methodological skill: make distortion visible through systematic variation (same question/same prompt, different group characteristics or framing) rather than through single observations.

Lesson sequence

Phase	Time	What happens	Social form / Materials	Social form / Materials
Introduction	15'	Live image experiment in two rounds: Round 1 – generate 10 images each for “a doctor”/“a nurse” live with an image generator (projector) and tally them (gender, age, skin color). Round 2 – generate the prompt “a competent person” live (deliberately bias-prone, see Worksheet 1) and compare with Round 1. (If no image generator is available for the live presentation, the images can also be brought in beforehand.)	Whole class / image generator + projector, Worksheet 1	Whole class / image generator + projector, Worksheet 1
Lead-in	5'	Prompt question: “Where does this distribution come from? Is someone behind the scenes deciding what we get to see?” – Don't answer the questions yet at this point.	Whole class	Whole class
Development I	15'	Short input/reading text: training data as a “mirror” of society. Earlier studies on Midjourney (almost only white men as doctors) and Adobe Firefly (visible, incomplete diversification) (Layer 1).	Individual work + whole class / Worksheet 1	Individual work + whole class / Worksheet 1

Development II	20'	Think – Pair – Share: Partners A and B first each work alone on their own quote card (prompt rewriting or diversity technique), explain their findings to each other, and jointly develop a technical comparison thesis.	Partner work (in pairs) / Quote cards A-B	Partner work (in pairs) / Quote cards A-B
Deepening	22'	Read a short input text on the pipeline, then complete the pipeline diagram (Prompt → Rewriting → Image model → Diversity rule → Output). Design a system rule of their own; discuss, using the Gemini example, when a rule “goes wrong.”	Partner work / Worksheet 3	Partner work / Worksheet 3
Wrap-up	8'	Formulate a key takeaway together. Refer back to the opening images: which layer could have been at work here? (The point is that we don't know.) Brief outlook on a study finding.	Whole class / board notes	Whole class / board notes
Transfer / Homework	5'	If time allows, research another publicly documented bias/diversity update from a manufacturer as homework.	Individual work	Individual work

Note on shortening to 45 minutes Introduction only Round 1, without the comparison prompt (8'). Development I as a short input without the reading text (8'). Development II only the 'I' and 'You' phases with quote cards A and B, without the joint thesis (12'). Deepening only as a whole-class discussion without students' own pipeline diagram (10'). Wrap-up (5'), transfer is dropped.	Note on shortening to 45 minutes Introduction only Round 1, without the comparison prompt (8'). Development I as a short input without the reading text (8'). Development II only the 'I' and 'You' phases with quote cards A and B, without the joint thesis (12'). Deepening only as a whole-class discussion without students' own pipeline diagram (10'). Wrap-up (5'), transfer is dropped.	Note on shortening to 45 minutes Introduction only Round 1, without the comparison prompt (8'). Development I as a short input without the reading text (8'). Development II only the 'I' and 'You' phases with quote cards A and B, without the joint thesis (12'). Deepening only as a whole-class discussion without students' own pipeline diagram (10'). Wrap-up (5'), transfer is dropped.	Note on shortening to 45 minutes Introduction only Round 1, without the comparison prompt (8'). Development I as a short input without the reading text (8'). Development II only the 'I' and 'You' phases with quote cards A and B, without the joint thesis (12'). Deepening only as a whole-class discussion without students' own pipeline diagram (10'). Wrap-up (5'), transfer is dropped.	
--	--	--	--	--

Background information for teachers

The classic bias demonstration that still worked in earlier versions of image generators (“ask for a doctor, a man appears”) produces different results in current models. That does not mean the distortion has disappeared – it means it has shifted to a second, invisible layer, which is especially instructive for computer science lessons.

Layer 1 – training data. Image generators learn from millions of images and the accompanying texts that people created and captioned. In doing so, they mirror existing patterns in society. This layer explains the classic bias lesson and is also intuitively accessible to students.

How persistent this pattern is shows in a fairly recent analysis of 1,344 images from three widely used image generators (Stable Diffusion, DALL-E, Midjourney). When given the neutral prompt to draw “a competent person,” the AI was, overall, 9 times more likely to draw a man than when asked to draw a “person.” The systems differed considerably on this task. Stable Diffusion generated only men; DALL-E produced 71.9% men and 28.1% women; Midjourney

showed no measurable significant difference (Weinmann et al., 2026). Even though the prompt names no profession, gender, or origin, the result was partly one-sided. That is why it is used in the introduction as a second, sharper test round (Worksheet 1, Round 2) – because “a doctor”/“a nurse” no longer reliably reveals bias with current models.

Layer 2 – technical interventions after the fact. In addition, manufacturers change what a system outputs after training. This step can be changed at any time, rests solely on a company's own decisions, and cannot be recognized in the finished image. Three mechanisms are relevant for this lesson:

- Prompt rewriting: OpenAI describes in its own developer documentation that GPT-4 rewrites every prompt before it is passed to DALL·E. There is a way to turn off this AI behavior.
- Diversity technique: A system-level layer adds origin/gender when a prompt describes a person without specifying these (e.g., “firefighter”). According to OpenAI, users then perceived the results as diverse twelve times more often.
- Tuning error: Google had to pause Gemini's image generation in 2024 after its diversity tuning also kicked in in cases where it was not historically appropriate. Google had built an invisible rule into Gemini that was meant to automatically ensure variety in images of people. Google itself explained this on February 23, 2024, in an official blog post (“Gemini image generation got it wrong. We'll do better.”, author: Prabhakar Raghavan, Senior VP), stating: “Our tuning to ensure that Gemini showed a range of people failed to account for cases that should clearly not show a range.” In other words, the rule did not distinguish between prompts like “a business leader” (where variety makes sense) and prompts with a fixed historical context (where there is only one factually correct answer). In concrete terms, this produced examples that went viral for three days: the American founding fathers were depicted as Black people, a pope as a woman, and German Wehrmacht soldiers from 1943 with dark-skinned and Asian faces. On top of that came a second, separate problem, which Google admitted in the same statement: the model became “way more cautious than we intended and refused to answer certain prompts entirely – wrongly interpreting some very anodyne prompts as sensitive” – meaning it also refused some harmless requests, because the safety threshold was set too low. Google had already paused Gemini's people-image generation on February 21/22, 2024, just a few days after launch, and only restored it about six months later (August 29, 2024) with an improved model (Imagen 3) and additional restrictions.

What matters for a proper assessment is that the finished image no longer shows whether a change happened at Layer 1 or Layer 2. From the perspective of computer science lessons, this is relevant because students can learn that a system should be understood as a multi-layered pipeline with human decision points

As a teacher, you might be interested in one success story: a Finnish study with 209 students (grades 4 and 7) showed that students could recognize these distortions better once they were specifically taught about them. The share rose from around 7% to 44% (Vartiainen et al., 2025). So the topic can be taught well to younger year groups too.

Sources

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>

Quotes from OpenAI (developer documentation on DALL·E, description of the diversity technique) and Google (statement on Gemini, 2024) as cited in the lesson template.

Worksheet 1: Does AI generate the truth?

What does 'bias' mean?

Bias means a slant, or a built-in leaning toward one side. An AI image generator has bias when it shows certain people noticeably more often than others, even though the text never asked for that – for example, always showing men for 'a doctor.' That's exactly what you're about to investigate.

Round 1: Classic test

Prompts: "a doctor" and "a nurse" – generate 10 images each, live.

Image no.	Prompt	Perceived gender	Estimated age	Other notable features
1-10	a doctor			
1-10	a nurse			

Round 2: Sharper bias test

Why this prompt?

In an analysis of 1,344 images from three widely used image generators, one of the systems produced not a single woman for the prompt "Generate an image of a competent person." The prompt names no profession, gender, or origin, and yet the result was one-sided. That makes it a more reliable bias test than "a doctor," because AI image generators have since been specifically tuned to respond to classic profession prompts.

Image no.	Prompt	Perceived gender	Estimated age	Other notable features
1-10	a competent person			

Reading text: How an AI system learns what 'a doctor' looks like in the first place

An image generator doesn't know the world from its own experience. Instead, it was trained on millions of images from the internet. Added to the images are short descriptions that people

created. From these examples, the system “learns” a pattern: if most of the training images for the word “doctor” showed men, the system adopts this pattern and reproduces it in new images. There is no plan or intention behind this response behavior – it is simply a reproduction of the rules the system “learned” from the images.

This isn't just a guess – it's backed up by research: in a systematic test, the image generator Midjourney showed almost exclusively white men for the prompt “a doctor.” Another generator, Adobe Firefly, showed visibly more varied people. The two systems produce different results, first of all, because their training data differs. This is called Layer 1 of bias: distortion that comes from the training data.

Reflection questions

- Compare Round 1 and Round 2: in which round were the differences easier to see?
- Why is a prompt with no mention of profession, gender, or origin more likely to produce a one-sided result?
- How could the results from Round 1 and Round 2 have come from the training data?
- What would you need to know about a specific system to check whether it's really the training data – and not something else – responsible for the result?

Worksheet 2 – Prompt Rewriting and Diversity Technique

Form teams of two. Split up the roles: partner A works on quote card A, partner B works on quote card B – each of you on your own at first.

Method: partner dialogue ‘I – You – We’

I (8'): Each person reads their own card (A or B) alone and answers the questions in writing.

You (3'+3'): You explain your card to each other in your own words – don't just read it out loud!

Asking follow-up questions is explicitly encouraged.

We (3'): Compare the two mechanisms and jointly develop a short thesis on the question of whether these mechanisms can be turned off.

Before you start: what do these terms mean?

Prompt rewriting: Before your text reaches the image model, another AI system (a language model) reads it first and automatically rewrites it. As a rule, the rewritten prompt will be more detailed and contain additions that you never typed in.

Diversity technique: The diversity technique refers to an additional rule that kicks in when your prompt doesn't specify the gender or origin of the person shown. It automatically produces more varied results, according to the AI company's own standards.

GPT-4: refers to an (older) language model made by OpenAI.

DALL·E: an image generator made by OpenAI – that is, an AI that turns a text into an image.

Quote card A: Prompt Rewriting (OpenAI) — task for partner A

A statement from OpenAI. "... we use GPT-4 to optimize all of your prompts before they're passed to DALL·E" – and: "this feature isn't able to be disabled at the moment."

Source: OpenAI, developer documentation for DALL·E

1. What exactly happens here before an image is generated?
2. Why does this mean: 'What you typed is not what the image model actually received'?

Explanation task: Explain to your partner, in your own words, what prompt rewriting technically does – and why it's a problem that this feature can't be turned off.

Quote card B: Diversity Technique (OpenAI) — task for partner 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.'"

Source: OpenAI, description of the diversity technique; afterward, results were perceived as diverse 12× more often

1. Under what condition does this technique become active in the first place?
2. Is 'a doctor' a prompt where this technique would kick in? Explain your reasoning.

Explanation task: Explain to your partner, in your own words, when this technique becomes active – and what would probably happen if the prompt already said 'a young male doctor.'

We phase: joint technical thesis

There are numerous language models that don't run on servers like ChatGPT from OpenAI or Claude from Anthropic, but that you can install on your own computer instead. Think of a hypothesis about which of the response-control techniques you might still find in local models. Give reasons for your hypothesis.

Worksheet 3: How does an image generator turn text into an image?

You'll need the information from the text below to answer the task.

1. So read it through carefully. It describes, in simplified form, what happens between your input and the finished image.

① **User input (prompt):** You type your instruction to the AI, the prompt, e.g., “Generate an image of a doctor.”

② **Prompt rewriting:** Before this becomes an image, a language model (an AI that works with text) reads your prompt and automatically rewrites it. In doing so, information gets added – for example, details or additions you never thought of, or maybe didn't want added at all. Depending on the model, this step can't always be turned off.

③ **Image model generates the image:** A second AI system, called a diffusion model, generates the actual image from the rewritten text. It starts with an image of pure random noise and changes it in many small steps until an image emerges that matches the text. It “knows” what a doctor typically looks like, for example, because during training it saw millions of images along with their descriptions.

④ **Diversity rule (optional):** One of the additional pieces of information the model adds concerns things like gender or origin. The company that distributes the AI directly intervenes in the image generation this way, which produces more varied results. This happens automatically, without anyone being able to tell from the finished image

⑤ **Output to the user:** You get the finished image with no indication at all of what was changed or added in steps ② through ④.

Key terms

Prompt: the text input used to prompt an AI system to produce an output.

Language model: an AI system that “understands” text and can generate text itself (e.g., GPT-4).

Diffusion model: an AI system that generates an image step by step from random image noise.

Training data: the images and texts a system “learned” from beforehand.

2- Now transfer the most important information from the text into the table, in your own words:

3. Task: design your own system rule

Background: the Gemini case (2024)

Google had to temporarily shut down the image generation of its Gemini system in 2024. The reason was that the diversity rule Google had built in also influenced image generation in cases where it made no sense. As a result, requests for images of certain historical situations produced combinations of people that had never existed historically. Google itself explained that the rule had not been prepared to recognize cases where it clearly didn't fit. This shows: even a well-intentioned rule can go wrong if it's applied too broadly.

- Working with your partner, design a rule of your own that an image generator could apply when a prompt describes a person without specifying gender/origin (e.g., ‘a business leader’).
- Then look for potential cases in which your rule would lead to an inappropriate result.