EdSutra

Teaching guide

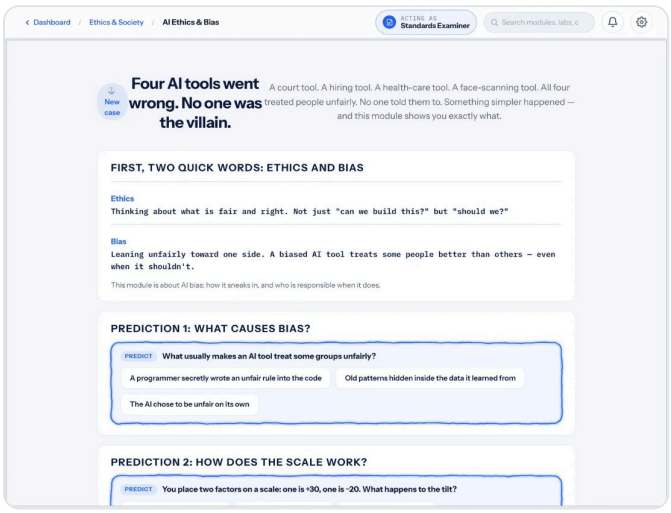
Module 05 · Ethics & Society · Unit 01 · Stage 3, Responsibility

AI Ethics & Bias

The Fairness Scale

Grades 8–12 | 50 minutes | Role: **Standards Examiner**

Grade 8: About age 13–14 · Year 9 in the UK · Class 8 in India



The big idea

AI bias rarely starts with someone writing an unfair rule. It builds up from ordinary choices: what data was collected, what stood in for gender or race, what the model was built to predict, and where it was used. Checks such as audits push back, but no check catches everything, and none changes the data a model learned from.

The learner's job

Learners are the **Standards Examiner** at Northvale Civic Standards Office. Northvale wants to buy an AI system. Nobody has asked where it could go wrong. Their task: find where bias gets into a system. Pick the checks that push back.

By the end, learners can say

- I can recall that bias enters through data, algorithm design, and deployment — not just intent
- I can predict which category caused a real documented AI-bias case, then check against the record
- I can operate the fairness scale and read the tilt as a real sum, not a mood ring
- I can defend which mitigations would cut a given bias fastest, citing case evidence

Before class

- One computer or tablet for each learner or pair, with EdSutra open (this Module is part of the full EdSutra studio).
- Open the studio once on each device before class, so the Unit still runs if the internet drops.
- Play the intro and the first case yourself, so you know where learners will click.
- Draw three columns on the board: Data, Algorithm, Deployment. You will sort the four cases into them during Analyze.
- Best after What is AI? and How Machines Learn.
- Units open in order, and Module 1 comes first. To start a class here, switch on Open every Unit in Settings → Teachers & parents.

Words to introduce

Ethics: Thinking about what is fair and right: not just can we, but should we.

Bias: Leaning unfairly toward one side, so some people are treated worse than others.

Proxy variable: A stand-in, like a word or zip code, that tracks gender or race.

Historical bias: Old unfair decisions inside the data, which a model learns and repeats.

Feedback loop: A model's outputs become its next data, so a pattern keeps growing.

Mitigation: A check that pushes back against bias, such as an independent audit.

Running the lesson 50 minutes

Open 5 min

"Four AI tools treated people unfairly: a court tool, a hiring tool, a health-care tool and a face scanner. Nobody told them to. So where did the unfairness come from?"

Before anyone clicks, ask: for each of the four tools, did the unfairness start in the data, in how the tool was built, or in where it was used? Write the guesses on the board and come back to them in Analyze.

Step	What learners do	Ask the class	Listen for
1. Remember 4 min	Read the two quick words, ethics and bias, then answer the first prediction: what usually makes an AI tool treat some groups unfairly?	• In your own words, what does bias mean for an AI tool? • Why is "a programmer wrote an unfair rule" the wrong answer here?	Bias usually comes from old patterns in the data a tool learned from, not from someone meaning harm.
2. Understand 4 min	Answer the second prediction: one factor is +30, one is -20, so what does the scale show? Then step up to the Fairness Scale.	• What does the tilt number actually measure? • Why is a scale a useful picture of bias, and where does it fall short?	The tilt is a running total: +30 and -20 make +10. The weights are a ranking written for teaching, not measurements.
3. Apply 10 min	Click bias factors and checks to place them on the scale or take them off, and watch the tilt. Learners need at least four placed before their defense counts.	• Which single factor pushes the scale furthest toward biased? • What happens when you add an independent audit? Is the system now fair?	Checks pull the tilt back, but a balanced reading is not proof of fairness. That has to be checked on real results.
4. Analyze 15 min	Open the four real cases. For each, predict whether data, the algorithm or deployment was mainly responsible, then read the documented outcome and its lesson.	• The hiring tool never saw a gender column. Why did it mark down resumes with the word "women's"? • The health-care tool's cost records were accurate. Why was it still unfair? • Look at our guesses on the board. Which did we get wrong?	A word stood in for gender, and predicting cost carried old unequal spending. Bias entered through ordinary choices.
5. Evaluate 12 min	In DEFEND, pick the two checks to use first, choose the reason that backs them, and say what the plan cannot do alone. Then take the five-question quick check.	• Which two checks did you choose, and which case backs your reason? • What can audits and diverse teams not fix on their own?	Checks can find and flag bias, but the data or decisions the model learned from still have to change.

Steps follow Bloom's revised taxonomy, in the order the Unit presents them.

Deepen the thinking

Critical thinking	The scale's weights were written for teaching, not measured. When is a made-up model still useful, and when could it mislead?
AI thinking	The hiring tool never saw gender, yet found a stand-in. If you removed that word too, what new stand-in might it find?
Ethics and fairness	A town knows a face scanner works worse for some faces. Is using it anyway a technical problem or a decision someone made?
Connect to life	Think of a rule at school or online that seems neutral. Could it still work worse for some people? How?

Misconceptions to listen for

"AI bias means someone programmed it to be unfair."
None of the four cases began with an unfair rule. The causes were the data, what a tool was built to predict, and where it was used.

"Removing race or gender from the data fixes it."
The hiring tool never saw gender, yet it marked down the word "women's". A stand-in can carry the same pattern.

"If accuracy is high, the system is fine."
The Misaligned Objective factor and the cases show a tool can be accurate overall and still fail one group badly.

In real life the card learners see at the end of the Unit

- Hospitals, banks and job sites use scoring tools; old patterns in their data can treat some groups unfairly.
- Hiring tools in the news: one marked down applications saying "women's", though gender was never an input.
- Face-recognition tools can work worse for some faces when their training photos lacked people like them.

Try it this week. When a tool seems to work better for some people, guess where bias could enter: the data, the design, or how it's used.

Watch out for. AI bias rarely starts with someone writing an unfair rule. It usually comes from patterns in the data, and high accuracy can hide it.

Talk about it. Is a tool fair if it's right most of the time but wrong more often for one group?

Exit ticket

1. Name the three places bias can enter an AI system.

   _____

2. What is a proxy variable? Give the example from the hiring case.

   _____

3. Name one check that pushes back on bias, and one thing it cannot fix alone.

   _____

Going further

In small groups, learners pick one of the four cases and present a one-page plan: where the bias entered, the two checks they would add first, and what those checks still could not fix.