EdSutra

Teaching guide

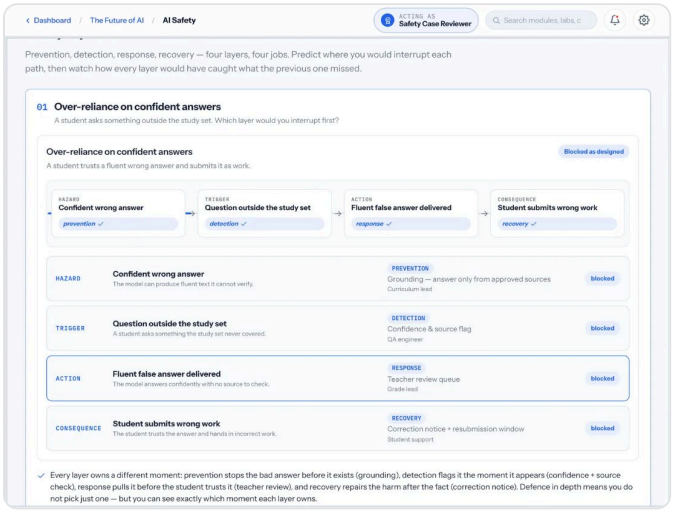
Module 08 · The Future of AI · Unit 03 · Stage 3, Responsibility

AI Safety

The Safety Case Lab

Grades 7–12 | **30 minutes** | Role: **Safety Case Reviewer**

Grade 7: About age 12–13 · Year 8 in the UK · Class 7 in India



The big idea

An AI system is not safe because it is smart or because one check passed. Safety means tracing each path from a hazard to harm, and checking that prevention, detection, response and recovery each have evidence and a named owner. When something fails, look for the first layer that broke, not the final output.

The learner's job

Learners are the **Safety Case Reviewer** at Fernfield Safety Bench · Review Office. The school board wants to know if the Study Assistant is safe to keep. You cannot hand them a yes or a no — you can only show them every path, and exactly where each one stopped. Their task: trace how a system can fail — and name the first layer that was supposed to stop it.

By the end, learners can say

- I can match hazard, risk, alignment, misuse, robustness, monitoring, containment, and recovery to plain-language examples
- I can predict where a failure path should be interrupted and explain how the four control layers differ
- I can assemble a safety case by linking each claim to supporting evidence and a named control owner
- I can run failure injections and locate the first missing or failed control
- I can decide deploy / pilot / revise / do-not-deploy under explicit constraints using residual paths, evidence gaps, and reversibility
- I can redesign an architecture or operating policy to close a hazardous path and document the remaining uncertainty

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 Remember and Understand yourself, including the Layer Dam, where you switch layers off and watch the hazard move.
- Write the four layers on the board in order: Prevention, Detection, Response, Recovery. Point to them all lesson.
- Best after Reinforcement Learning and AI & Climate.
- 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

Hazard: A situation or property that can cause harm.

Risk: A hazard combined with a chance of it being reached.

Alignment: The system pursues the real goal, not a shortcut such as pleasing the user.

Misuse: A person using the system in a way it was not designed for.

Defense in depth: Prevention, detection, response and recovery layers, each catching what the last missed.

Safety case: Claims about a system, each backed by evidence and a named owner.

Recovery: Returning to a known-good state and learning from the failure.

Running the lesson 40 minutes

Open 4 min

"The school board wants a yes or a no: is the Study Assistant safe to keep? Why might a plain yes or no not be enough?"

Before anyone clicks, ask: a study bot gives a confident wrong answer, and a learner hands it in as homework. Where could someone have stopped it? Write the ideas on the board and come back to them in Understand.

Step	What learners do	Ask the class	Listen for
1. Remember 4 min	Learners match eight plain-language examples to eight safety terms, from hazard and risk to containment and recovery.	• What is the difference between a hazard and a risk? • Which term fits a tutor that corrects you instead of just agreeing?	A hazard can cause harm; a risk is that hazard plus a real chance of reaching it.
2. Understand 7 min	For two failure paths, learners predict which layer they would interrupt first, then watch each layer catch what the last one missed. On the Layer Dam they switch layers off.	• What moment does each of the four layers own? • Look at our board. Which layer does each of our ideas belong to?	Each layer has a different job. With layers switched off together, the hazard reaches its consequence.
3. Apply 7 min	Learners build the safety case for the fictional Fernfield Study Assistant. For each of three claims, they attach the evidence that supports it and the owner who enforces it.	• Why is a claim with no named owner "just a hope"? • Why does attaching every record make the case weaker, not stronger?	A claim is only as strong as the evidence behind it and the person responsible for it.
4. Analyze 6 min	Three incidents, each with one control down or unstaffed. Learners trace the path and name the layer that failed first. They need two of the three, and can retry.	• The flags fired, but nobody was allowed to act. Which layer failed? • Why look for the first broken layer, not the final output?	Response failed: a flag with no one to act on it is a broken layer, even though detection worked.
5. Evaluate 6 min	Two decisions: roll out to every classroom, and stay live after a breach drill. Learners pick deploy, pilot, revise or do not deploy, then choose at least two pieces of evidence that support it.	• Which evidence supported your verdict, and what did you leave out? • Why can both "deploy" and "do not deploy" pass?	The verdict is a judgment. What counts is evidence that really supports it, and none that argues against it.
6. Create 6 min	Learners choose one redesign that closes a failure path and re-run the worst-case drill: four reachable paths drop to three. They name what is still uncertain and pick a review date.	• What is still uncertain after your redesign? • Why does the Unit refuse a 12-month review while a gap is open?	A path closed on paper still needs testing, so the open questions and a review date stay visible.

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

Deepen the thinking

Critical thinking	A company says its AI passed every safety test. What would you want to see before you believed it was safe to use?
AI thinking	The flattery loop came from rewarding agreement. How is that like the Reward Lab's robot circling the beacon?
Ethics and fairness	If a safety layer fails, which learners are most likely to be harmed, and who should tell them?
Connect to life	Think of a time something went wrong and was fixed. What was the first thing that should have caught it?

Misconceptions to listen for

"If the model is smart enough, it won't fail."
A model answering confidently and nobody stopping it are separate failures. Intelligence is not a safety layer.

"One checklist makes a system safe."
On the Layer Dam, switching layers off lets the hazard pass further. Each layer catches what the last one missed.

"The problem is wherever the bad output appeared."
In Analyze, every incident is traced back to the first missing or failed control, not the final answer.

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

- A school fire plan has layers (alarms, drills, exits, a meeting point); AI safety uses layers too.
- Chatbots use filters, warnings and report buttons, because no single check catches every problem.
- When a company pauses or rolls back an AI feature after problems, that's response and recovery.

Try it this week. Pick an everyday safety system, like crossing a road. Name what prevents, detects, responds to and recovers from a mistake.

Watch out for. A smarter AI isn't automatically safer, and one checklist isn't enough. When something goes wrong, look for the first layer that failed.

Talk about it. Who should decide whether an AI tool is safe enough to use at school?

Exit ticket

1. Name the four safety layers in order.

2. What two things must back every safety claim?

3. Why look for the first layer that failed?

Going further

For a class project, groups pick an AI tool learners use and draw one failure path from hazard to consequence. For each step they name a layer, a control and an owner, then present the weakest point.