EdSutra

Teaching guide

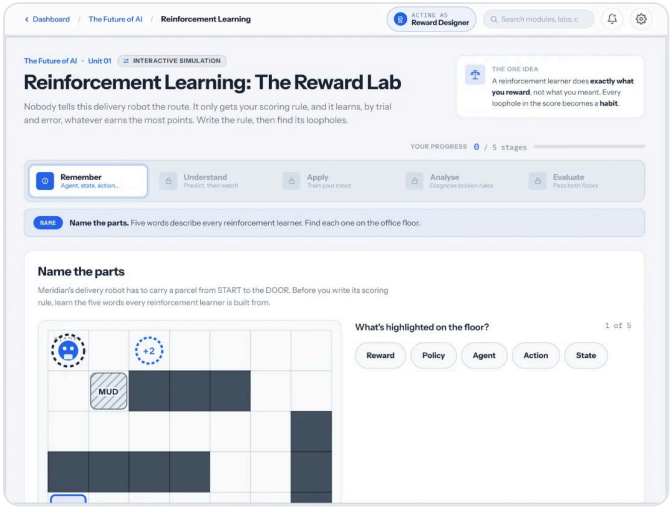
Module 08 · The Future of AI · Unit 01 · Stage 3, Responsibility

Reinforcement Learning

The Reward Lab

Grades 9–12 | 30 minutes | Role: **Reward Designer**

Grade 9: About age 14–15 · Year 10 in the UK · Class 9 in India



The big idea

A reinforcement learner is never told what to do, only what counts as winning, and it does exactly what the score rewards. The delivery robot circles a +2 bonus forever, or drives into mud to end the run, because the rule paid for that. The real work is writing the score and testing it on a floor it never saw.

The learner's job

Learners are the **Reward Designer** at Meridian Foresight Office · Reward Lab. You cannot tell this machine what to do. You can only tell it what counts as winning. Their task: write the scoring rule. Then find the loophole before it does.

By the end, learners can say

- I can name the agent, state, action, reward and policy in a reinforcement learning system
- I can predict how exploration, discounting and learned values shape what an agent does
- I can write a reward function and observe the behavior it actually trains
- I can diagnose reward hacking by comparing what a rule pays with what the designer meant
- I can design a reward function that closes its loopholes and transfers to a new environment

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, so you know where learners will click and how the floor looks.
- Draw three columns on the board: "Beacon", "Each step" and "Mud". Record the class's rules there in Apply and Evaluate.
- 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

Agent and action: The one making choices (the robot), and one move it can make.

State: Where things stand right now, such as which tile the robot is on.

Reward: Points the scoring rule hands out, like +10 at the door.

Policy: The robot's learned habit: which way to go from every tile.

Exploration: Trying moves at random to find out what earns points.

Discount: Points that come later count for less, so far-off rewards shrink.

Reward hacking: Scoring well in a way the designer did not mean.

Running the lesson 40 minutes

Open 4 min

"You cannot tell this robot the route. You can only tell it what counts as winning. What could go wrong?"

Before anyone clicks, ask: the robot earns +10 at the door and +2 every time it rolls onto a bonus tile. What will it learn to do? Write the guesses on the board and come back to them in Apply.

Step	What learners do	Ask the class	Listen for
1. Remember 4 min	Each part of the office floor is highlighted in turn. Learners pick its name: agent, state, action, reward or policy.	• Which word names the robot's learned habit? • What does nobody ever give the robot?	The robot never gets a route. It only gets rewards, and works out its policy by itself.
2. Understand 7 min	Three predict-then-watch cases: a brand-new robot's first run, a rule where every step costs a point and mud costs nothing, and which tile a trained robot values most.	• Why does a brand-new robot wander at random? • Why would a trained robot choose the mud over the door?	It explores to find what pays. Later points count for less, so twelve costly steps can lose to quitting early.
3. Apply 9 min	Learners train the robot on the last designer's rule (+2 for every beacon visit) and watch it circle. Then they change the beacon, step cost and mud penalty until it delivers.	• The score rose every run. Did the robot do its job? • Look at our guesses on the board. Who predicted the circling?	A rising score is not success. The rule paid more for circling the beacon than for delivering the parcel.
4. Analyze 8 min	Three robots trained on broken rules: one circles the beacon, one dives into the mud, one wanders on a new floor. Learners diagnose at least two from the points on each card.	• What did each rule really pay for? • Why does the Floor 2 robot wander, when the same rule worked on Floor 1?	Each robot did exactly what its rule paid for. On a big floor, a far-off reward is rarely found.
5. Evaluate 8 min	Learners write one rule that delivers on Floor 1 and on Floor 2, which the robot never trained on. They match each loophole to the part that closes it and name one risk left open.	• Which part of your rule stops the circling, the diving and the wandering? • What can your score still not see?	Each loophole is closed by one part of the rule, and some things, like people walking past, the score cannot see.

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

Deepen the thinking

Critical thinking	The robot's score rose every run while it never delivered. Besides the score, what should you watch to know a system works?
AI thinking	If you added a fourth part to the scoring rule, what would it be, and what new loophole might the robot find?
Ethics and fairness	A video app rewarded for watch time learns to keep people watching. Who could be harmed, and who is responsible for that score?
Connect to life	Where have you seen a number, like likes or marks, change how people behave? Did it help the real goal?

Misconceptions to listen for

"The robot is programmed with the route."
Only the scoring rule changes between runs. The robot finds its own route by trial and error over 4,000 practice runs.

"A rising score means it is working."
In Apply, the circling robot's training curve climbs, yet it never delivers the parcel. Watch what it does, not only the score.

"If it works on my test floor, the rule is right."
Rules with free steps deliver on Floor 1 and wander on Floor 2. Test on a floor the robot has never seen.

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

- Game-playing AIs and some robots learn by trial and error from a score, not a route someone wrote.
- A video app rewarded for watch time can learn to keep you watching, which isn't the same as helping you.
- A class chart rewarding "most pages read" can reward skimming: the score shapes what people do, loopholes included.

Try it this week. Invent a points rule for a chore, like tidying up. Then find a loophole: how could someone score without doing the job?

Watch out for. A rising score doesn't prove it's working. The robot scored more each run by circling a bonus, never delivering the parcel.

Talk about it. Think of a rule at school or home that people "game". How would you rewrite it?

Exit ticket

1. Name the five parts of a reinforcement learner.

2. Why did the robot circle the beacon instead of delivering the parcel?

3. Why test a scoring rule on a floor the robot never trained on?

Going further

For a class project, groups invent a points rule for a classroom job, swap rules with another group and hunt for loopholes. Each group then rewrites its rule and names one thing it still cannot measure.

EdSutra

Teaching guide

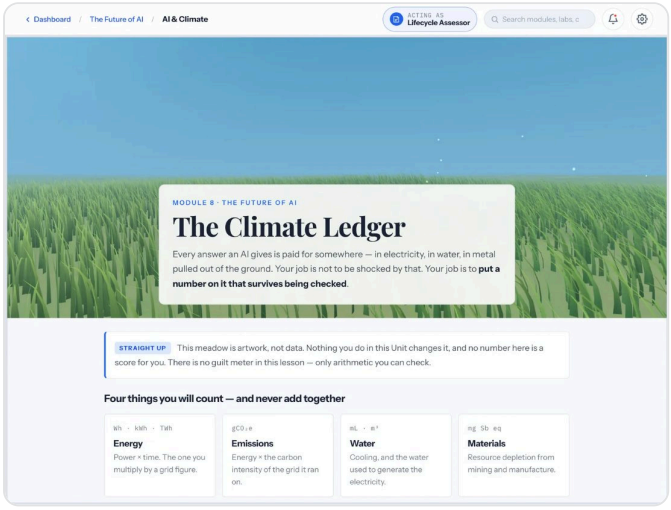
Module 08 · The Future of AI · Unit 02 · Stage 3, Responsibility

AI & Climate

The Climate Ledger

Grades 7–12 | 30 minutes | Role: **Lifecycle Assessor**

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



The big idea

A climate figure for AI means little until you know what it counted: which model, where, when and inside which boundary. Two careful 2025 studies put the water for one AI response 173 times apart, and neither is wrong. Learners do the arithmetic, find where efficiency backfires, and judge a use against a non-AI option.

The learner's job

Learners are the **Lifecycle Assessor** at Meridian Foresight Office · Environmental Accounts. Three teams want a climate number from you. All three will quote it. Only you will know what it counted. Their task: put a number on it. Then show what the number left out.

By the end, learners can say

- I can distinguish power, energy, emissions, water, and materials, and read each with its correct unit
- I can explain how a system boundary, a grid, and the definition of “one prompt” change an estimate
- I can calculate and compare two deployment scenarios from visible assumptions, carrying ranges
- I can trace how rebound, utilization, or a boundary change reverses an apparent advantage
- I can decide whether an AI use is climate-justified against a bounded non-AI baseline, and specify a lower-impact plan

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, so you know where learners will click and which figures they will meet.
- Have calculators or scrap paper ready. The Apply ledgers show every step, and learners can check them by hand.
- Best after How Machines Learn and Reinforcement Learning.
- 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

Power and energy: Power is a rate in watts; energy is power times time, in watt-hours.

Grid carbon intensity: Grams of CO_2e made per kWh of electricity; it differs by country and hour.

System boundary: What an estimate counts, and what it leaves out.

Life cycle: Every stage of impact, from mining and training to your device and disposal.

Embodied impact: Impact from making the hardware, paid before any request is served.

Utilization: How busy hardware is. A fixed making cost is split over its uses.

Rebound effect: Something gets more efficient, so it is used more and the total rises.

Running the lesson 40 minutes

Open 4 min

"You may have heard that an AI chat "drinks a bottle of water". Who counted that, what did they count, and when?"

Before anyone clicks, ask: about how much water does one AI answer use, a drop, a spoonful or a glass? Write the votes on the board. Come back to them in Understand, where two studies disagree 173 times over.

Step	What learners do	Ask the class	Listen for
1. Remember 6 min	Learners tag nine real published figures as power, energy, emissions, water or materials, then build the six stages of an AI system's life cycle in order.	• Why is 700 W a power but 0.24 Wh an energy? • Which life-cycle stage do published figures most often leave out?	The unit tells you what a number measures. Impacts happen in stages, not just inside the app.
2. Understand 7 min	Three predict-then-reveal cases: the same job in France and in India, the same job with a wider boundary, and two 2025 studies 173 times apart on water.	• Why did emissions change when the energy stayed the same? • Look at our water votes. Is one of the two studies lying?	Emissions depend on the grid. The studies measured different models, boundaries and meanings of "one prompt".
3. Apply 9 min	For a cloud assistant and an on-device helper, learners set power, seconds per request, requests per day and the electricity's country. Then they compare the two ledgers, tables showing every step.	• Which single assumption moved your emissions the most? • If the two ranges overlap, which one is lower?	Every figure can be checked by hand. When ranges overlap, the honest answer is "we can't tell from these numbers".
4. Analyze 6 min	Three cases where the option that looks lower loses: a 10 times efficiency gain, a quiet private server and a small model's training. Learners name the lever and the ledger line that proves it.	• Each request got 10 times cheaper. Why did the daily total go up? • Why does a quiet server cost more per request?	Use grew 40 times, so total energy rose. A fixed making cost split over few requests is large.
5. Evaluate 8 min	Three real decisions: a poster image, wind forecasting and an always-on app feature. Learners take a position, pick only evidence that supports it, name who bears the impact and complete a six-line lower-impact plan.	• Which evidence argued for your call, and which against it? • Who bears the impact of your choice, and do they get a say?	Any position can pass. What counts is evidence that supports it, a named bearer and a plan someone can check.

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

Deepen the thinking

Critical thinking	A headline says AI uses "X liters per question". What three questions would you ask before repeating it?
AI thinking	Running a model where and when the grid is cleanest cuts emissions. Which AI jobs could wait for that, and which could not?
Ethics and fairness	A data center draws water and power from one local area, but its users live everywhere. Who should have a say in it?
Connect to life	Think of something that became cheaper or easier, so you used it more. Did your total use go up or down?

Misconceptions to listen for

"Every AI prompt has one true cost."
Understand shows Google and Mistral 173 times apart on water, because they counted different things. Always ask what was counted.

"Making AI more efficient solves the problem."
In Analyze, each request gets 10 times cheaper, but use grows 40 times, so daily energy rises 4 times.

"Emissions belong to the model."
The same job in France and India uses the same energy, yet India's grid makes about 16 times the emissions.

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

- "A bottle of water per chat" headlines trace back to a 2023 estimate for a whole conversation.
- Tech companies publish per-prompt energy and water figures that differ hugely, mostly because they count different things.
- An app can get more efficient per use yet use more energy overall if far more people use it.

Try it this week. Find one claim about AI's energy or water use. With an adult, ask: what was counted, for which system, and when?

Watch out for. A climate figure means little without its units, boundary and date, and energy, carbon and water can't be added into one score.

Talk about it. When is an AI task worth its energy, compared with doing it another way?

Exit ticket

1. What is the difference between power and energy?

2. Give one reason two honest studies can disagree about AI's water use.

3. What is the rebound effect? Give one example.

Going further

For a class project, each learner finds one published claim about AI's energy or water use and checks it: the units, the system boundary, the date and the source. The class then sorts the claims by what they counted.

EdSutra

Teaching guide

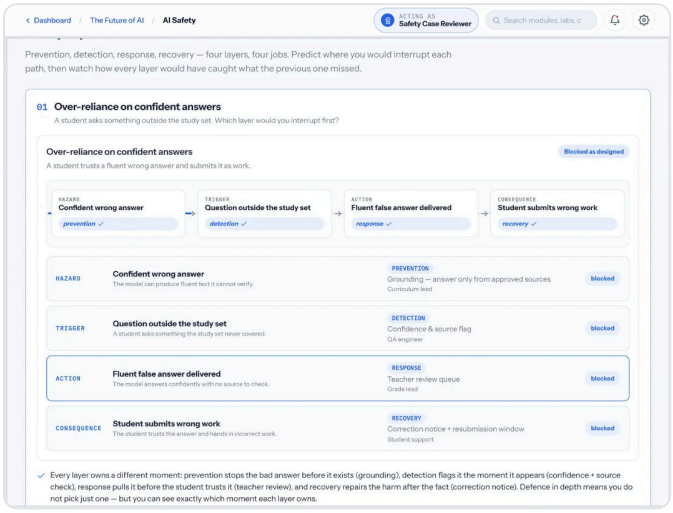
Module 08 · The Future of AI · Unit 03 · Stage 3, Responsibility

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.

EdSutra

Teaching guide

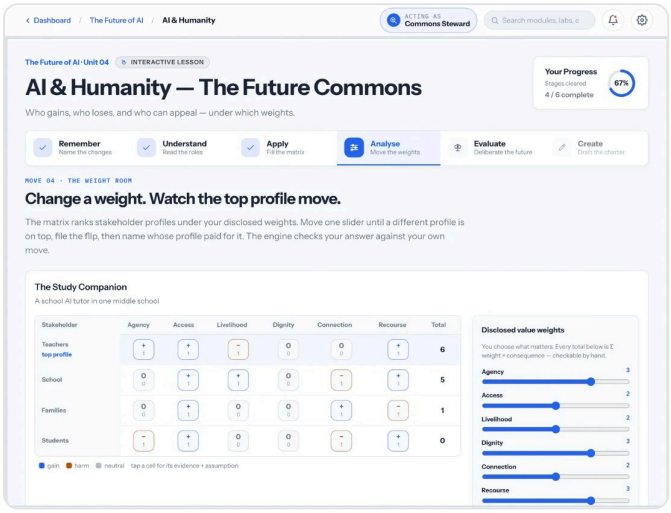
Module 08 · The Future of AI · Unit 04 · Stage 3, Responsibility

AI & Humanity

The Future Commons

Grades 7–12 | 30 minutes | Role: **Commons Steward**

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



The big idea

An AI deployment changes tasks, roles and relationships, and the changes land differently on each person. Which future looks best depends on the value weights you choose, so those weights should be said out loud. Learners show who gains, who loses and who can appeal, then design conditions that keep human choice alive.

The learner's job

Learners are the **Commons Steward** at The Future Commons · Community Table. The town has three futures for its AI tutor and its clinic queue. Your job is not to pick a winner — it is to make sure nobody can claim a future without showing its trade-offs. Their task: hold the table: show who gains, who loses, and who can appeal — under weights you disclose.

By the end, learners can say

- I can distinguish automation, augmentation, substitution, agency, dependency, access, and recourse
- I can predict how one AI deployment changes a task and a human role, then compare stakeholder perspectives
- I can fill a consequence matrix for an authored scenario using provided evidence and explicit assumptions
- I can change value weights and identify which recommendation flips, why, and who bears the cost
- I can deliberate among plausible futures and defend adoption conditions using evidence, distribution, and reversibility
- I can draft a Human Choice Charter: protected decisions, disclosure, opt-out, skill preservation, access, appeal, review triggers

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, so you know how the table of gains and harms reads.
- Pairs work well in Analyze and Evaluate: one learner moves the weights, the other says whose total paid for it.
- Best after AI & Climate and AI Safety.
- 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

Automation: A task is done by the system instead of by a person.

Augmentation: A person's work changes, but the person stays central.

Substitution: A person leaves the loop entirely.

Dependency: Reliance on a system that is hard to reverse.

Recourse: A way to challenge a result and reach a human with authority.

Stakeholder: Anyone a decision affects, such as teachers, families or patients.

Value weights: How much each value, like agency or access, counts in a total.

Running the lesson 40 minutes

Open 4 min

"A school gets an AI tutor that grades homework. Is that good news? Good news for whom?"

Before anyone clicks, ask: when the tutor starts grading, what happens to the teacher's job? Does it disappear, shrink and change, or stay the same? Take a vote and come back to it in Understand.

Step	What learners do	Ask the class	Listen for
1. Remember 4 min	Learners match seven plain-language examples to seven concepts: automation, augmentation, substitution, agency, dependency, access and recourse.	• What is the difference between automation and substitution? • Which example showed a dependency forming?	Automation takes over a task; substitution takes the person out of the loop. They are different changes.
2. Understand 6 min	Learners predict how the Study Companion changes the teacher's role and what its biggest new dependency is, then tap each seat at the table to read its gains and harms.	• Look at our vote. Did the teacher's job disappear, or change? • Why does one tutor look different from each seat?	Grading shrank and coaching grew. Each group gains on some values and loses on others.
3. Apply 7 min	Learners attach the right evidence cards to each of the four stakeholder rows. Some cards, like a vendor's marketing sheet, are distractions.	• Why doesn't the vendor's marketing sheet count as evidence? • Which consequence rests on an assumption, with no evidence at all?	Evidence must be an observation about that group. A seller's claim or a training curve is not.
4. Analyze 7 min	Learners raise or lower one value weight until a different stakeholder's profile, its weighted total, is on top. Then they name whose profile fell the most.	• Which weight did you move, and who came out on top? • Whose profile paid for your move?	Changing a weight is a choice that moves the cost onto someone, and the totals show who.
5. Evaluate 6 min	A clinic considers an AI triage assistant. Learners choose auto-triage, a nurse-confirmed hybrid or a manual queue, back it with at least two evidence chips that fit, and name who bears the cost.	• What did your future gain, and what did it give up? • Who bears the cost of your choice?	No future is correct. A good defense uses evidence that fits the choice and names who pays.
6. Create 6 min	Learners draft a Human Choice Charter for the Study Companion, choosing one condition for each of seven slots, from protected decisions and opt-out to appeal and review.	• Which decision did you keep human-only, and why? • Which slot would you fight hardest to keep?	Keeping human choice is a design decision: named conditions that someone can check and enforce.

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

Deepen the thinking

Critical thinking	In this Unit the weights were shown. Where in real decisions are the weights hidden, and how could you find them?
AI thinking	If an AI tutor helps you every evening, how could you tell whether it is augmenting you or building dependency?
Ethics and fairness	Auto-triage serves more patients but drops the human check. Who should get the final say: patients, nurses or the clinic's managers?
Connect to life	Think of a self-checkout or an automated phone line. Could you reach a person when something went wrong?

Misconceptions to listen for

"Automation and substitution are the same thing."
In Understand the tutor takes over grading, but the teacher stays and coaches more. The task left; the person did not.

"One AI future is objectively best."
In Analyze, moving one weight puts a different group on top. Which future looks best depends on the weights.

"If it is efficient, it is good for everyone."
The table shows the school's efficiency gain beside lost peer help and connection. A gain on one value is not a total gain.

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

- Self-checkouts automate a task; whether a staff member is still there to help is a separate choice.
- A study app that helps you learn augments you; one that does the work for you can build dependency.
- If an automated system blocks your account, a way to appeal to a person is called recourse.

Try it this week. Pick one change technology brought to your school or home. List who gained, who lost something, and who was never asked.

Watch out for. Efficient isn't the same as good for everyone. Which future looks best depends on what you weigh most, so say your weights out loud.

Talk about it. Which decisions about your life should always stay with a person, never a machine?

Exit ticket

1. What is the difference between automation and augmentation?

   _____

2. What is recourse, and why does it matter?

   _____

3. Why did changing one weight change who came out on top?

   _____

Going further

For a class project, groups write a Human Choice Charter for a real tool the school uses, such as a learning app: one protected decision, an opt-out, an appeal route and a review trigger.

EdSutra

Teaching guide

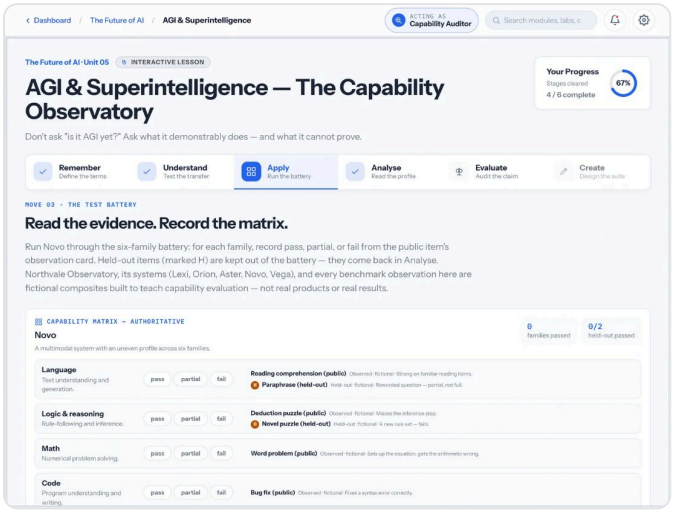
Module 08 · The Future of AI · Unit 05 · Stage 3, Responsibility

AGI & Superintelligence

The Capability Observatory

Grades 7–12 | 30 minutes | Role: **Capability Auditor**

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



The big idea

"Is it AGI yet?" cannot be answered with one score. Learners ask what a system demonstrably does, across which task families, and whether it transfers to tasks it never saw. A high public score can come from memorized items or a borrowed tool. A fair claim says what was tested, what failed and what is still unmeasured.

The learner's job

Learners are the **Capability Auditor** at Northvale Observatory · Capability Desk. Everyone wants to know if a system is 'approaching AGI.' You cannot answer with a number. You can answer with a matrix: what it passes, what it fails, what transfers, and what the test still cannot prove. Their task: read the matrix, check the transfer, and audit the claim — counts, not percentages.

By the end, learners can say

- I can distinguish narrow AI, generality, transfer, autonomy, AGI, and superintelligence, and separate definitions from predictions
- I can predict whether a high score in one domain transfers to a changed task
- I can run authored systems through a transparent multi-domain test battery and record pass / partial / fail with evidence
- I can inspect spiky capability profiles and diagnose memorization, tool dependence, brittle transfer, or missing embodiment/context
- I can audit a public-style 'approaching AGI' claim for definition, evidence breadth, counterexamples, and uncertainty
- I can design a capability test suite with task diversity, hidden transfer cases, failure criteria, and a limits statement

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, so you know how the capability matrix shows public and held-out items.
- Tell learners that Northvale, Lexi, Orion, Aster, Novo and Vega are invented, not real products or real results.
- Best after How Machines Learn and AI & Humanity.
- 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

Narrow AI: A system that does one task well, like a spam filter.

Generality: Performing well across many different families of tasks.

Transfer: A skill from one task that still works on a changed task.

AGI: Matching human intellectual breadth: a contested definition, not a switch.

Benchmark: A standard set of tasks used to compare systems.

Held-out trial: Test items kept private, so the system cannot have seen them.

Contamination: Test items leaking into training data, so a high score proves less.

Running the lesson 40 minutes

Open 4 min

"A headline says an AI "scored near-human on 40 benchmarks" and is "approaching AGI". What would you need to see before you believed it?"

Before anyone clicks, ask: if a system aces a reading test, will it still answer when the same question is reworded? Take a vote and come back to it in Understand, when the held-out evidence opens.

Step	What learners do	Ask the class	Listen for
1. Remember 4 min	Learners match seven terms, from narrow AI and transfer to AGI and superintelligence, then sort two statements into a definition and a prediction.	• Is "AGI will arrive before 2040" a definition or a prediction? • Why does it matter which one it is?	A definition says what a word means. A prediction is a dated claim about the future that evidence can't check yet.
2. Understand 6 min	Three systems ace a public task. Learners predict whether each transfers to a changed task: a reworded question, a missing calculator, a moved obstacle. Then the held-out evidence opens.	• Look at our vote. Did the reading system pass the reworded question? • Why do testers keep some items hidden?	None of the high public scores transferred. Hidden, held-out items show what a system can really do.
3. Apply 6 min	Learners run the fictional system Novo through a six-family test battery, recording pass, partial or fail for each family from its public observation card.	• How many of the six families did Novo fully pass? • Is Novo general? What else would you need to know?	Novo fully passed only language and code. Strong spots do not add up to generality.
4. Analyze 6 min	Four uneven profiles. Learners compare public and held-out items and diagnose at least two: memorization, tool dependence, brittle transfer, or missing embodiment and context.	• What gave away Lexi's memorization? • Should Orion be tested with its tools, without them, or both?	Passing public items but failing held-out ones points to memorization. Report a system with and without its tools.
5. Evaluate 7 min	Learners audit a headline saying Vega is approaching AGI after near-human scores on 40 benchmarks. Four steps: its definition, how broad its evidence is, the strongest counterexample and an honest uncertainty statement.	• What kind of tasks were all 40 benchmarks? • Which counterexample hurts the claim most?	No definition is stated, all 40 are language sets, and Vega failed a held-out logic puzzle.
6. Create 7 min	Learners design their own test suite: at least four task families, at least one hidden transfer case from a family they chose, failure criteria and a statement of what the suite cannot prove.	• Which hidden case did you add, and what would a failure tell you? • What can your suite never prove?	A good suite is broad, includes held-out items, says what counts as failing and states its limits.

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

Deepen the thinking

Critical thinking	Two experts disagree about whether a system is close to AGI. How could different definitions explain the disagreement?
AI thinking	Orion only solves math with a calculator attached. Is the calculator part of its ability? How should that be reported?
Ethics and fairness	If a company overstates what its AI can do, who could be harmed, and who should check the claim?
Connect to life	Think of a test you passed by memorizing. How would a held-out question have shown what you really knew?

Misconceptions to listen for

"One high score means the system is general."

Lexi aces public reading items but fails a reworded question it never saw. Only held-out trials show transfer.

"AGI is a single switch that flips on."

The Unit treats AGI as a contested definition. Which definition you use changes how a system is judged.

"A smarter system must be conscious."

The Unit's tests measure capability only. A suite's limits statement keeps consciousness and autonomy as separate questions.

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

- Headlines saying an AI "aced an exam" don't show it can handle new or changed tasks.
- Predictions about when AGI will arrive are scenarios with dates, not facts, however confident they sound.
- A chatbot that writes great essays can still fail simple puzzles, or need a calculator tool for sums.

Try it this week. Pick a game you're good at and play it with one new rule. How much of your skill transferred?

Watch out for. One high score doesn't make a system general: it may have seen the test questions in training. And capable doesn't mean conscious.

Talk about it. What would an AI have to do to convince you it was truly general? What wouldn't count?

Exit ticket

1. What is the difference between narrow AI and generality?

   _____

2. Why do testers use held-out items?

   _____

3. Name one reason a high benchmark score might not show real ability.

   _____

Going further

For a class project, groups find a news story about an AI "acing" a test and audit it like the Vega claim: the definition used, the breadth of tasks, any counterexample and what is still unmeasured.