EdSutra

Teaching guide

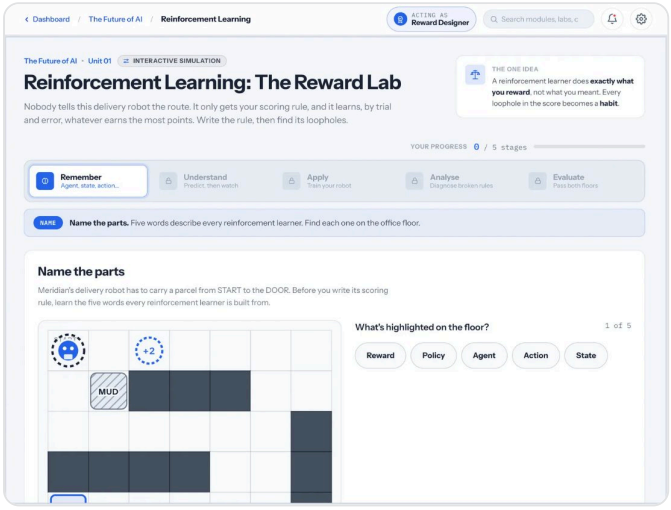
Module 08 · The Future of AI · Unit 01 · Stage 3, Responsibility

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.