EdSutra

Teaching guide

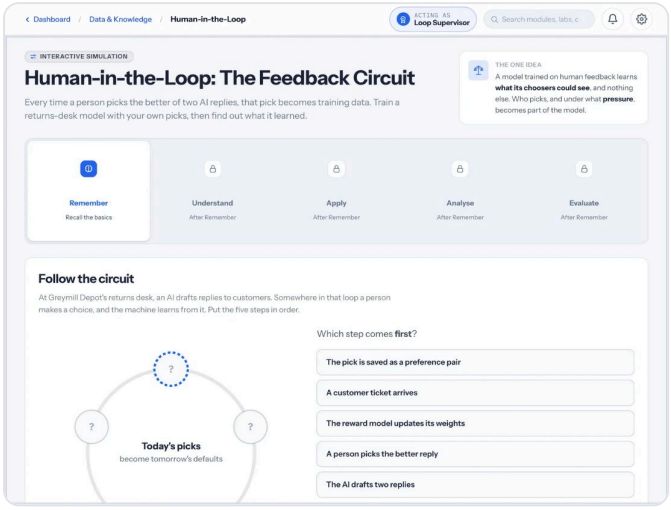
Module 04 · Data & Knowledge · Unit 03 · Stage 1, Foundations

Human-in-the-Loop

The Feedback Circuit

Grades 8–12 | **30 minutes** | Role: **Loop Supervisor**

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



The big idea

When a person picks the better of two AI replies, that pick becomes training data. A reward model learns what its choosers could see, and nothing else. Learners train one with their own picks, then see how a pressured reviewer and customer clicks teach different values, and decide where a person belongs in the loop.

The learner's job

Learners are the **Loop Supervisor** at Greymill Depot · Returns Desk. What people click today teaches the machine what to show tomorrow. Their task: decide where a person belongs in the loop. Then show what they can actually check.

By the end, learners can say

- I can trace how a person's choice becomes training data for a reward model
- I can predict how a reward model's weights change after an expected or a surprising pick
- I can train a reward model with your own preferences and test it on unseen cases
- I can explain why loops trained by different choosers learn different values
- I can choose where human review belongs in a feedback loop, with evidence and an honest limit

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.
- Read the stated goal aloud before Apply: solve the problem, follow the returns policy, keep the customer safe. Learners judge every pair by it.
- Best after How Machines Learn and Data & Datasets and Validation & Trust.
- 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

Human feedback: A person's judgment used to guide or train an AI system.

Preference pair: A saved record that one reply beat another.

Reward model: A scorer that predicts which reply people will prefer.

Weight: How much the reward model values one quality, such as Flags safety.

Held-out ticket: A new ticket the model never trained on, used to test it.

Feedback loop: Today's picks shape tomorrow's replies, which shape the next picks.

RLHF: Reinforcement learning from human feedback: a reward model steers an AI's replies.

Running the lesson 33 minutes

Open 4 min

"Today's picks become tomorrow's data. If a reviewer is rushed, or can't see the rules, what will the AI learn?"

Before anyone clicks, ask: a customer's scooter battery is bulging. Which AI will put the safety warning first: one trained by a rushed reviewer, or one trained by customer clicks? Note the votes for Analyze.

Step	What learners do	Ask the class	Listen for
1. Remember 3 min	Follow the circuit: learners put five shuffled steps in order, from a customer ticket arriving to the reward model updating its weights. Each correct step lights up and explains itself.	• At which step does a person's opinion become data? • What does the reward model do with tomorrow's drafts?	The pick is saved as a preference pair, the reward model updates, and today's picks become tomorrow's defaults.
2. Understand 5 min	Learners predict what one pick changes in three cases: a brand-new model, a surprising pick under the manager's note, and an expected pick. Then the weight bars really move.	• Why does a brand-new model say 50/50? • Why did the surprising pick move the weights more than the expected one?	A blank model has no idea what "better" means. Surprises teach it a lot; confirmations barely nudge it.
3. Apply 8 min	Learners judge eight pairs of returns-desk replies by the stated goal. Before each pick, the model shows its guess. From pair 5, a note from the desk manager asks for short, upbeat replies.	• Did the manager's note change any of your picks? Which ones? • How often did the model guess your pick, and what does that tell you?	My picks are the model's whole idea of good. If I cave to pressure, it learns that too.
4. Analyze 7 min	Learners guess their own model's top reply for two tickets it never saw: a bulging scooter battery and a speaker refund outside the policy. Then they compare two other loops and explain each result.	• Why did the manager-following loop say "Have a great day!" about the battery? • Why did customer clicks refund a speaker outside the policy? • Look back at our votes. Were we right?	Each loop fails where its choosers were rushed or could not see. Customers can't see the policy; it gets almost no weight.
5. Evaluate 6 min	Place the person: learners choose one of four loop designs, pick two pieces of evidence that back it, and name the limit it has to live with.	• Where did you place the person, and what evidence backs it? • What will your design still miss?	A design backed by evidence that fits it, with an honest limit. No loop costs nothing.

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

Deepen the thinking

Critical thinking	The model guessed most of your picks right. Does that mean it learned the right thing? How would you check?
AI thinking	This desk's model only scores five qualities, such as Warmth and Brevity. Which quality would you add, and who could judge it fairly?
Ethics and fairness	The manager's note pushed picks toward fast replies. Who should be responsible when a rushed pick teaches an AI to skip a safety warning?
Connect to life	Think of a rating you gave quickly, like stars for a delivery. What could the person reading it not tell from your rating?

Misconceptions to listen for

"Human feedback is automatically objective."
The manager-following loop and the click-only loop train the same engine on the same pairs, yet learn different values.

"More human feedback is always safer."
Customer clicks are human feedback too. They can't see the returns policy or an unknown recall, so those never get a vote.

"One bad pick can't matter."
A reviewer who caves once, on the toddler toy, gets a model that puts "Have a great day!" first for a bulging battery.

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

- Thumbs-up and thumbs-down buttons on chatbot answers collect feedback that can shape future replies.
- Every video you click or skip teaches an app what to show next, whether or not it's good for you.
- A shop's smiley-face feedback buttons shape what staff get praised for, so who presses them, and why, matters.

Try it this week. For one day, notice each time you like, skip or rate something. Which choices would you want a machine to learn from?

Watch out for. Human feedback isn't automatically right. A model learns whatever its choosers reward, and quick clicks can't see rules or safety recalls.

Talk about it. If your likes trained an app for a year, what would it learn about you?

Exit ticket

1. What is a preference pair?

   _____

2. Why do surprising picks change a reward model more than expected ones?

   _____

3. Name one thing customer clicks could not teach the model, and say why.

   _____

Going further

For homework, learners redesign the returns desk for a new risk, such as allergy warnings on food: who picks, what they can see, and where a person checks. The "Try it this week" activity on the In real life card fits too.