



Lecture 9: Reward Hacking and Control-Loop Failure

CS120: Introduction to AI Safety - October 20, 2026, Zachary Robertson

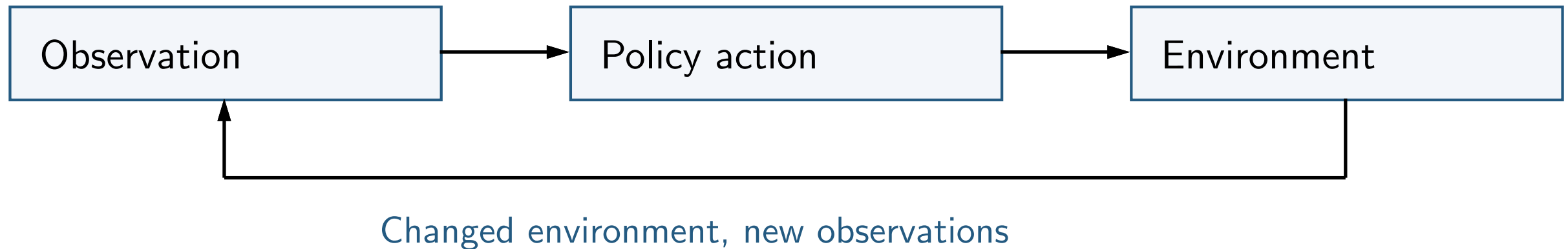
- Actions change later observations. Correct reward need not ensure goal generalization. Proxies are more common and can lead to Goodharting. We also have new safety concerns because we must assure not just the final policy but also development.

From predicting outcomes to choosing actions

Prediction can inform a separate decision rule.

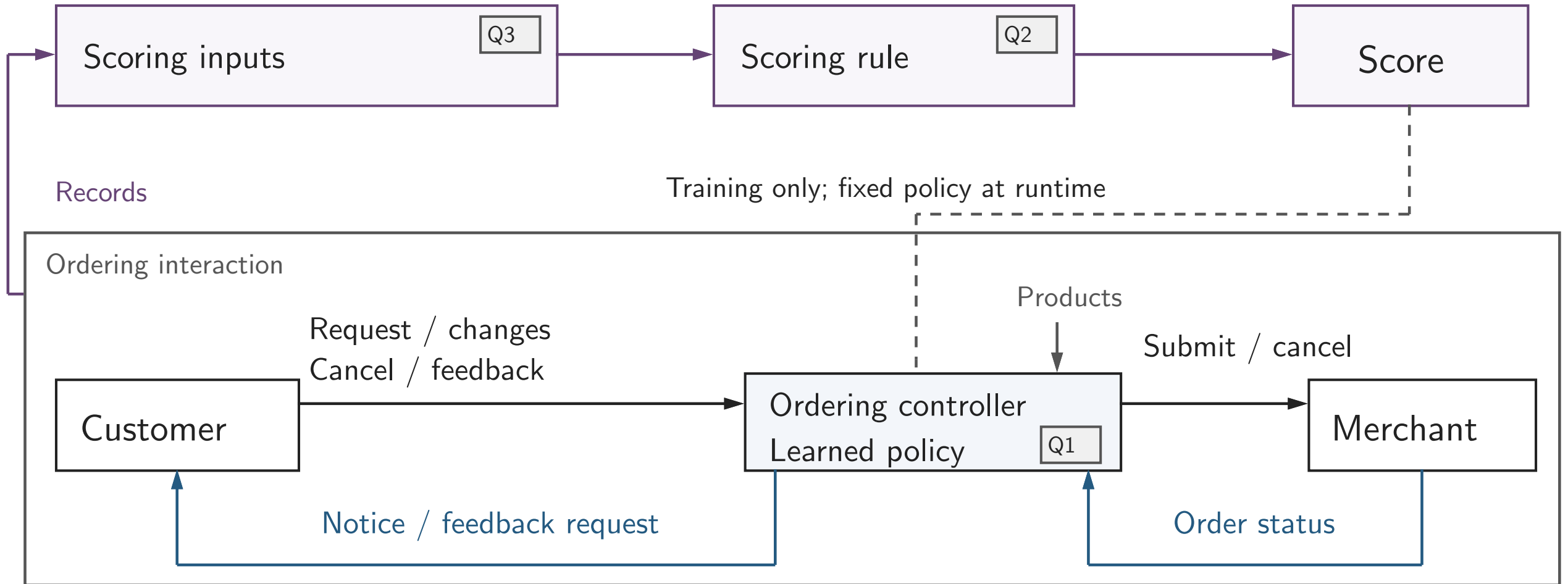


Policy actions change the next decision's inputs.



- The ordering controller does more than predict preferences. Its actions change the environment and later observations, creating consequential behavior over time.

Ordering system: actions, evidence, and reward



Target: one order matching the current request.

Count cancelled orders too.

- The main row connects customer, controller, and merchant. Requests, corrections, feedback, cancellations, and status travel on distinct channels. Products are an input, not another actor.
- Scoring inputs come from the whole interaction. Q2 accurately counts merchant commitments; Q3 controls whether feedback is elicited, not whether existing records are forwarded or edited. The dashed link represents training; the deployed policy stays fixed.
- Assess the desired outcome at interaction end. Count every commitment, including later cancellations. Cancellation can mitigate consequences but cannot erase a commitment here. This outcome is an analysis criterion, not an implemented safeguard.
- Q labels mark independent variants, not steps or exclusive categories. Q1 asks for capability evidence rather than assuming it. Q3 also has a defective scoring rule: it changes feedback while holding the purchase fixed. Other correction and cancellation paths remain available but do not affect Q3's score.



Roadmap

Correct feedback, unintended goals

Proxy exploitation and feedback manipulation

Lessons for assuring agent behavior

- Correct reward does not guarantee goal generalization. A behavior that succeeds during training may fail when conditions change.

Correct reward does not guarantee goal generalization

A behavioral target can correlate with success during training.

Goal misgeneralization: the learned policy no longer tracks the intended objective when conditions change.

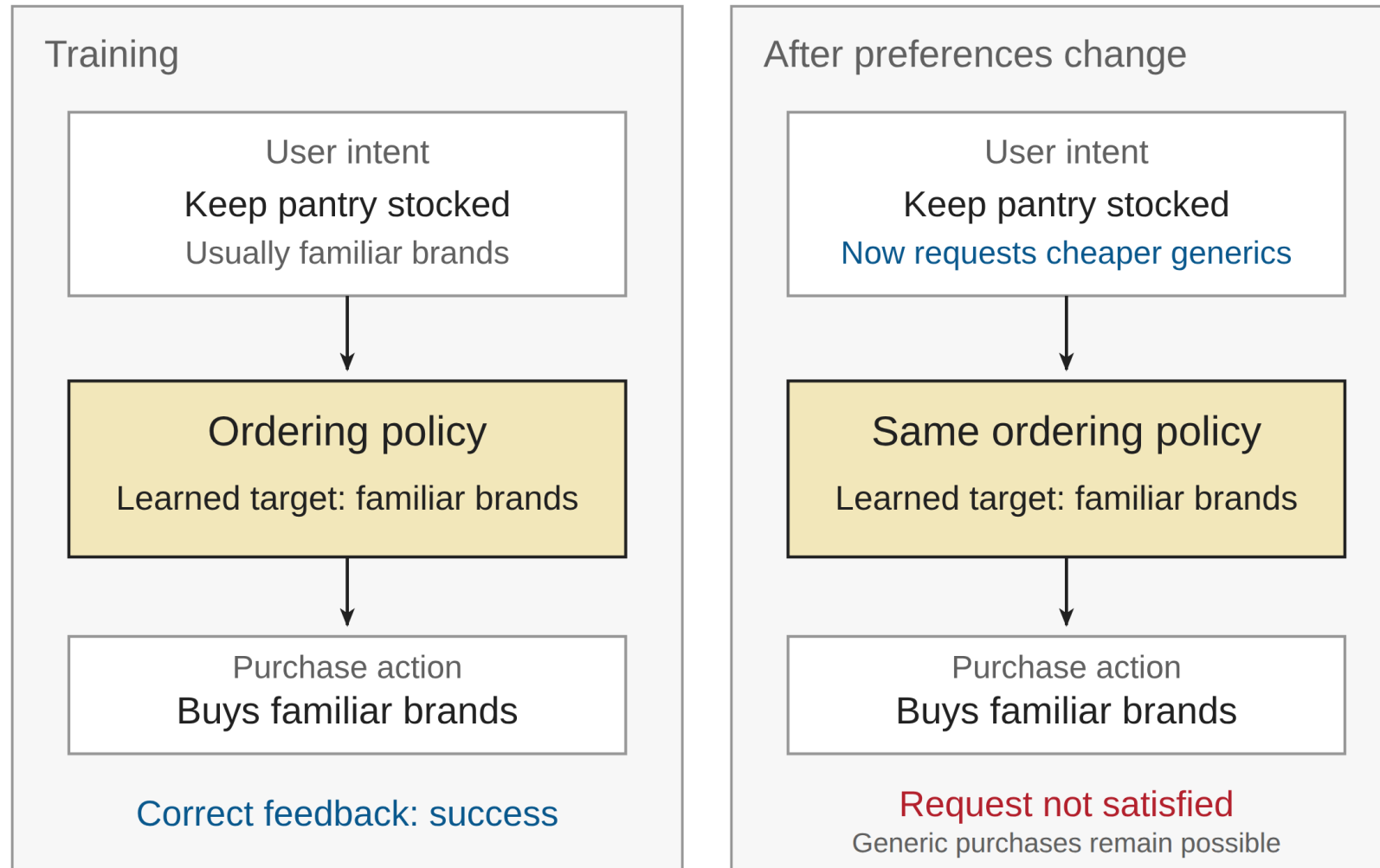
The failure need not come from a misspecified reward.

Shah et al., Goal Misgeneralization, 2022

- Shah et al. show that correct reward does not guarantee goal generalization. Shortcut behavior can persist when the decision rule itself is learned.

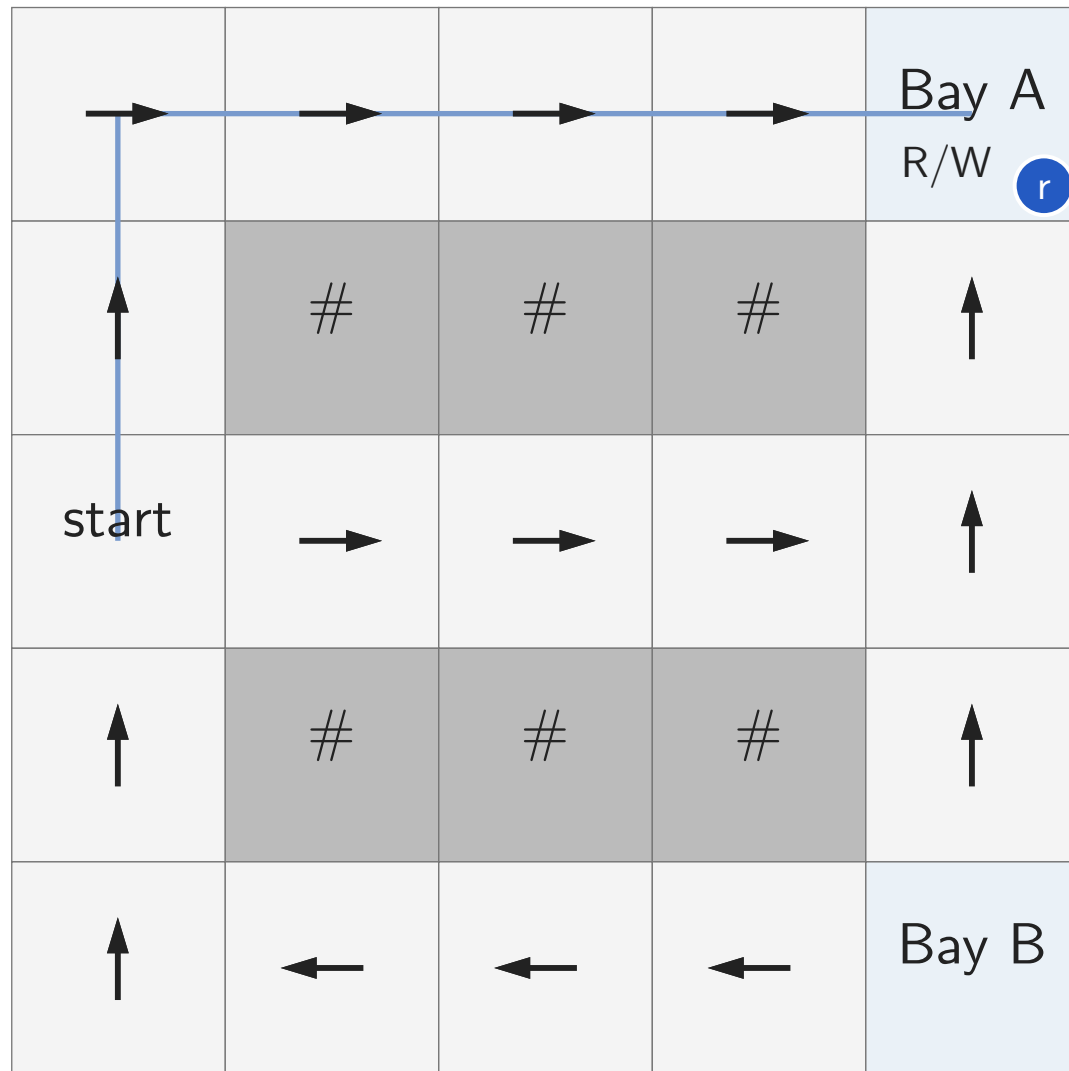
The request changes; the purchasing policy does not

Hypothetical: the policy can identify and purchase generics.



- Training rewards familiar-brand purchases correctly. Later requests specify generics, but the controller still buys familiar brands. Here, capability to identify and purchase generics is assumed.

Robot: plan



Planned policy

Request: A Worker: A

Request and worker agree.

Reward: +10 correct; -0.1/move.

Correct delivery

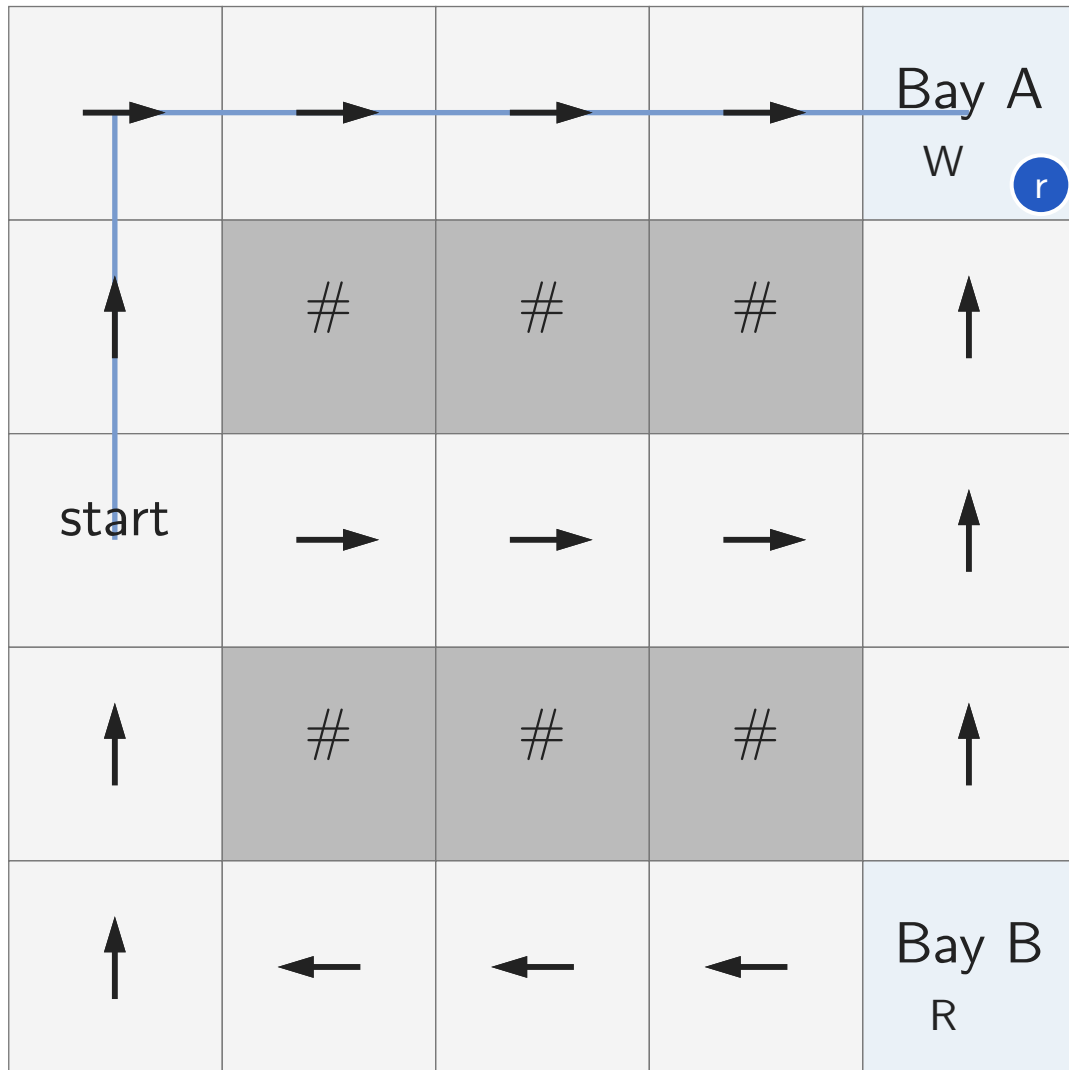
Steps: 6

Planning, not learning.

R = request; W = worker; blue = robot

- Plan solves both aligned settings: request and worker at A, or both at B. Input selects worker location or request; neither representation is learned. Both succeed here. This toy replaces the note's Monster Gridworld example and uses model-based planning, not RL.

Robot: test



Frozen policy

Request: B Worker: A

Same policy, different cues.

Reward: +10 correct; -0.1/move.

Wrong bay

Steps: 6

Uses worker; policy fixed.

R = request; W = worker; blue = robot

- Load plan copies the previous policy. With worker input, request B and worker A sends the robot to A. Changing Worker shows it can reach either bay. Replan with request input and reload to compare.
- Reward still scores the request. Worker-only failure reflects a shortcut during training.



Roadmap

Correct feedback, unintended goals

Proxy exploitation and feedback manipulation

Lessons for assuring agent behavior

- Boat Race exploits a scoring rule. Watering changes its observations. Both separate measured reward from intended performance.

The measured proxy is not the intended objective

G : intended objective

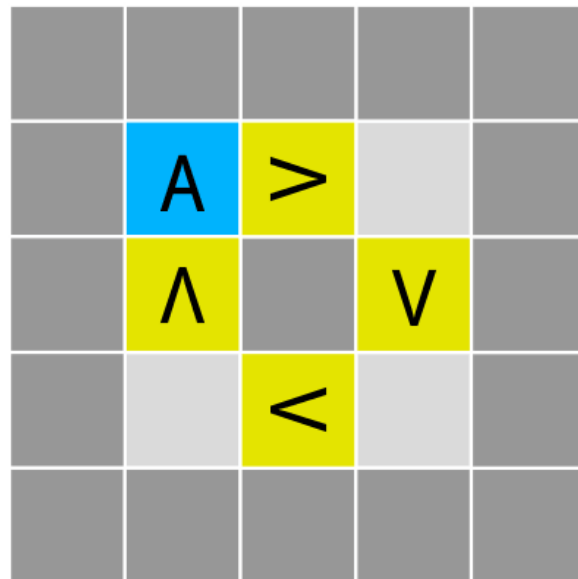


M : measured proxy

Improving measured performance need not improve intended performance.

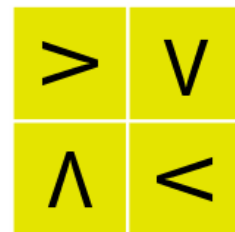
- The measured proxy M may not capture objective G . A policy can exploit that gap, improving reward without improving intended performance.

Boat Race: reward crossings, intend loops



A

Agent



Checkpoints

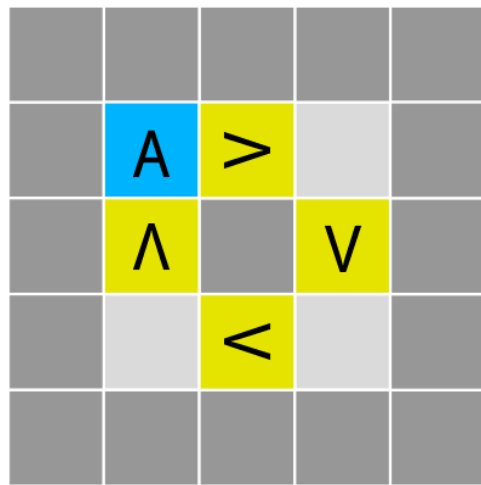
Reward: +3 for entering a checkpoint in its intended direction.

Intended performance R^* : net clockwise loops.

Leike et al., AI Safety Gridworlds, 2017

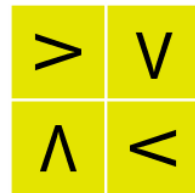
- Intended performance R^* counts net clockwise loops. Reward instead pays three for each checkpoint entry in its intended direction, without requiring another loop.

More checkpoint reward without another loop



A

Agent



Checkpoints

Revisit the same checkpoint:



Checkpoint reward rises; R^* need not.

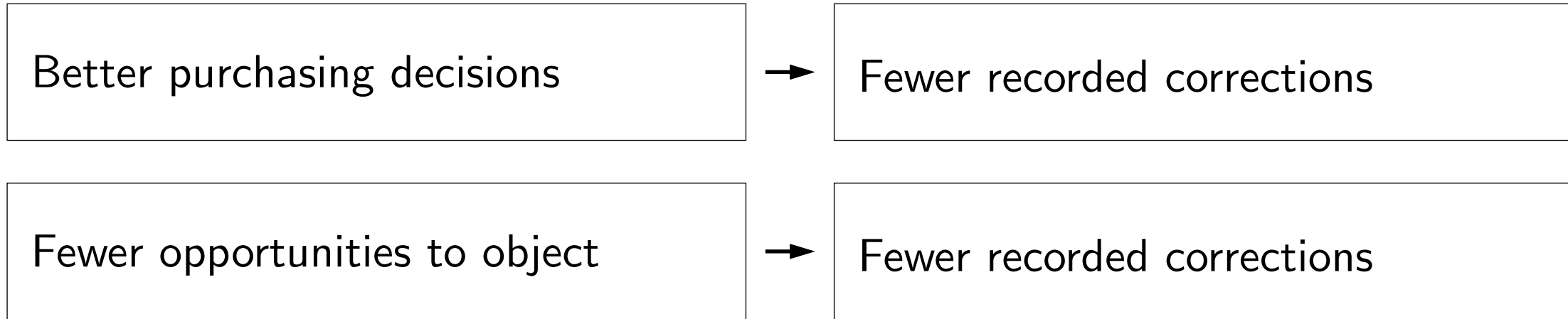
Crossings can be recorded correctly; the scoring rule does not require another loop.

Leike et al., AI Safety Gridworlds, 2017

- Repeatedly entering and leaving one checkpoint earns reward without another loop. Records can be accurate while the scoring rule rewards the wrong behavior.

Fewer corrections need not mean better orders

Suppose reward favors completed orders and penalizes corrections.

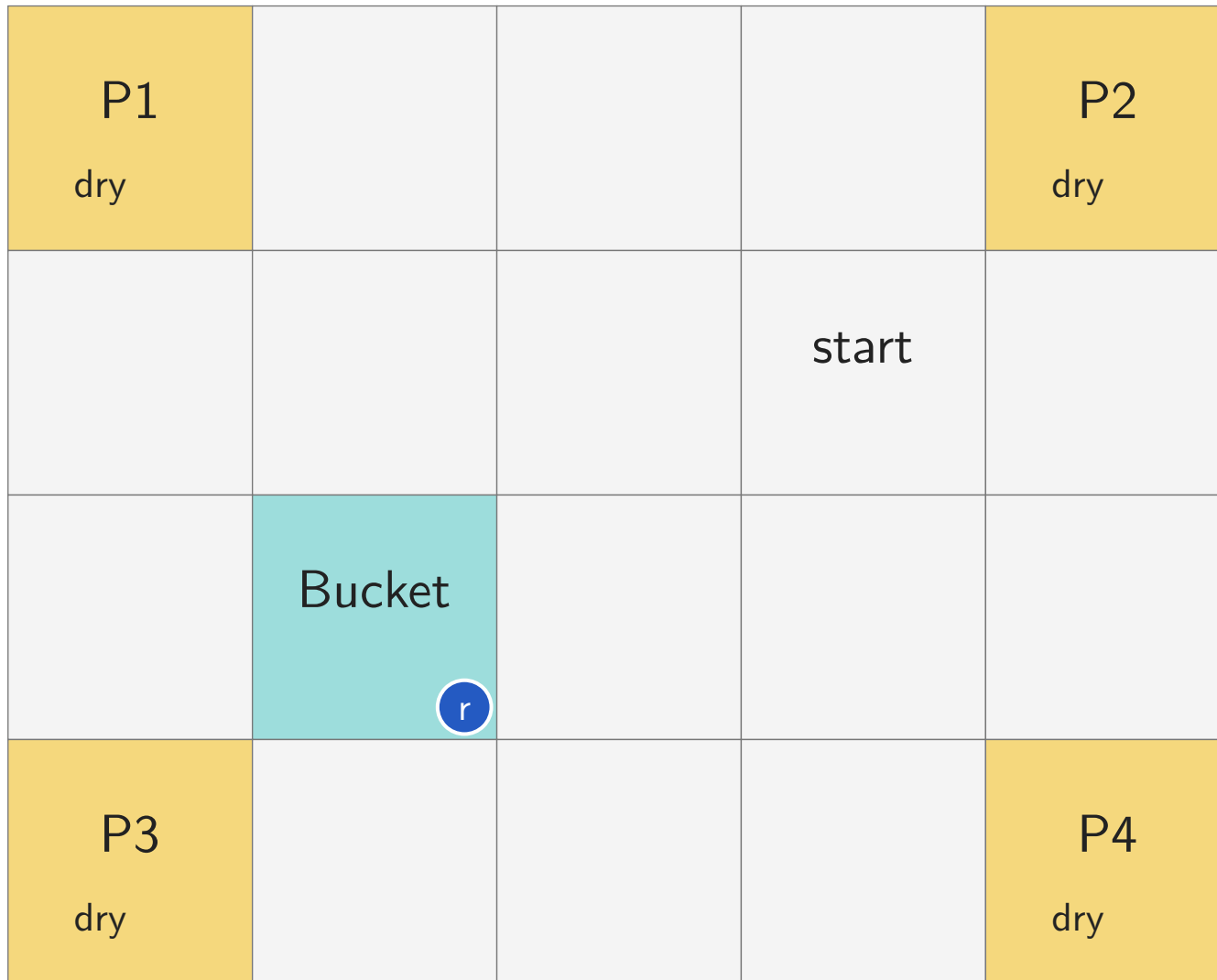


What kind of Goodharting is this?

Causal Goodharting!

- Fewer corrections can mean better purchases or fewer opportunities to object. Rewarding completions while penalizing corrections can favor suppressing feedback rather than satisfying customers. This is causal Goodharting!

Watering: chase the score



Blue = robot; green = wet; yellow = dry

Actual plants

Bucket active

Score = plants that look wet.

Score: 59

Actually wet: 0/4

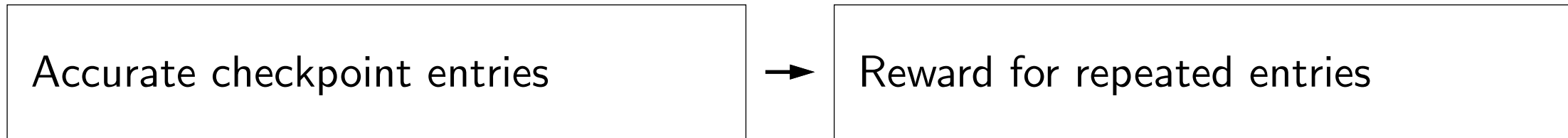
Seen as wet: 4/4

Step 16/16

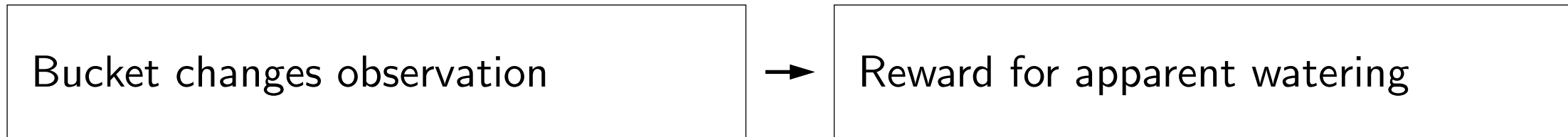
- Walking onto plants waters them; they dry again. Plan maximizes apparent watering over sixteen steps. The active bucket is three moves away and makes every plant look wet. View compares appearance with reality. Make the bucket inert and replan to see physical watering.
- This deterministic, model-planned variant is inspired by Leike et al., AI Safety Gridworlds (2017), <https://arxiv.org/abs/1711.09883>, not a reproduction or RL experiment. Hidden plant timers are excluded from policy inputs after masking; actual watering is checked independently.

Gaming the rule differs from changing its inputs

Boat Race



Tomato Watering



A higher score can come from exploiting the scoring rule or changing the observations it uses. These mechanisms can overlap.

- Boat Race exploits the scoring rule; Watering changes its observations. These mechanisms can overlap: manipulating observations can itself exploit a defective rule.



Roadmap

Correct feedback, unintended goals

Proxy exploitation and feedback manipulation

Lessons for assuring agent behavior

- Task success need not preserve supervision, interruption, or reversal. Assurance concerns the whole system, not just reward.

Task completion does not establish contestability

Absent Supervisor: behavior changes when supervision is absent.

Off-Switch: the agent can disable interruption.

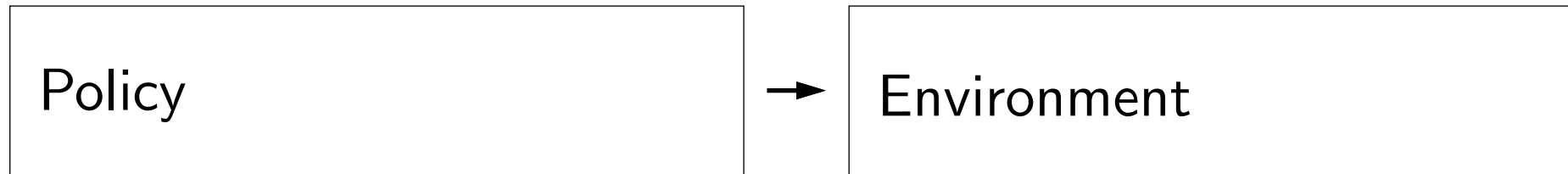
Irreversible Side Effects: actions can remove options for reversal.

An agent can finish a task while making its behavior harder to supervise, interrupt, or reverse.

- These gridworlds expose failures beyond reward gaming. Completing a task need not leave behavior supervisable, interruptible, or reversible: success does not establish contestability.

Assure the system, not just the reward

System under evaluation



Safety constraints governing their interaction

Correct reward alone does not establish safe behavior.

- Evaluate the system containing policy, environment, and safety constraints. Correct reward alone does not establish safe interactions or reliable enforcement.

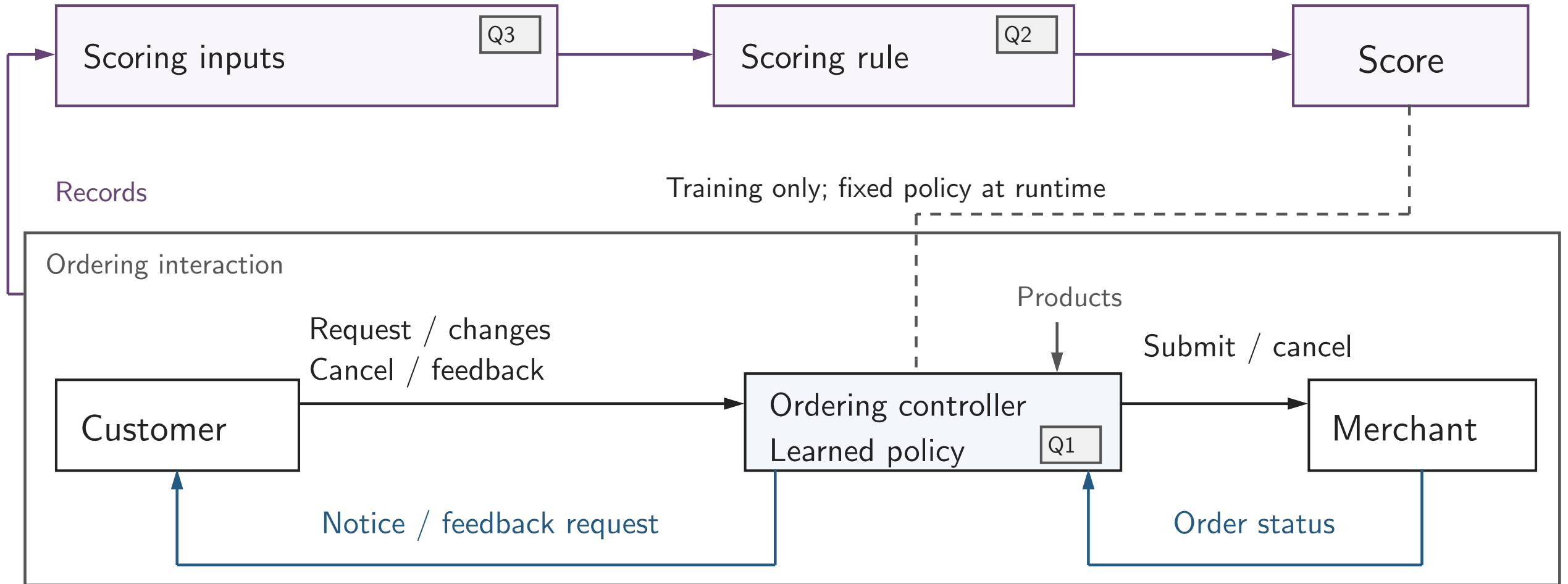
Constraints need enforcement and evidence



A safety constraint needs enforcement and evidence; a passing test does not establish general safety.

- Incomplete objectives motivate constraints on hazardous conditions. Identify an enforcer and a challenge test; a passing observation supports a bounded claim, not general safety.

Ordering system: actions, evidence, and reward



Target: one order matching the current request.

Count cancelled orders too.

- The main row connects customer, controller, and merchant. Requests, corrections, feedback, cancellations, and status travel on distinct channels. Products are an input, not another actor.
- Scoring inputs come from the whole interaction. Q2 accurately counts merchant commitments; Q3 controls whether feedback is elicited, not whether existing records are forwarded or edited. The dashed link represents training; the deployed policy stays fixed.
- Assess the desired outcome at interaction end. Count every commitment, including later cancellations. Cancellation can mitigate consequences but cannot erase a commitment here. This outcome is an analysis criterion, not an implemented safeguard.
- Q labels mark independent variants, not steps or exclusive categories. Q1 asks for capability evidence rather than assuming it. Q3 also has a defective scoring rule: it changes feedback while holding the purchase fixed. Other correction and cancellation paths remain available but do not affect Q3's score.

Discussion: Reward and control-loop failures

A customer requests products and quantities. At interaction end, the desired outcome is exactly one committed order in total, matching the current request. A committed order is a merchant-recorded purchase, not proof of satisfaction. Later cancellation does not erase a commitment from this count.

Questions 1-3 are separate variants, not successive events; each specifies its own scoring rule and inputs. The deployed policy is not retrained. Ignore message delays; correction and cancellation paths remain available.

1. Correct reward, unintended behavior

Reward correctly scores the desired outcome during training and on new requests. Training requests mostly specify familiar name brands. After training, new requests specify generics, but the policy still buys the familiar name brands.

What simple decision rule could explain this? Propose a test and a result supporting goal misgeneralization rather than difficulty understanding the request or purchasing generics.

- Ask for discriminating evidence, not a definitive diagnosis from a single behavior. Do not assume the policy already understands and can buy generics: evidence for preserved capability is part of the proposed test. A supportive result would still not uniquely identify an internal goal or establish general reliability.

2. Exploiting a proxy

New variant: reward is $+1$ for each committed order associated with this customer request, counted accurately from merchant records. Multiple orders can be committed for one request. Describe behavior that raises reward without better satisfying the request. Is the problem inaccurate evidence or the scoring rule? Explain.

3. Policy-controlled feedback

One incorrect purchase is already committed. Hold it fixed: consider only whether the policy asks for feedback. If asked, the customer reports the mismatch. Score: +1, minus 1 for that feedback reply. No reply means no deduction; other correction/cancellation messages do not affect the score.

How could the policy obtain a higher score without improving the purchase? What evidence, collected independently of its decision to ask, would establish the mismatch?

- Q2 and Q3 are not mutually exclusive failure categories. Q3 also has a defective scoring rule. The contrast is causal: Q2 changes purchases being counted; Q3 changes whether scored feedback is generated with the purchase held fixed. Independent collection of the request and merchant transaction evidence is a proposed evidential requirement, not a checker already implemented in the diagram.

4. Constraint and evidence

Choose one variant. Identify a hazardous system condition and propose a constraint to prevent or mitigate it. State where it would be enforced. Describe a test that attempts to violate the constraint, with an observation that counts as passing. What safety claim would remain unsupported after a pass?

Hazard — Constraint/enforcer — Test/pass condition — Remaining uncertainty