Lab 3: Reward Design Worksheet

Group Members: _____
Scenario: _____

Part A: Problem Definition

1. Environment Description

In plain language, describe the task your agent is trying to solve.

2. State Space (S)

What information does the agent observe at each step?

3. Action Space (A)

What actions can the agent take?

List or describe:

Part B: Reward Design Phase 1

4. Reward Function (R)

Write your reward function explicitly.

Use bullet points or a formula.

- +_____ for _____
- −_____ for _____
- −_____ for _____

5. Discount Intuition (γ)

Should the agent prioritize:
☐ Immediate rewards
☐ Long-term outcomes

Why?

6. Success Criteria

How will you decide if the agent is “working”?

Explain:

How would you measure or evaluate this?

Part C: Adversarial Review

Reviewing Group (not the same group): _____

7. How could the agent maximize reward while violating intent?

8. Pathological or unintended behaviors

What weird behavior might emerge?

9. What is missing from the reward?

What important signal is not incentivized?

10. Likely Learned Policy

Assume the agent is well-trained and strongly optimizes your reward. Describe, step by step, what the agent would actually do.

- What actions would it take most often?
- What actions would it avoid?
- What edge cases would it exploit?

Write a short behavioral description, not intentions.

Part D: Redesign

11. Revised Reward Function

Update your reward to mitigate the issues found above.

- +_____ for _____
- −_____ for _____
- −_____ for _____

12. Remaining Risks

What could *still* go wrong, even after your fixes?

13. Monitoring Plan

If you trained this agent, what would you monitor?

Explain:

Final Report

Scenario and Task Description

(2–3 sentences)

Original Reward Design

Briefly describe the initial reward function.

Observed Failure Modes

List the unintended behaviors or risks identified during critique.

Revised Reward Design

Explain what you changed and why.

Remaining Risks

What behaviors are still possible or hard to prevent?

Monitoring & Evaluation Plan

What signals would you track during training and deployment?

Key Takeaway

One sentence answering: *Why is reward design harder than it looks?*