

AI Alignment 11/17/2025

60 Points Possible

Attempt 1



In Progress

NEXT UP: Submit Assignment



Add Comment

Unlimited Attempts Allowed

▼ Details

AI Alignment Challenge: The Cleaning Robot

In this assignment, you will explore concepts in AI alignment by programming a cleaning robot that must efficiently clean dirt while avoiding unintended behaviors. This challenge demonstrates key alignment concepts including reward specification, goal completion, and avoiding reward hacking.



LINK TO INTERFACE FOR TESTING: <https://alignment-robot-railway-production.up.railway.app/> ↗(https://alignment-robot-railway-production.up.railway.app/)

The Challenge

You need to program a cleaning robot to clean 20 pieces of dirt on an 8x8 grid in the fewest possible steps. The robot:

- Can move up, down, left, right (actions 0-3)
- Can clean dirt at its current location (action 4)
- Can see its own position and the full dirt map
- Must clean all dirt within 200 steps to succeed
- Will compete against other students on the leaderboard

Your Task

1. You are provided with a naive baseline policy. Analyze why the baseline policy is inefficient
2. Design and implement a better policy
3. Test your policy on the web interface (you must enter in an identifier so I know you did this step! Each successful attempt is logged in a database I have set up and is tied to you only through your identifier (name you enter)).
4. Document the alignment challenges you encountered
5. Explain how your solution addresses them

Deliverable: Report

1. **Analysis of Baseline Policy**
 - What makes it inefficient?
 - What alignment issues does it demonstrate?
2. **Your Solution**
 - Explain your improved policy

- Include your code with comments
- Discuss your design choices

3. Alignment Discussion

- What alignment challenges did you encounter?
- How does your solution address them?
- What limitations remain?

4. Performance Analysis

- Compare your policy to the baseline
- Analyze failure cases
- Suggest further improvements

Submission

1. Submit your report as PDF (or as a link to a publicly accessible blog)
2. Your successful attempts (<200 steps to clean 20 dirt blocks) will be logged in the database. Make sure I can identify you in the database via your name!
3. Include your policy code as an appendix to your PDF report



Tips for Success

- Consider how different reward structures affect behavior
- Think about the trade-off between exploration and exploitation
- Test your policy against different dirt configurations
- Look for opportunities to optimize path planning
- Consider edge cases and failure modes

Choose a submission type



(<https://canvas.duke.edu/courses/62464/modules/items/572880>)

(<https://canvas.duke.edu/courses/62464/modules/items/572926>)