WORKSHEET

U12.5 Reward is a specification

Learning, and the capstone · University · about 35 min

Name _____ Class _____ Date _____

What this lesson is about

The optimiser answers the question you asked, which is rarely the question you meant.

- The failure has a name
- The two you will write yourself
- Shaping, and how it goes wrong
- Writing one that survives

Questions

7 marks in all

1. The reward is distance travelled towards a mark 60 cm ahead. What is it actually rewarding? [1 mark]

- ☐ **A** Going fast and far, so the best policy charges straight past the mark
- ☐ **B** Stopping on the mark
- ☐ **C** Getting close to the mark at some moment
- ☐ **D** Driving straight

2. Four commands were each driven for 4 s from the same line. What does this print? [1 mark]

```
MARK = 60.0
travel = {55: 49.5, 70: 63.8, 85: 79.2, 100: 93.6}
by_progress = max(travel, key=lambda c: travel[c])
by_error = min(travel, key=lambda c: abs(travel[c] - MARK))
print(by_progress, by_error)
```

3. What is the name for a policy scoring well on its reward while the behaviour is not what was wanted? [1 mark]

4. A robot is rewarded for a small depth reading, meant to encourage parking close to its charger. What is the likely result? [1 mark]

- ☐ **A** It drives up against whatever is nearest, because the reward is a proxy for the thing wanted
- ☐ **B** It parks at the charger, because that is the smallest reading
- ☐ **C** It learns nothing, because the reward is sparse
- ☐ **D** It learns to avoid obstacles

5. Which form of shaping reward is guaranteed not to change the optimal policy? [1 mark]

- ☐ **A** The difference of a potential between states, F(s') - F(s)
- ☐ **B** A bonus for being near the goal
- ☐ **C** A bonus for facing the goal
- ☐ **D** A small reward for every step taken

6. Why does a sparse reward of 1 for arriving and 0 otherwise teach almost nothing, even though it specifies the task perfectly? [1 mark]

- ☐ **A** A random policy never arrives, so every trial scores the same and there is no signal to follow
- ☐ **B** It rewards progress instead of arrival
- ☐ **C** It is a proxy for the thing wanted
- ☐ **D** The optimiser cannot handle integer rewards

7. A search's best score has risen steadily for fifty trials. What should you do before believing it? [1 mark]

- ☐ **A** Watch the behaviour that produced the winning score
- ☐ **B** Run fifty more trials to confirm the trend
- ☐ **C** Add a shaping term to speed it up
- ☐ **D** Report the best score with its trial number

The task: two scores, two winners

Try commands 55, 70, 85 and 100 for four seconds each from the same line, homing between trials. Print `by progress:`, the command that travelled furthest, and `by error:`, the command that finished nearest the mark 60 cm ahead. Then run the one the error score chose, and stop there.

```
from bugbot import *
connect()

DT = 0.1
MARK = 60.0
RUN_S = 4.0
COMMANDS = [55, 70, 85, 100]
```

Plan your program here, then type it in and press Run.

Do it on the robot

www.bugbotlab.com/learn/u12-5-reward-is-a-specification/

The simulator checks it and tells you when it passes. Nothing to install, no account.

Challenges

1. Add a third score: nearest approach to the mark at any moment during the trial. Which command wins that, and why is it the worst of the three specifications?
2. Write a score that wants the robot near the mark and stopped, and check that the command it picks is the same one.

3. Describe, in two sentences, how a policy could score well on your combined reward while doing something you would refuse to accept.