Dr. Taylor

Evaluating the Alignment of Reward Functions in Reinforcement Learning

Sarah Khandoker, Calarina Muslimani, Matthew E.Taylor

The Intelligent Robot Learning Lab, Department of Computing Science, University of Alberta

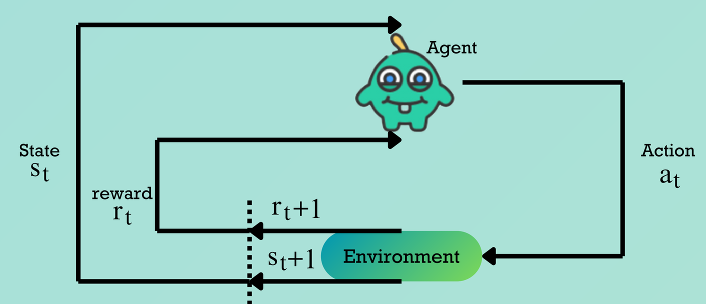
Introduction

- In AI, it becomes critical that agents learn to interact with the real world in dynamic scenarios and teach itself. Reinforcement Learning is one method of that.
- Reinforcement Learning (RL) is a Machine Learning (ML) technique in which an agent learns through a process of trail and error by maximizing rewards.

Rewards = Set by human engineers
= Feedback of performance for Agent

- **Agent-Environment Interface:**

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots$$



- **What happens when the reward function is designed poorly?** RL Agents using faulty reward functions cause them to perform sub-optimal actions that are misaligned with the preferences of the human stakeholders. [1]
- The *Reward Hypothesis* + poorly designed Reward Functions = value functions misrepresenting main objectives. [1]
- Faulty reward functions incentivize agents to follow a faulty policy, [2] assigning higher values to state-action pairs that the human stake-holders had not intended for. We want the optimal policy of the agent to be aligned with what stakeholders had intended.

$$q_*(s,a) \doteq \max_{\pi} q_\pi(s,a)$$

- The purpose of this project is to **understand alignment of human designed reward functions.**

Setting up the Environment

We set up our environment as a 4x4 grid world of a Hungry-Thirsty Domain (Singh, Lewis, and Barton 2009). Within our testbed, we have an article of food and water, and walls between squares. Our agent is placed at a random square at the beginning of every trajectory.

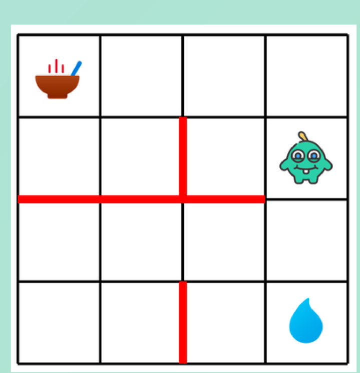
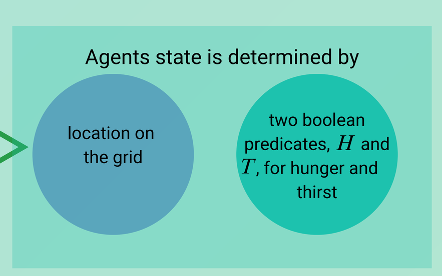


Figure 1.1 4x4 Hungry-Thirsty World [3] (positions generated at random)

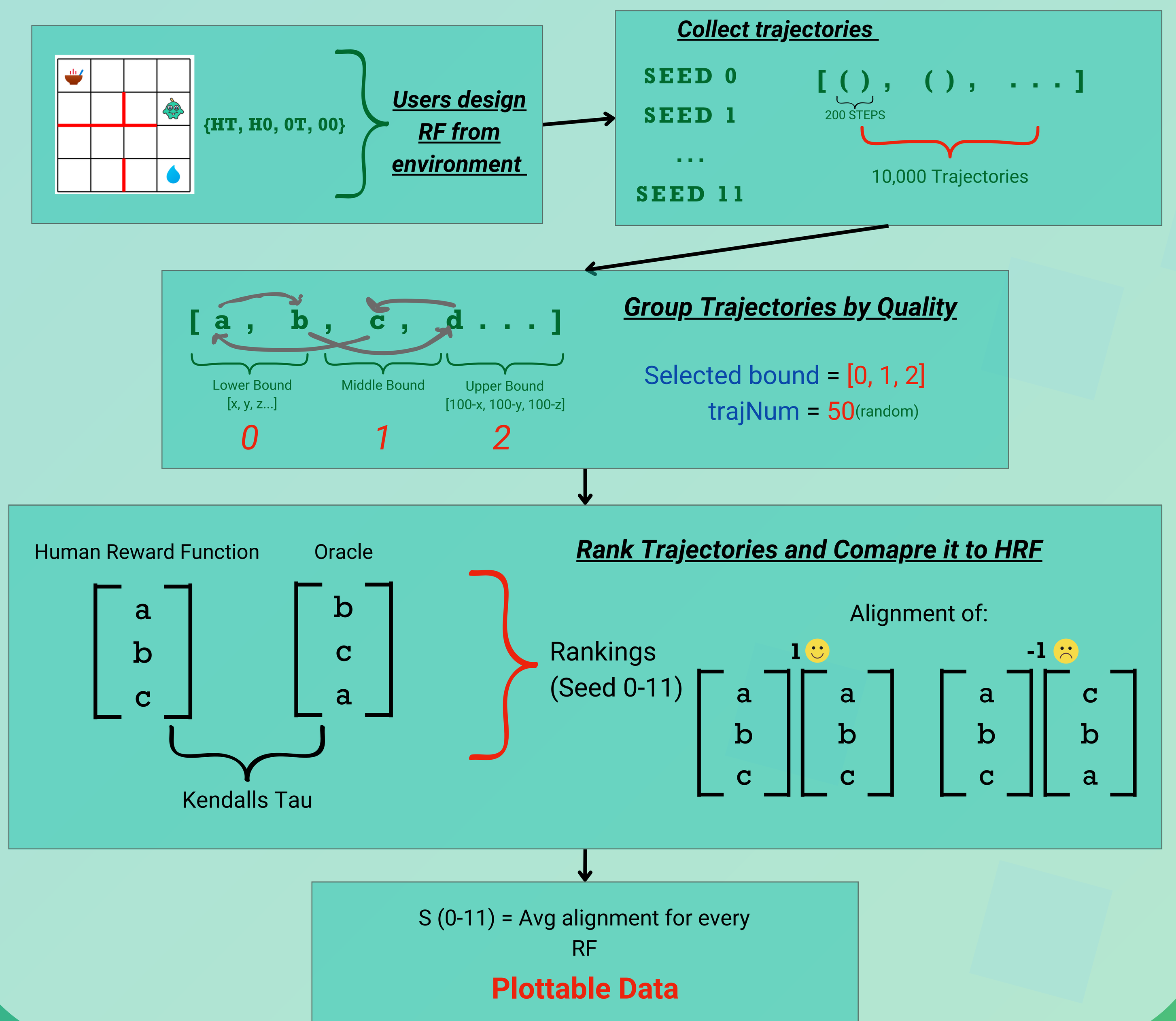
Objective: Stay satiated at as many time steps as possible



Research Question(s)

How do we characterize alignment?
How aligned are human designed reward functions?
How does trajectories quality influence alignment?

Methodology



Results (I)

How does trajectories quality influence alignment?

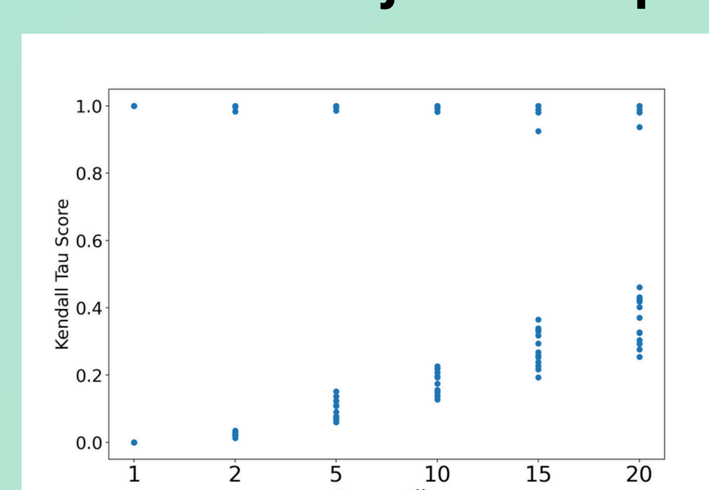


Figure 2.1 Lower-performing trajectories alignment data.

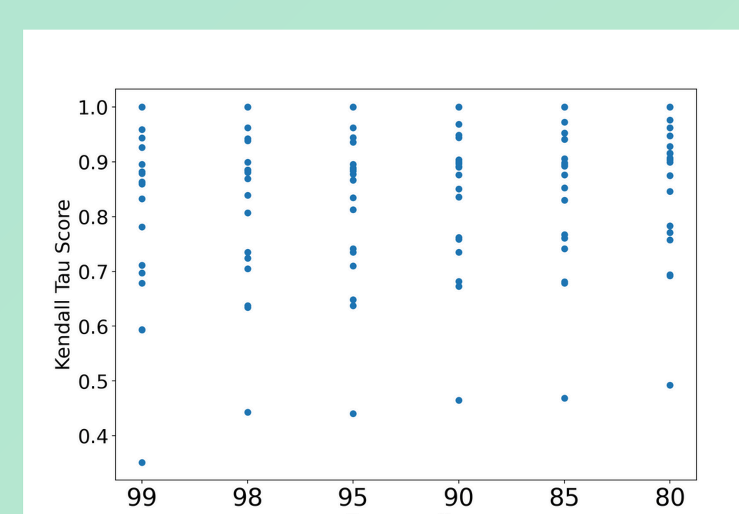


Figure 2.2 Higher-performing trajectories alignment data

- A very well-designed reward function has alignment roughly =1 through all bounds
- Only 5% of the RF were misaligned for the upper bound while 78% were misaligned for the lower bound.
- As we consider a larger set of trajectories (eg, 20 in 2.1, 80 in 2.2), we have less variance amongst alignment.

Results (II)

How aligned are human designed reward functions?

- Figure 3.1 is a well aligned reward function (-0.5, -0.5, 10, 10) that was set with values that resembled the pattern in our oracle.
  - HT, H0, had equal lesser values
  - 0T, 00, had equal greater values

  Regardless of if the bound was lower, higher, or had more differentiation, the alignment was always roughly =1.
- Figure 3.3 is a poorly aligned reward function (-10, 0, 10, 0) that was set with values that contradict the order of the values found from the oracle.
  - The top 99th percentile of high performing trajectories was more misaligned than the wider bounds.
  - The bottom 1st percentile was not aligned to any slight degree.

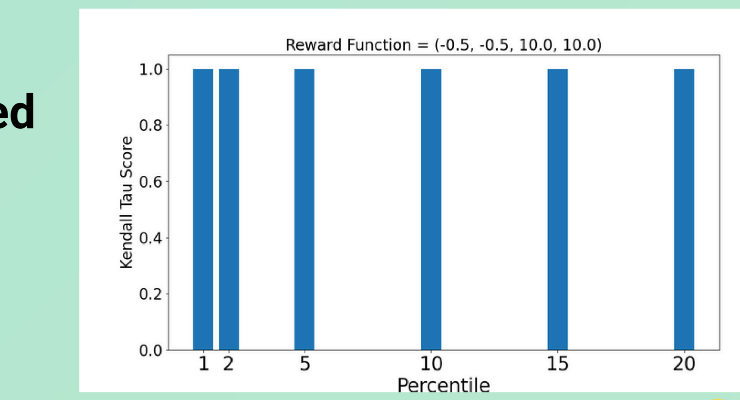


Figure 3.1 Alignment of RF (-0.5, -0.5, 10, 10) for lower bounds

(Oracle)

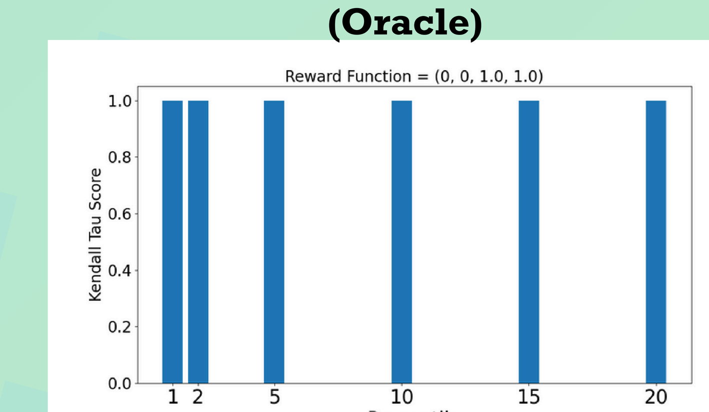


Figure 3.1 Alignment of RF (0, 0, 1, 1) for lower bounds

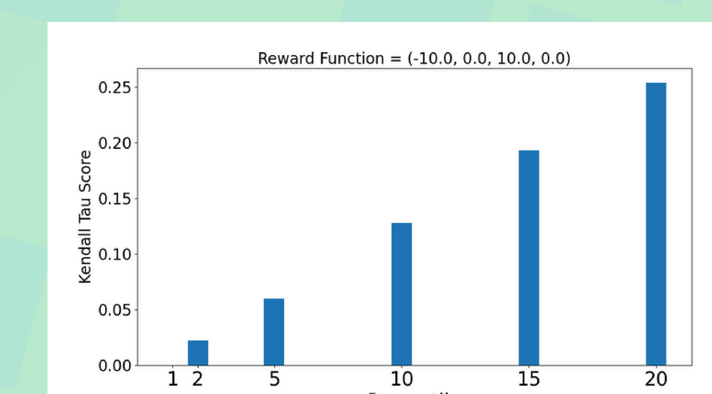


Figure 3.1 Alignment of RF (-10, 0, 10, 0) for lower bounds

Conclusions

- Humans may not create perfectly aligned reward functions.
- Humans are worse at designing reward functions that align with values in lower-performing trajectories.
- This work can assist in understanding how to design more aligned reward functions.
- Further understanding reward functions leads to further development of RL, assisting in creating alignment between agent and human preferences.

Acknowledgments and Citations

I would like to thank my direct supervisor, Calarina Muslimani, and my principal investigator Dr. Matthew E. Taylor for their guidance and support. For introducing me to Computer Science and inspiring my curiosity, I'd like to express my heartfelt gratitude to my Computer Science teacher, Mr.Said. I'd also like to extend my thanks to my school, Westwood Community High School, and Amii for their resources. I'd like to thank WISEST, Syncrude, and the Intelligent Robot Learning Lab for making this opportunity possible and sponsoring me.

Finally, I'd like to acknowledge the land that all work within this project was conducted on, Treaty 6 territory of Northern Alberta. Treaty 6 is home of the Dene Saline, Cree, Nakota Sioux, and the Saulteaux peoples. I live on Treaty 8 territory and am a visitor to Treaty 6. My research had benefitted from the land which provided me the grounds and spaces for innovation. Through current and future works, I hope my research can contribute to the preservation of Indigenous values, culture, and resources. I will continue to educate myself on the lands that I am on and its intersectionality with the work I conduct.

1. *1. Richard, S. Sutton., Andrew, G. Barto., (2018) Reinforcement learning: an introduction. Carnegie Melon University. https://www.andrew.cmu.edu/course/10-703/textbook/BartoSutton.pdf*
2. Bradley, W. Knox., James, MacGlashan., (2024) How to Specify Reinforcement Learning Objectives. BradKnox. https://bradknox.net/wp-content/uploads/2024/06/2024_How_to_Specify_RL_Objectives.pdf
Booth, S., Knox, W. B., Shah, J., Niekum, S., Stone, P., & Allievi, A. (2023). The Perils of Trial-and-Error Reward Design: Misdesign through Overfitting and Invalid Task Specifications. Proceedings of the AAAI Conference on Artificial Intelligence, 37(5), 5920-5929. https://doi.org/10.1609/aaai.v37i5.25733