CS 184: Reinforcement Learning
Embedded EthiCS @ Harvard Teaching Lab

Overview

Course:	CS 184: Reinforcement Learning
Course Level:	Upper-level undergraduate
Course Description:	"Modern Artificial Intelligent (AI) systems often need the ability to make sequential decisions in an unknown, uncertain, possibly hostile environment, by actively interacting with the environment to collect relevant data. Reinforcement Learning (RL) is a general framework that can capture the interactive learning setting and has been used to design intelligent agents that achieve super-human level performance on challenging tasks such as Go, computer games, and robotics manipulation. This course focuses on basics of Reinforcement Learning. The four main parts of the course are (1) multi-armed bandits, (2) Planning and Control in MDPs, (3) Learning in Large MDPs (function approximation), and (4) advanced topics."[1]
Module Topic:	Reward functions and Ethical Implications
Module Author:	Jenna L Donohue
Semesters Taught:	Fall 2022
Tags:	disability [phil], The difference principle [phil], maximin [phil], least well off [phil], structural injustice [phil], fairness [phil], utilitarianism [phil], Machine learning [CS], Reinforcement learning [CS], reward function [CS], utility function [CS]

Module Overview:	In this module, students were introduced to reward functions and their ethical implications. In particular we focused on and discussed how designing a reward function with an eye toward justice and the priority of the least well off might look differently from designing a reward function without those concepts in mind. The module was a mix of whole class discussion, whole class instruction, and group work. In particular, we discussed a RideShare case study and the ethical implications of setting a reward function in that context.	Unlike other kinds of ML, RL outputs an action rather than information. How the RL agent then determines whether the action was successful or not is by looking to the reward function.
Connection to Course Material:	Reward functions are one of the distinctive features of RL (Reinforcement Learning). They raise ethical questions in their own right, distinct from (but not wholly disconnected from) ethical questions raised by ML generally speaking. This course focuses on RL, so the connection between the ethical implications of reward functions and the course material is clear.	A module for this course could focus on ethical issues raised by ML generally, since RL is a kind of ML. Because this is a new course focused on RL in particular, the Embedded EthiCS instructor chose to focus on reward functions because they are a distinctive feature of RL not present in other versions of ML.

[1] https://shamulent.github.io/CS_Stat184_Fall22.html

Goals

Module Goals:	1. Recognize the importance of specifying a good / correct reward function. 2. Recognize that reward function specification is inherently value-laden. 3. Explain why different reward functions (in words and not yet formalized) benefit different groups, such as those in positions of power, those disadvantaged by the system itself, or the least well off in society, understood more generally.	
Key Philosophical Questions:	1. How should we design reward functions for RL systems? 2. What considerations, including technical, societal, and ethical, should we take into account when we are designing our reward functions? 3. How does reward function design change if we think about those who will be made worst off by the adoption of the system? How does it change if we think about prioritizing those who are least well off in society more generally?	Looking to Rawls' theory of justice and in particular the priority of the least well off helps to inform questions about how we should (ethically speaking) design our reward functions for reinforcement learning algorithms.

Materials

Key Philosophical Concepts:	● John Rawls & *Theory of Justice* ● least well off, maximin principle, difference principle ● structural injustice	When we design a reward function, we can think about the needs and interests of the least well off as a way of incorporating justice into our design. Considerations of structural injustice will also be relevant here.
Assigned Readings:	● Excerpt from Rawls, John (1985). "Justice as Fairness: Political not Metaphysical." *Philosophy and Public Affairs* ● Excerpt from Rawls, John (2001). *Justice as Fairness: A Restatement* (pages 59 - 63) ● Hawkins, Andrews (2022). "Uber Doesn't have to Provide Wheelchair-Accessible Vehicles in Every City, Judge Rules." *The Verge*. ● Lu, Donna (2020). "Uber and Lyft Pricing Algorithms Charge More in Non-White Areas." *NewScientist*.	The readings were chosen with two main ideas in mind: (1) to introduce students to Rawls' theory of justice without overwhelming them and (2) to introduce the idea that issues of justice might be at stake and relevant to setting the prices for rides on a RideShare. For (1), the Embedded EthiCS instructor assigned short excerpts of Rawls so that students could get a sense of his ideas and argumentation. For (2), the Embedded EthiCS instructor assigned news articles discussing Uber and disability and Uber and race, each helping to motivate the idea that justice can be at stake in this area. While the

		RideShare applications do use machine learning right now, they do not yet use RL. During class we imagined a RideShare application that employed RL. The readings helped them to connect the issues of justice to the issues discussed in class.

Implementation		
Class Agenda:	1. Brief introduction to Reward Functions 2. Case Study: Ride Share 3. John Rawls & Least Well Off 4. Return to Case Study and Apply Maximin	Class began with a brief introduction to reward functions, what they are, and why they pose both technical and ethical challenges. We then discussed the case study of using RL to set prices for a RideShare company. The Embedded EthiCS instructor gave an introduction to Rawls' *Theory of Justice* and the priority of the least well off. Finally, we engaged in a whole class discussion applying the priority of the least well off to the case study of setting RideShare dynamic prices.
Sample Class Activity:	The students were asked to talk in groups about ethical considerations relevant to designing a reward function for an RL algorithm setting prices for a RideShare app. They were also asked to talk in groups about how to apply Rawls' conception of priority for the least well off in this context. Group discussion of this case study was the main active-learning component of this module.	Frequent opportunities for group discussion can allow students who are less likely to talk during whole class discussion to have their voices heard.
Module Assignment:	The assignment was a question on a problem set for the students. This method was chosen as a way of making the Embedded EthiCS lectures feel like an integral part of the course: the students are used to problem sets and have them regularly. Assignment Directions: Write a paragraph of 5-7 complete sentences addressing the following question. "Consider the hypothetical RL algorithm that sets dynamic prices for rides on a RideShare app that we discussed in class. Choose **one** concept and explain why it **should or should not** be incorporated into a **just** utility function. Be sure to explain why it does or does not favor the interests of the least well off and whether they would fare better under a different utility function as part of your answer. It is	Incorporating a question into a problem set is one way to embed the content of the module within current assignment structures.

	ok to set aside / ignore some of the complicating factors we discussed in class, such as other RideShare company's prices, shareholder demands, etc. for the purposes of answering this question."	
Lessons Learned:	The module was well-received, and participation was high. A few lessons learned: - Reinforcement learning remains a very new area of study, so present-day applications are few and far between. In this module the Embedded EthiCS TA created a case-study for an application of the technology that doesn't exist yet. This went mostly ok, but it was crucial to include a specification for both state and action when asking students to think about reward functions. Otherwise the conversation is a bit too abstract and disconnected from the course content. - Students in this course were eager to talk and had a lot to say. Consider closing discussion a bit sooner to ensure sufficient time to discuss the application of the philosophy to the case study.	Marginal notes