STA3145

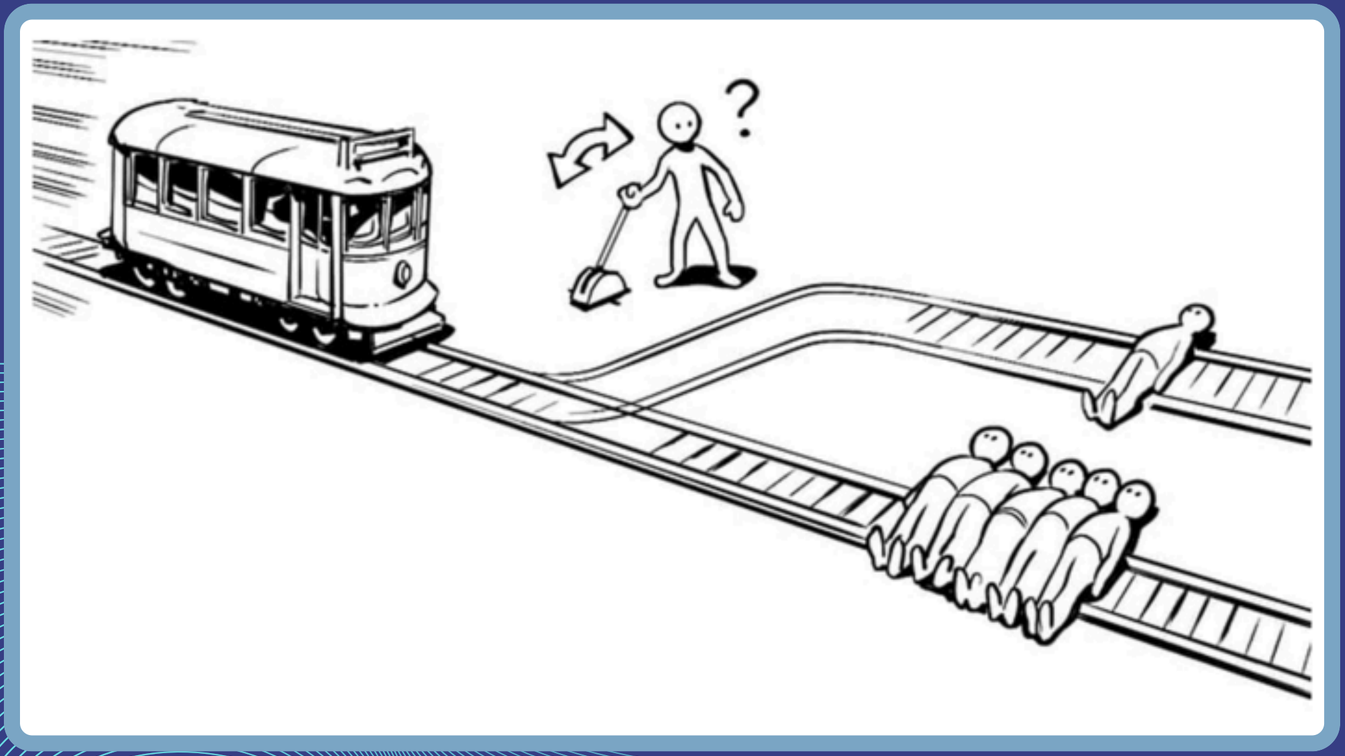
Solving Trolly Problem via Reinforcement Learning

Chaewon Yoo, Mingyu Kim, Seungjoo Yoo, Beomjun Shin

Speaker:
Beom Jun Shin

Contents

Background



https://theaxiom.ca/ethical-murder-the-trolley-dilemma/

Trolly Problem

- Would you kill one person to save five?

*Utilitarianism vs Deontology

Existing Works

<Reinforcement Learning Under Moral Uncertainty>

Multi-objective RL

solve the problem of finding the set of efficient trade-offs among competing objectives

does not address how to choose which trade-off policy to deploy

Which ethical theory should an agent follow?

A) By Voting system with credence

Existing Works

<Reinforcement Learning Under Moral Uncertainty>

Voting System

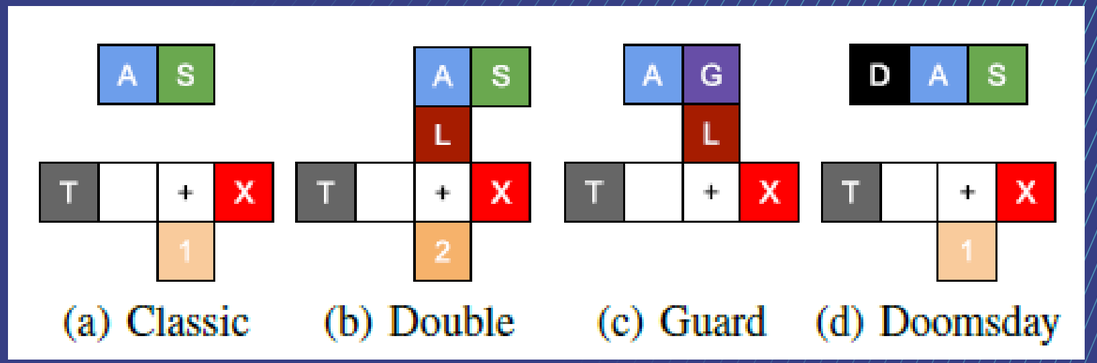
: Claiming each theory's opinion based on the credence and budgets

Nash Voting

Each theory provides
continuous vote value for actions
under their remaining cost budget

Variance Voting

Take the preferences of
each theory's action Q and then
transform to use Variance-SARSA



Adrien Ecoffet et al., "Reinforcement Learning Under Moral Uncertainty,"

Existing Works

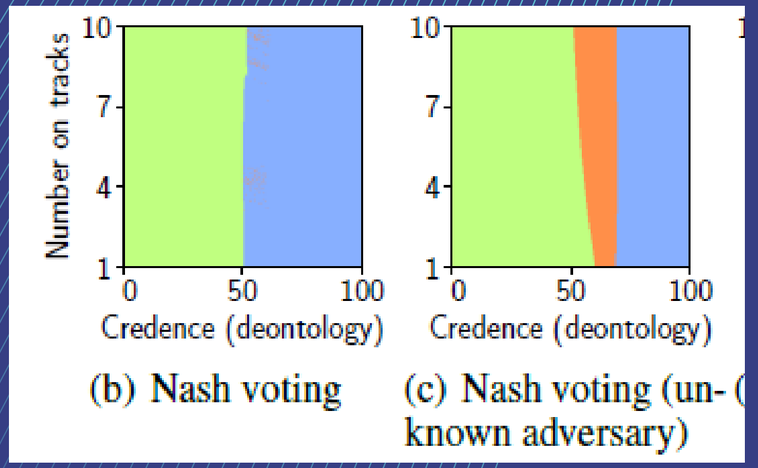
<Reinforcement Learning Under Moral Uncertainty>

Problems in Voting System

- Credence is determined by designer's beliefs → controversial
- Idea Under Multi-Agent (= Theory) RL : Computation complexity, …
- Incomparable theories → hard to extend in voting system

  (e.g. No Compromise in Nash voting)

How to handle these problems?



Adrien Ecoffet et al., "Reinforcement Learning Under Moral Uncertainty,"

Existing Works

<Ethical and Statistical Considerations in Models of Moral Judgments>

Goal : Having the agent 1) learn about its ethical objective function while
2) making decisions for maximizing rewards

Bayesian Approaches : Update beliefs by maximizing a combined utility function
represented as a linear combination of different ethical utility functions

Note : Motivating the agent to learn what is “ethical” minimizing our base belief

Apply Bayesian’s perspective for Moral Decision making!

Goal

Implement the situation of moral uncertainty using Bayesian Reinforcement Learning

Bayesian Reinforcement Learning

$$Posterior \propto (likelihood \times prior) \longrightarrow \text{able to update prior}$$

1. Prior Distribution
    a. Represent the inital beliefs

2. Posterior Distribution
    a. Represent the update belief after observing datas

3. Bayesian Update
    a. Update the prior distribution to the posterior using Bayes Theorm

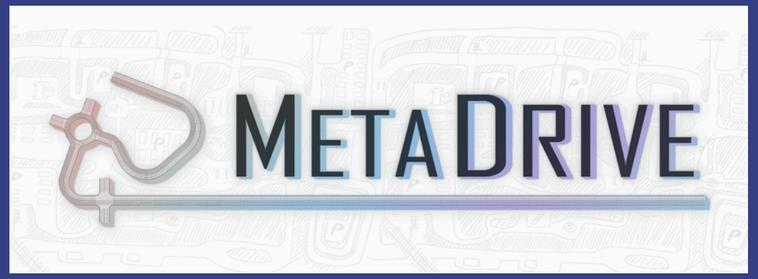
Bayesian Reinforcement Learning

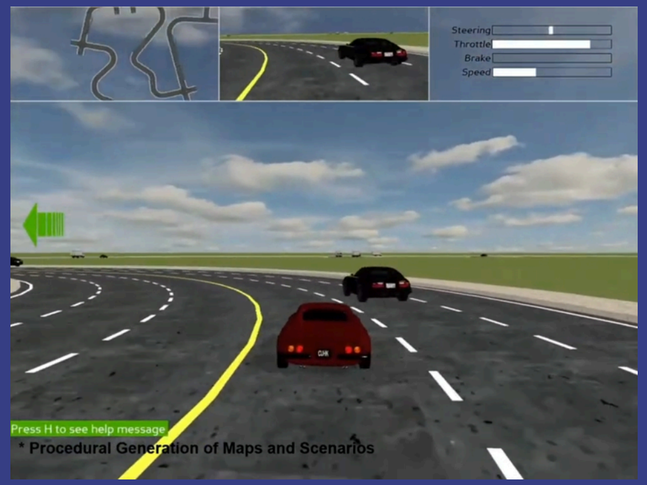
Implementation Steps

Initialization	Initialize prior parameter = [0.5, 0.5] credence for each theory
Observation	Observe data (state, action, and reward)
Update	Update the posterior distribution with the observed data
Decision Making	Use the updated posterior distribution to make decisions (Involves sampling from the posterior)
Iteration	Repeat step 2,3,4

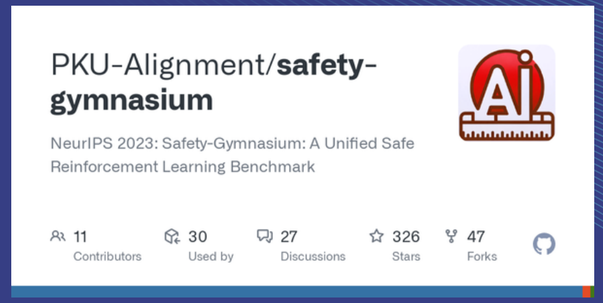
Implementation

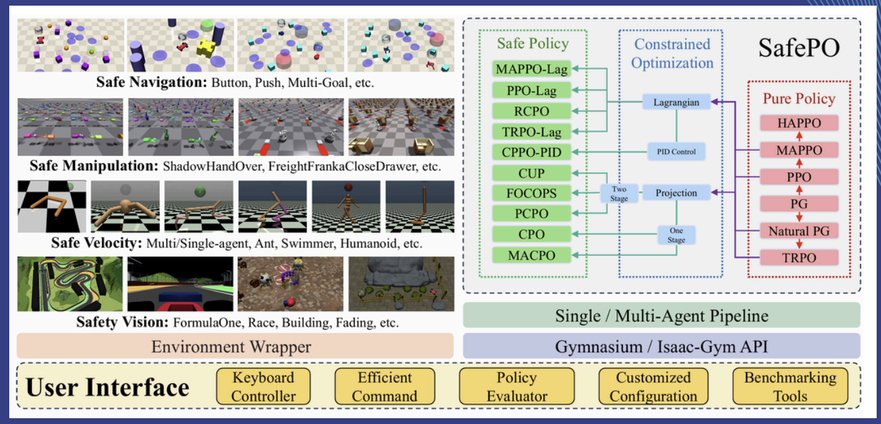
Environment





https://metadrive-simulator.readthedocs.io/





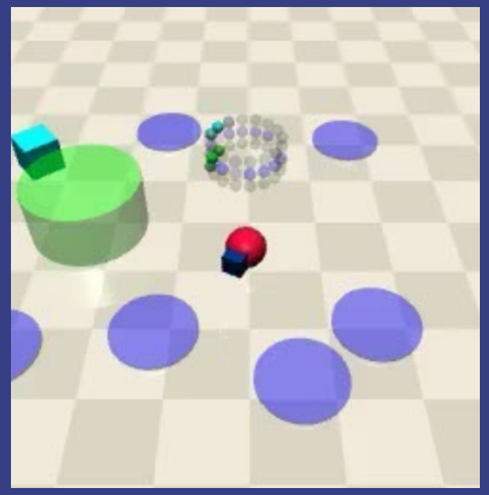
https://github.com/PKU-Alignment/safety-gymnasium

Implementation

Results

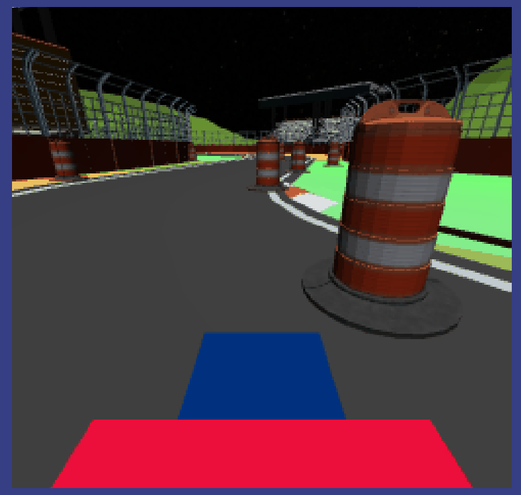
- Apply a bayesian approach at each episode

  ⟶ cost sensitive!

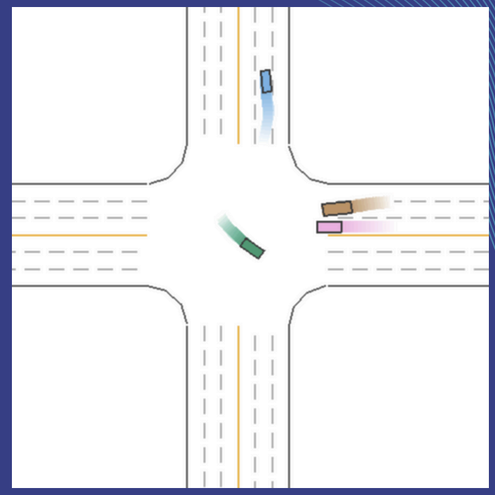


Safety-goal 1

Safety-goal 2



Safety-Formula One



MetaDrive 1

MetaDrive 2

Significance & Limitations

01

Consider more situation-appropriate prior distribution

02

Applicable in situations where we can’t determine prior parameter

03

Increased action flexibility compared to Voting

References

- Adrien Ecoffet et al., Reinforcement Learning Under Moral Uncertainty, ICML, 2021.
- David et al., Reinforcement Learning as a Framework for Ethical Decision Making, AAAI, 2016.
- T Sivill, Ethical and Statistical Considerations in Models of Moral Judgments
  Frontiers in Robotics and AI, 2019

Thank you

Solving Trolly Problem via Reinforcement Learning

STA3145 TEAM 2

June, 7, 2024