Brunel University London

College of Engineering, Design and Physical Sciences (CEDPS)
Department of Computer Science
BSc Computer Science (AI)

Reinforcement Learning vs. Cooperation: Inhibiting Selfishness in AI Agents

Philipp E. Bibik
Supervised by Dr Alina Miron

Introduction

- Artificial Intelligence has been increasingly integrated into various aspects of society
  - Becoming more **capable**, **ubiquitous**, and **autonomous**
- Reinforcement Learning optimises for specified objectives / goals
  - Agents are **inherently self-interested**
  - Pursue **only their own goals**
  - Do not consider anything beyond the reward function
- Can cause **negative side-effects** when interacting in complex systems
  - Recommender Systems causing social media addiction in teens
  - Cutting-off access to healthcare prematurely to maximise profits

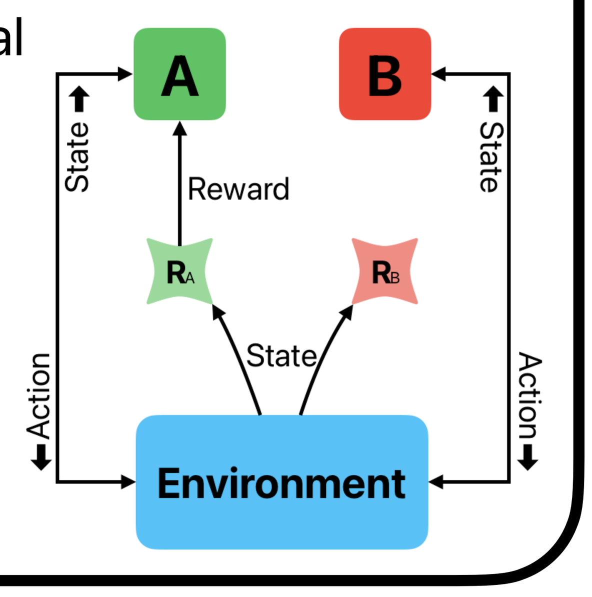
Aims and Objectives

Aim: Identify and evaluate effective approaches for inhibiting selfish behaviour in Reinforcement Learning-based agents, in multi-agent environments.

Objectives:
- Literature review
- Design and implement environments encouraging selfishness
- Train agents adopting selfish policies
- Develop multiple inhibiting approaches
- Train agents using inhibiting approaches
- Evaluate and compare effectiveness and drawbacks of approaches

Background

- **Reinforcement Learning is a Machine Learning paradigm**
  - Agents take actions in an environment
  - Feedback from reward function: Reward or Penalise states
    - e.g. in chess: taking/losing piece (+1/-1); winning/losing game (+50/-50)
  - Learns through trial-and-error: Starts randomly but improves
  - Approximates Optimal Policy (best State to Action mapping)
    - Optimal Policy is defined across an **MDP**, derived from an **Environment and Reward Function**
    - Only way to make agent non-selfish, is by **modifying / injecting into the Reward Function**, to change the MDP & Optimal Policy
- **Inherent Self-Interest**
  - Agent told to collect wood, faced with options:
    - Collect 50 logs without side-effects
    - Collect 51 logs, but murder 5 people in the process
  - Agent will prefer second option, unless told to care about humans
  - Learns to pursue **marginal increases in reward**, regardless of effects
- **Inverse Reinforcement Learning** (IRL) inverts classical RL problem
  - **Learn Reward Function** from States and Actions
  - Collect observations of 'expert' pursuing a goal
  - Learn a Reward Function that can explain the observed behaviour

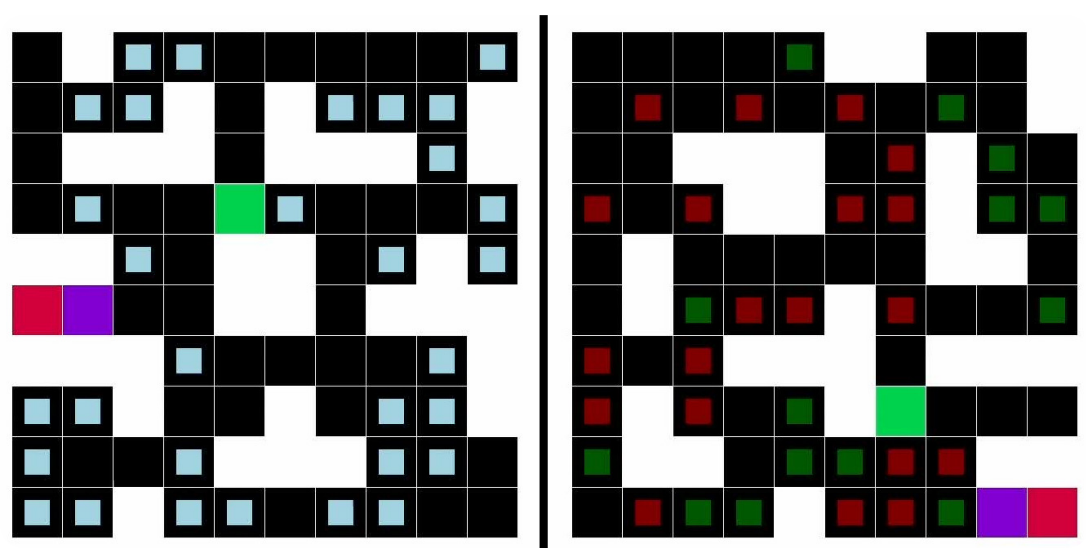


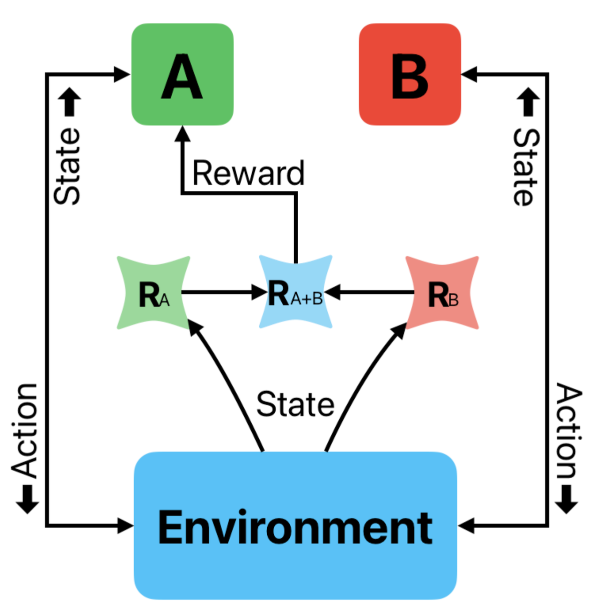
Regular RL-loop in a multi-agent system

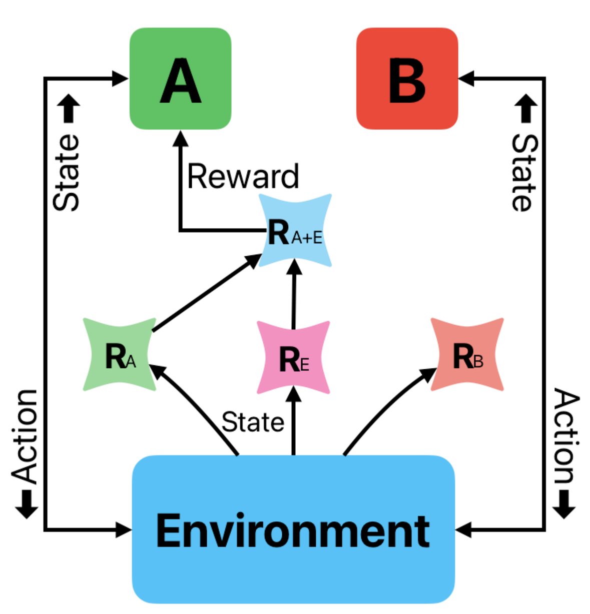
Methodology

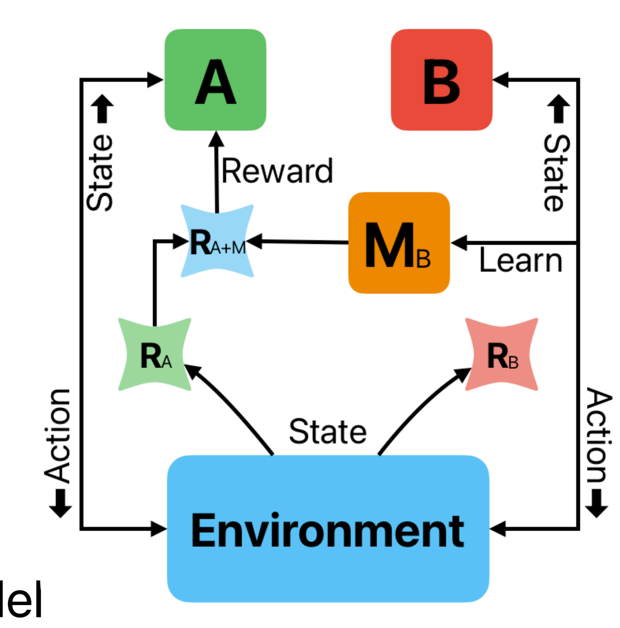
Environments:

Abstracted Girdworlds (2d board, with simple goals)
Goal is to collect own Resources, while other agent with own goals exists
Studying 3 scenarios:
- (i) Small cost to help (open door), with non-conflicting goals [right board]
- (ii) Small cost to help, with conflicting goals (zero-sum game) [left board]
- (iii) Large cost to help (prolonged actions)
  - (a) Touch other agent to speed-up
  - (b) Touch other agent to slow-down









Inhibition Approaches:

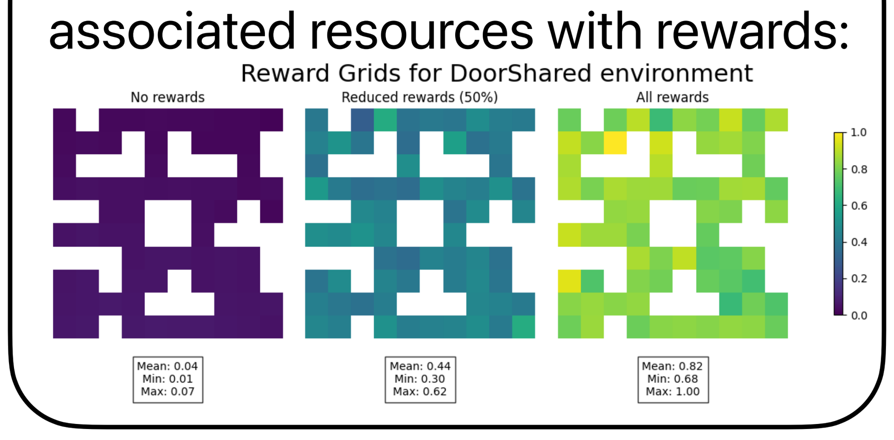
Add Secondary Agent's Reward [top]:
Add B's Reward Function to A's (with multiplier)
Best case scenario: Directly consider B's goals
Need to know B's Reward Function (a priori)

Environment-based custom Reward [middle]:
Write custom Reward function incentivising cooperation / penalising selfishness
Not scalable: Redo for each environment
Susceptible to not considered edge-cases

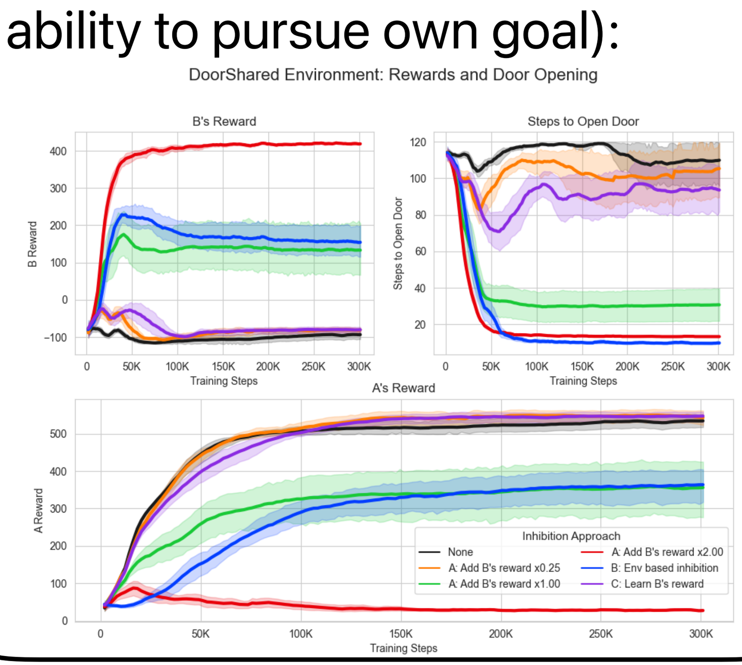
Learn Secondary Agent's Reward [bottom]:
Similar to first approach, but instead reward model is learned (using IRL)

Results

Reward model for approach 3 (IRL) associated resources with rewards:



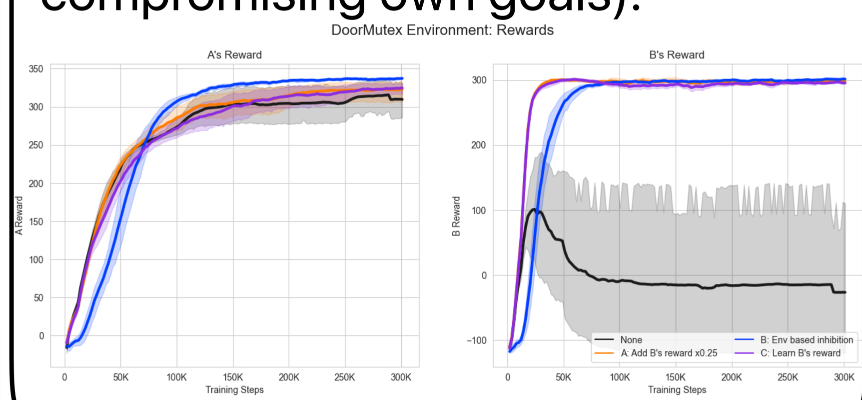
In zero-sum games, **only unreasonable weightings** for B's reward made agent cooperative (while removing ability to pursue own goal):



Environment-based inhibition scales worst, while also having the largest potential for unintentional side-effects.

Adding-based approaches performed similarly, with generalisability of **IRL-based inhibition** showing most promise for real-world scenarios.

All approaches perform similarly in low-cost scenario (i), massively outperforming no inhibition (help the other agent without compromising own goals):



In the speed-up scenario, environment-based inhibition **underperformed**, while others performed similarly.
A optimised for touching instead of helping *B* meaningfully, resulting in bad outcomes for both agents