On-Demand Coordination of Multiple Service Robots

Piyush Khandelwal

February 22, 2017

Dissertation Defense

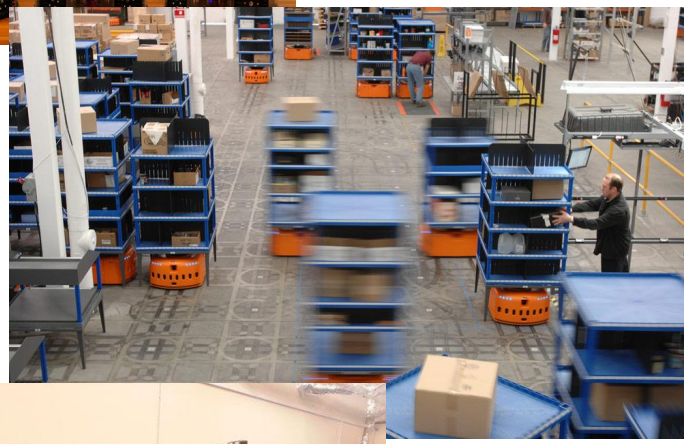
LARG
Learning Agents Research Group
The University of Texas at Austin

Motivation - Interactive Robot Systems

RP-VITA - iRobot

Cobots - CMU

Tally - Simbe Robotics



Amazon Robotics

OSHBot - Lowe's

STRANDS Project

LARG
Learning Agents Research Group
The University of Texas at Austin

Motivation - Video Demonstration

On-Demand Multi-Robot Coordination

Key aspects of interest for this problem:

- Temporarily coordinate multiple robots
- Unstructured indoor environment
- Stochastic outcome of actions

What is an *on-demand* multi-robot coordination task?

- Requires *real-time* planning
- Does not substantially deviate robots from independent *background* tasks

LARG
Learning Agents Research Group
The University of Texas at Austin

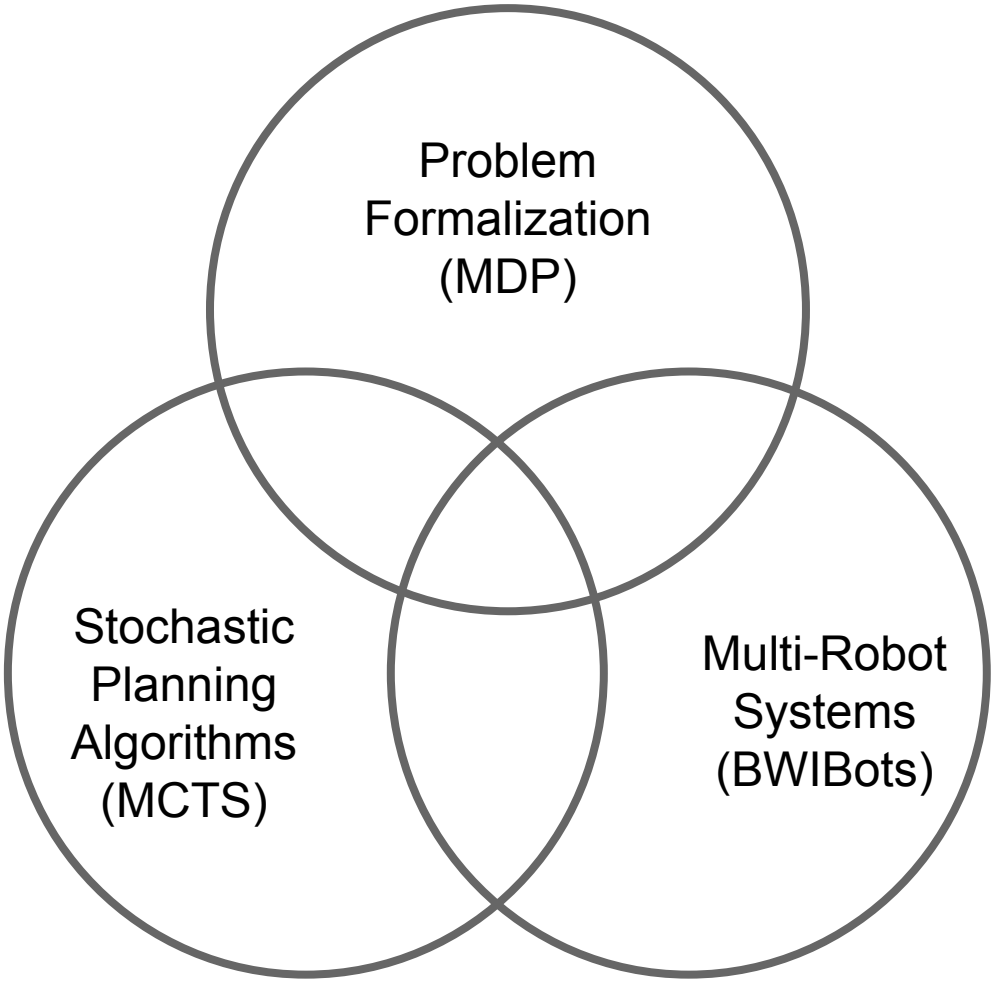
Thesis Question

How can multiple service robots be efficiently *interrupted*, *reassigned*, and *coordinated* to perform an on-demand task while

1. ensuring quick completion of the task
2. with minimal disruption to the robots' background duties?

LARG
Learning Agents Research Group
The University of Texas at Austin

Thesis Research Topics



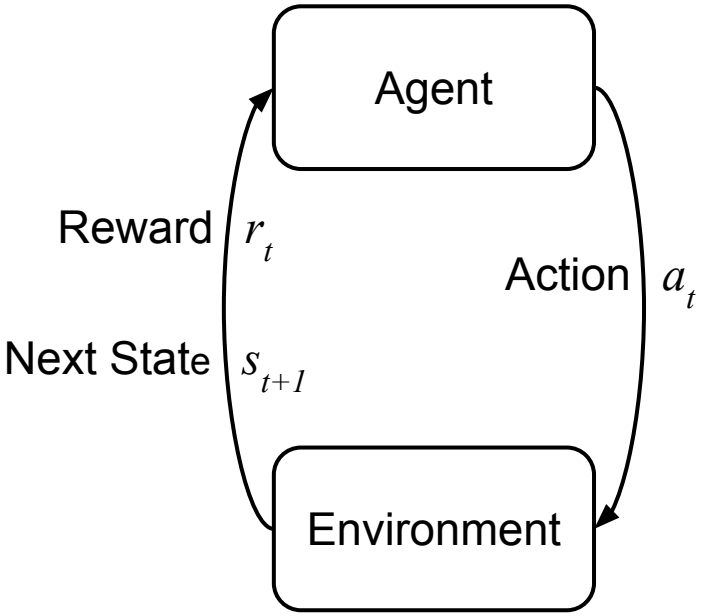
LARG
Learning Agents Research Group
The University of Texas at Austin

Talk Outline

- Background:
  - Markov Decision Processes
  - Monte Carlo Tree Search
- MDP Formalization of On-Demand Multi-Robot Coordination Problem
- Biased Backup in Monte Carlo Tree Search
- BWIBot Multi-Robot System + Conclusion

LARG
Learning Agents Research Group
The University of Texas at Austin

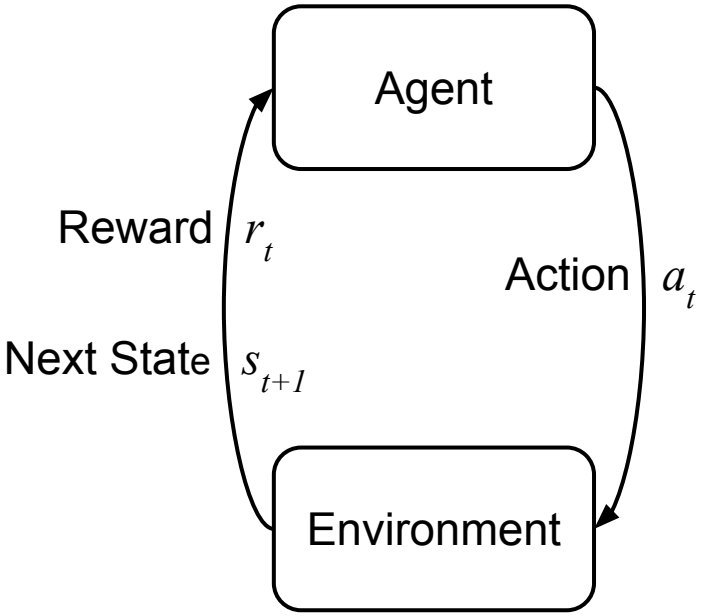
Markov Decision Process



Formally, an MDP comprises:

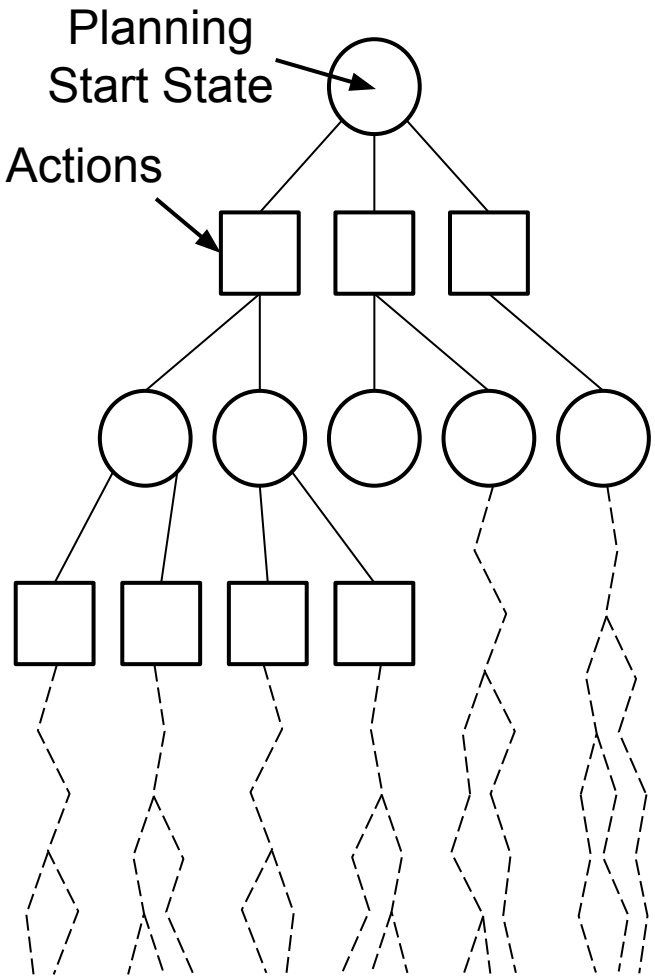
- State space
- Action space
- Transition Function
- Reward Function
- Discount Factor

Markov Decision Process



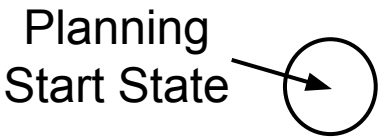
- Multiple actions available to agent at a state
- Each action can have a stochastic outcome
- Optimal solution can be computed using techniques such as *Value Iteration*, but it may be impractical to do so

Monte Carlo Tree Search



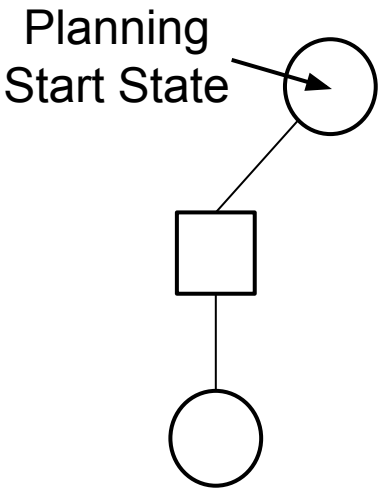
- Approximate search-based solver for MDPs
- Requires access to a model of the MDP for planning

Monte Carlo Tree Search



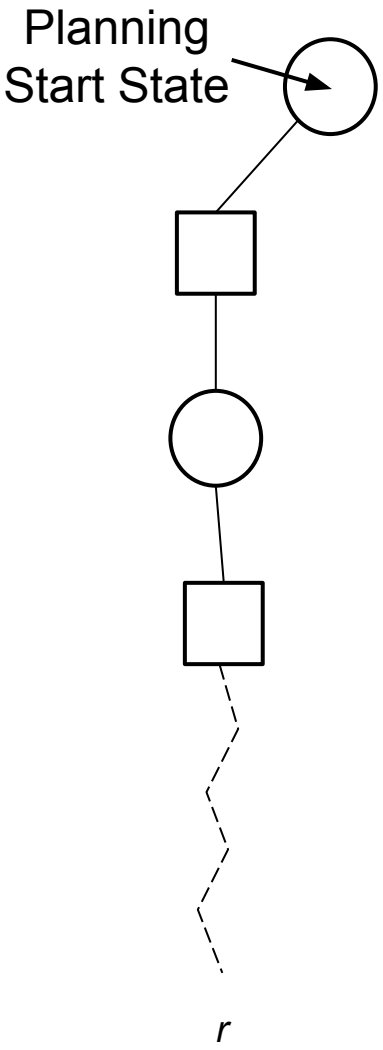
- 4 stages in MCTS:
  - Selection
  - Expansion
  - Simulation
  - Backup

Monte Carlo Tree Search



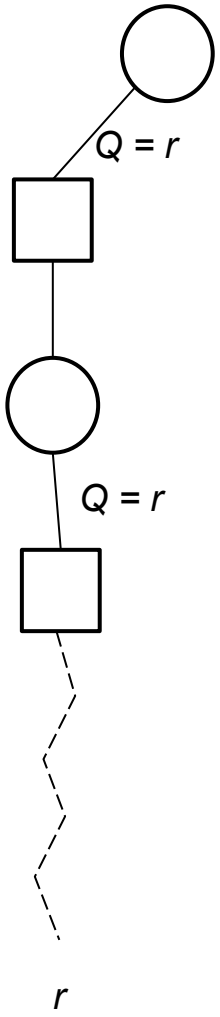
- 4 stages in MCTS:
  - Selection
  - ***Expansion***
  - ***Simulation***
  - Backup

Monte Carlo Tree Search



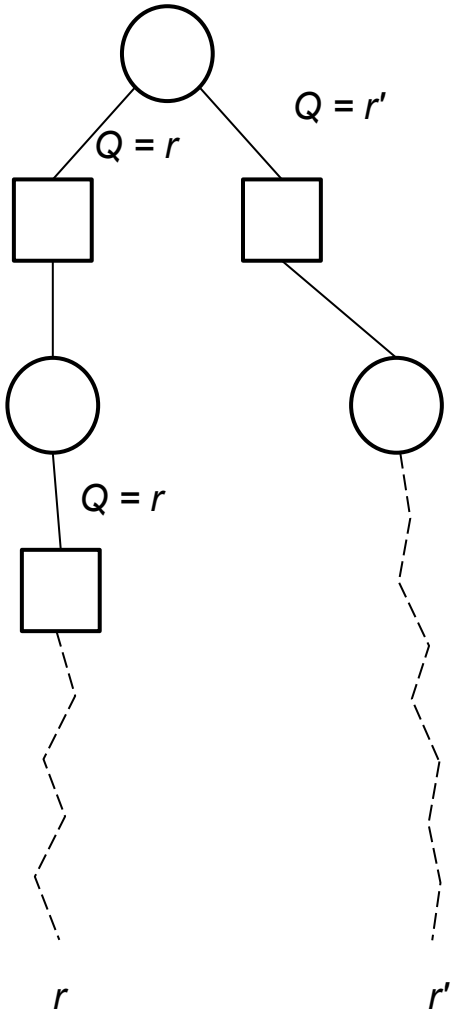
- 4 stages in MCTS:
  - Selection
  - ***Expansion***
  - ***Simulation***
  - Backup

Monte Carlo Tree Search



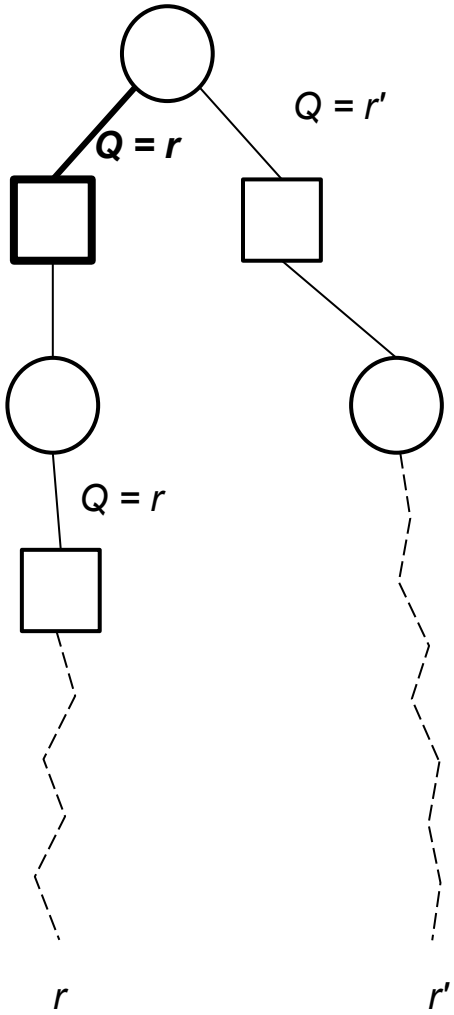
- 4 stages in MCTS:
  - Selection
  - Expansion
  - Simulation
  - ***Backup***

Monte Carlo Tree Search



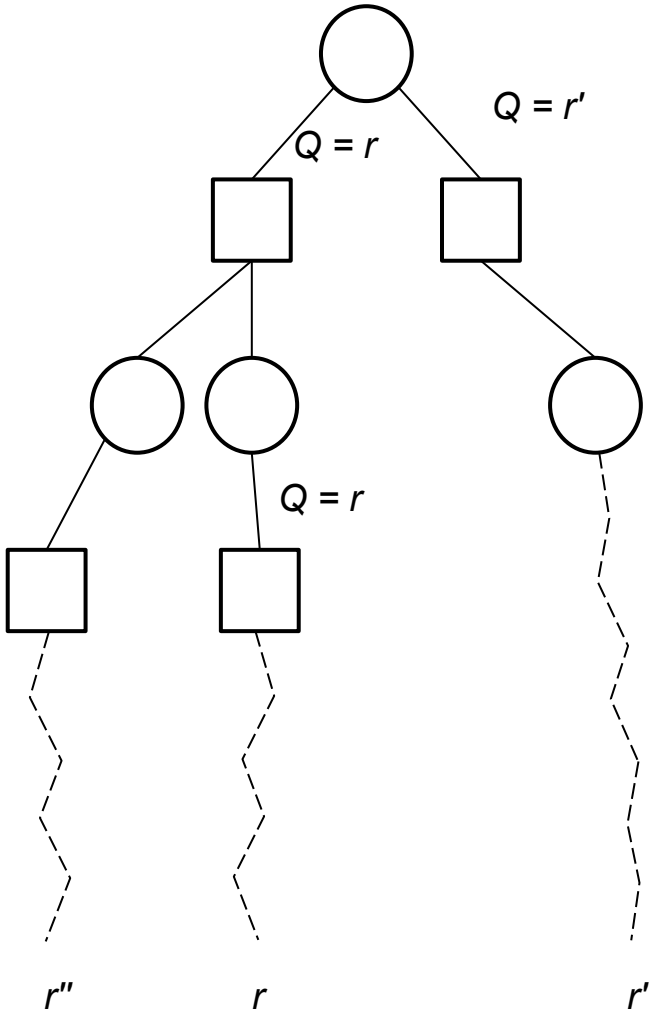
- 4 stages in MCTS:
  - Selection
  - ***Expansion***
  - ***Simulation***
  - ***Backup***

Monte Carlo Tree Search



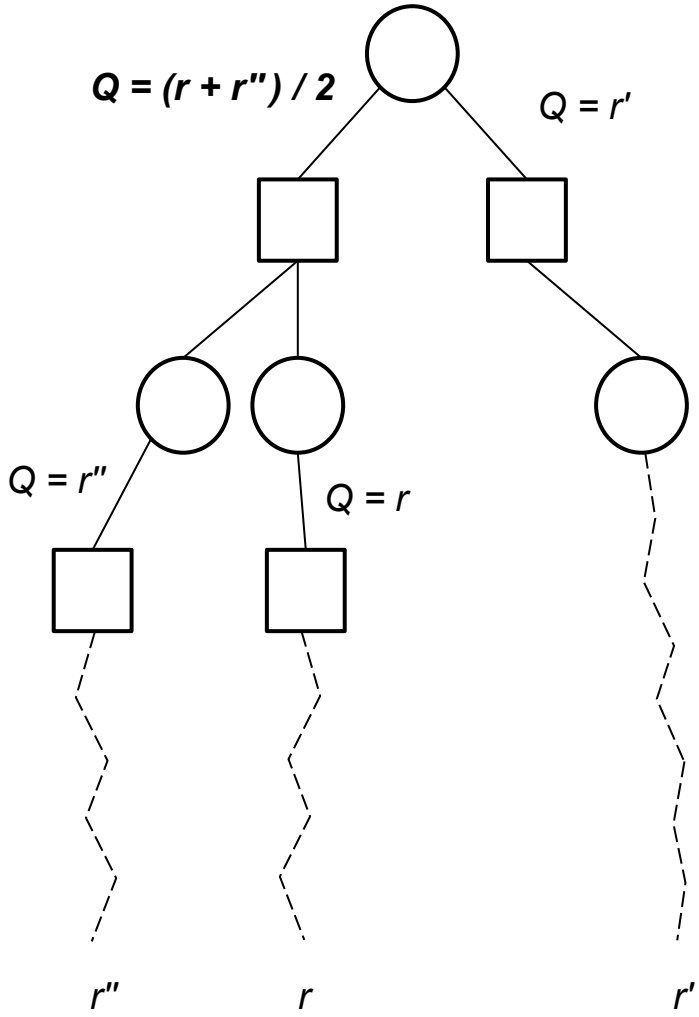
- 4 stages in MCTS:
  - ***Selection***
  - Expansion
  - Simulation
  - Backup

Monte Carlo Tree Search



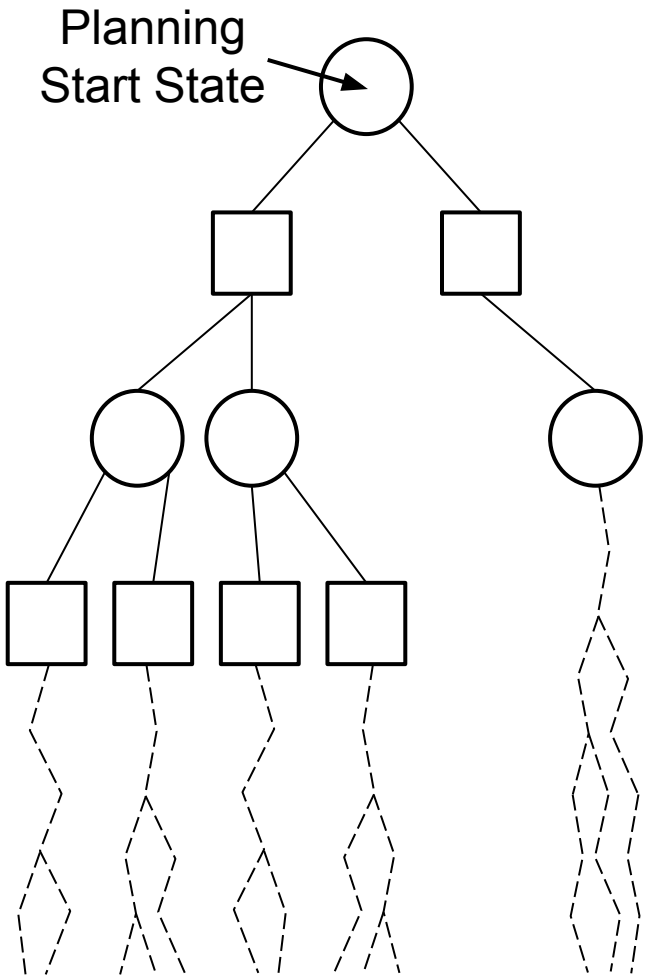
- 4 stages in MCTS:
  - Selection
  - ***Expansion***
  - ***Simulation***
  - Backup

Monte Carlo Tree Search



- 4 stages in MCTS:
  - Selection
  - Expansion
  - Simulation
  - ***Backup***

Monte Carlo Tree Search



- Use UCB1 for intelligent selection
- Run as many simulations as time permits
- Return action with highest Q value at the end of planning.

Overview

- Background
  - Markov Decision Processes
  - Monte Carlo Tree Search

- MDP Formalization of On-Demand Multi-Robot Coordination Problem

- Biased Backup in Monte Carlo Tree Search

- BWIBot Multi-Robot System + Conclusion

In this section

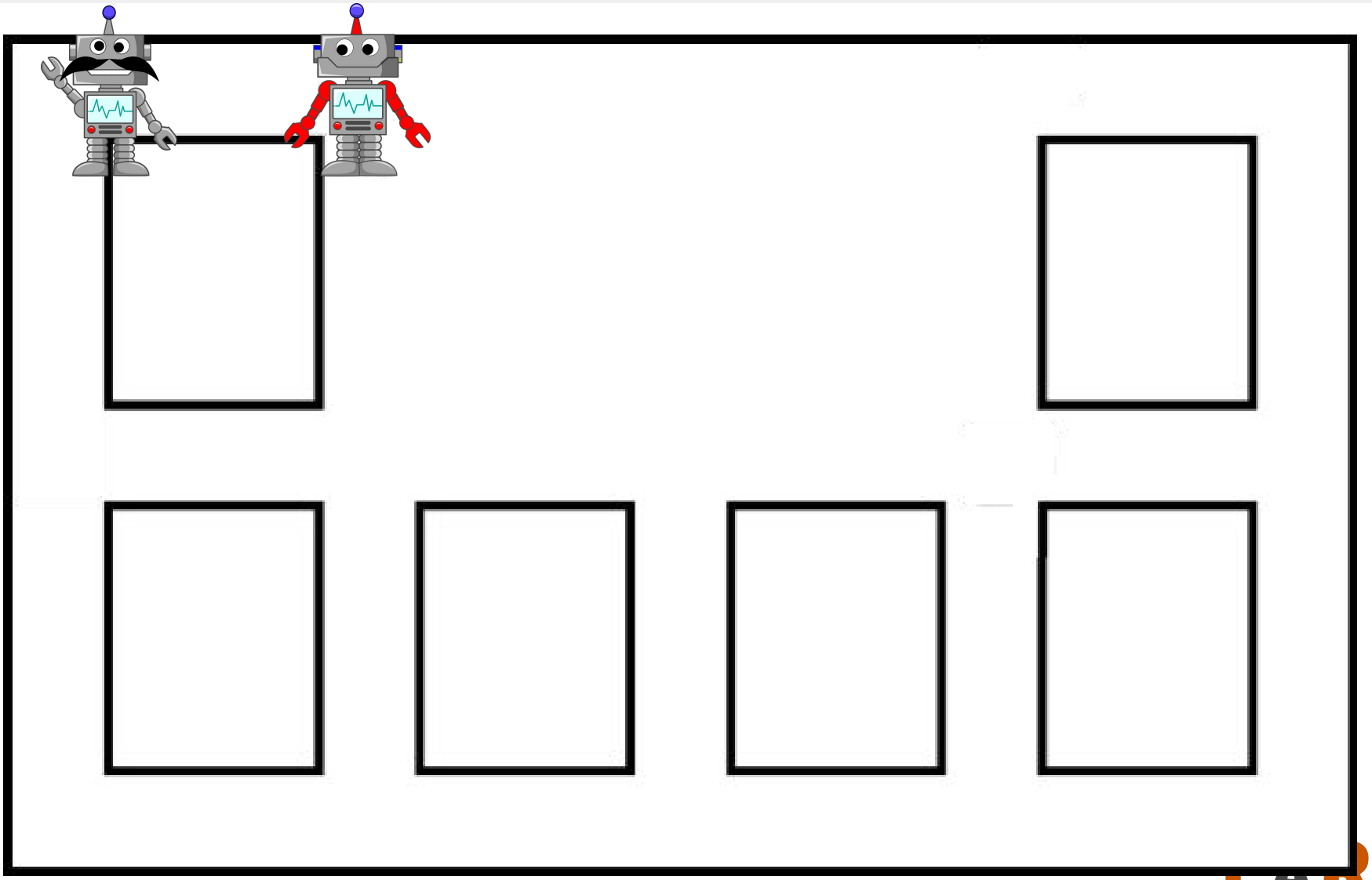
- We discuss the salient aspects of formalizing the on-demand multi-robot coordination problem as an MDP. [1][2]

- We will demonstrate that MCTS based approaches can outperform heuristic baselines under almost all domain configurations.

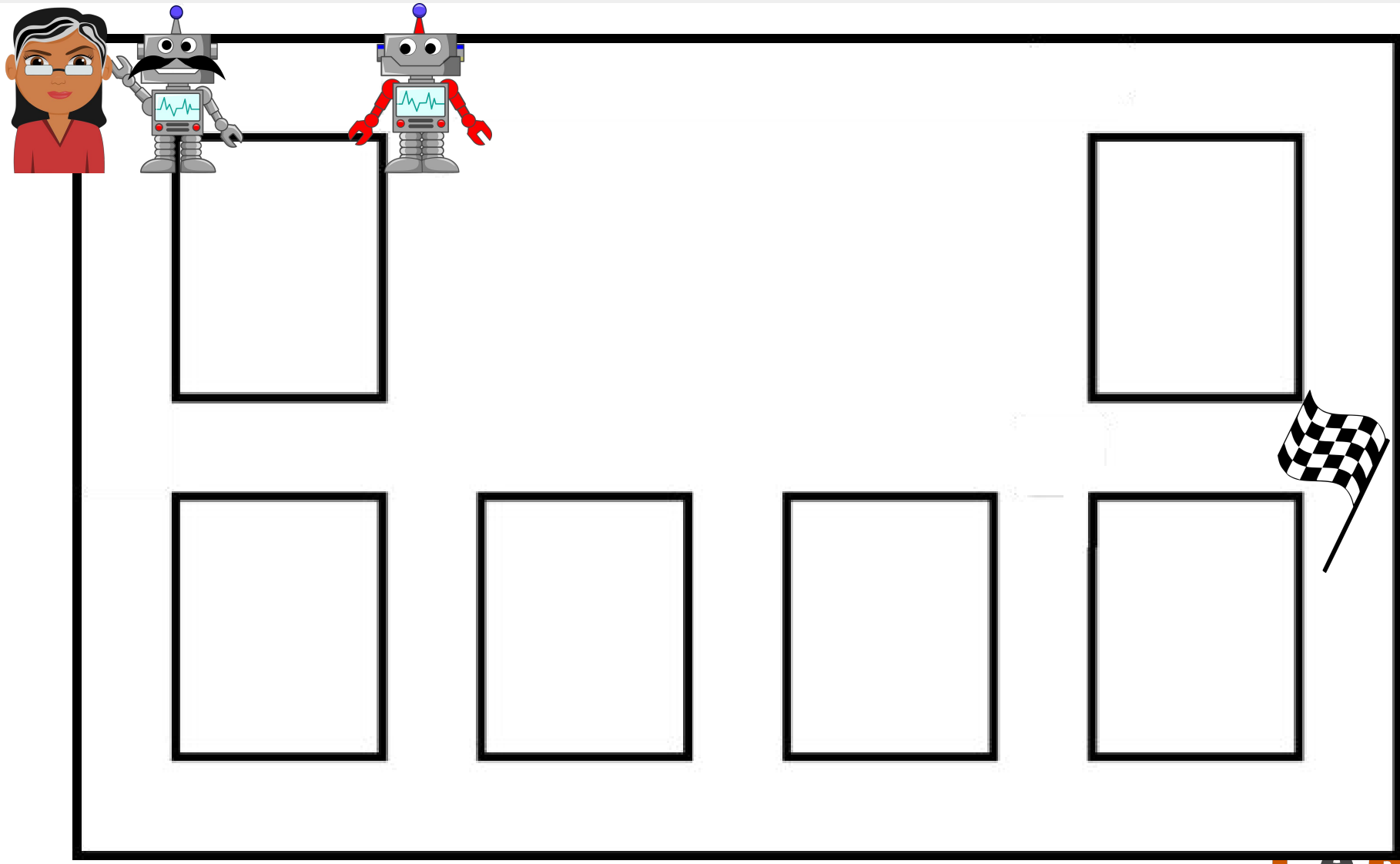
[1] Khandelwal and Stone. *Multi-Robot Human Guidance: Human Experiments and Multiple Concurrent Requests*. AAMAS 2017.
[2] Khandelwal, Barrett, and Stone. *Leading the Way: An Efficient Multi-Robot Human Guidance System.* AAMAS 2015.

LARG
Learning Agents Research Group
The University of Texas at Austin

Problem Introduction

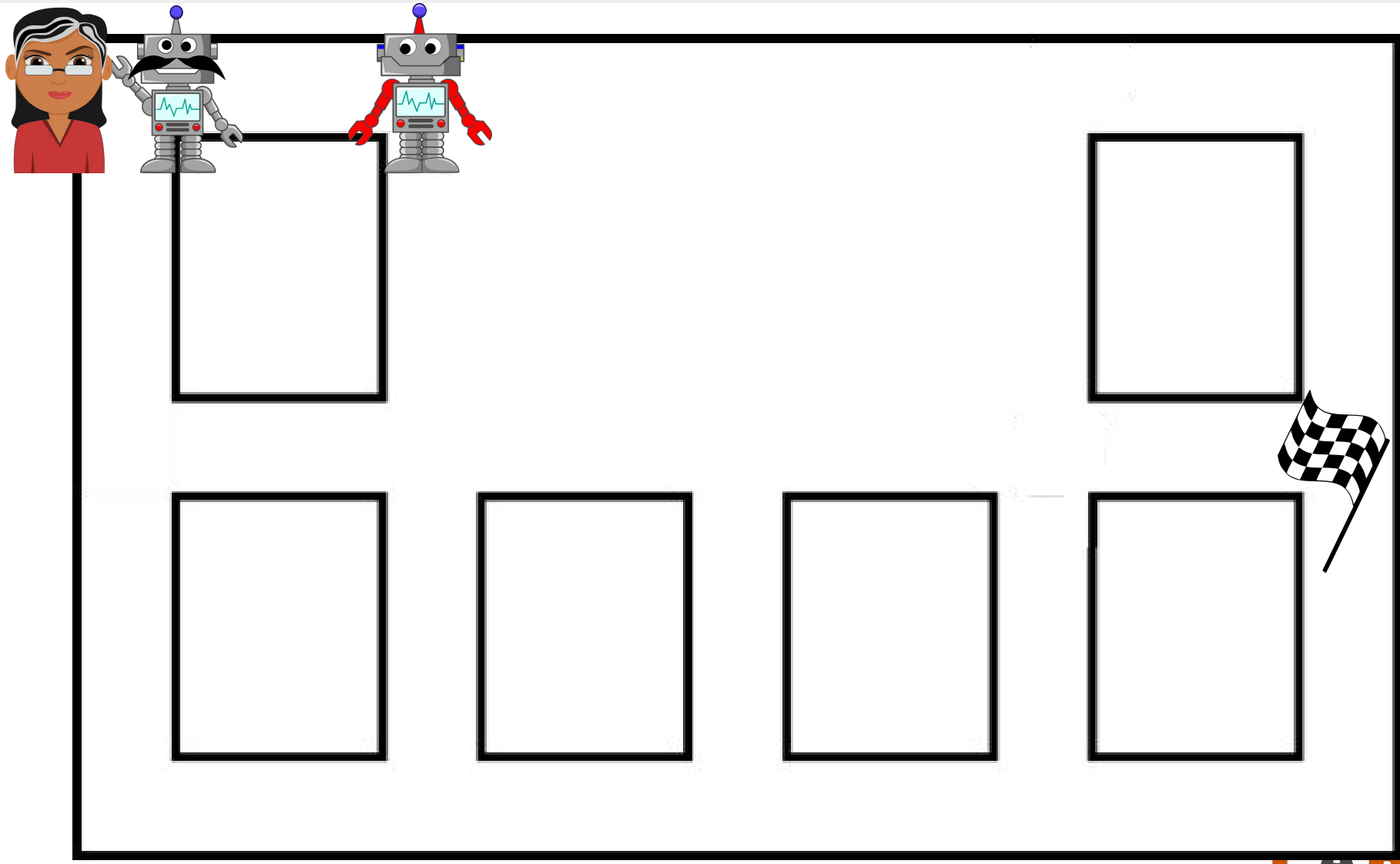


Problem Introduction

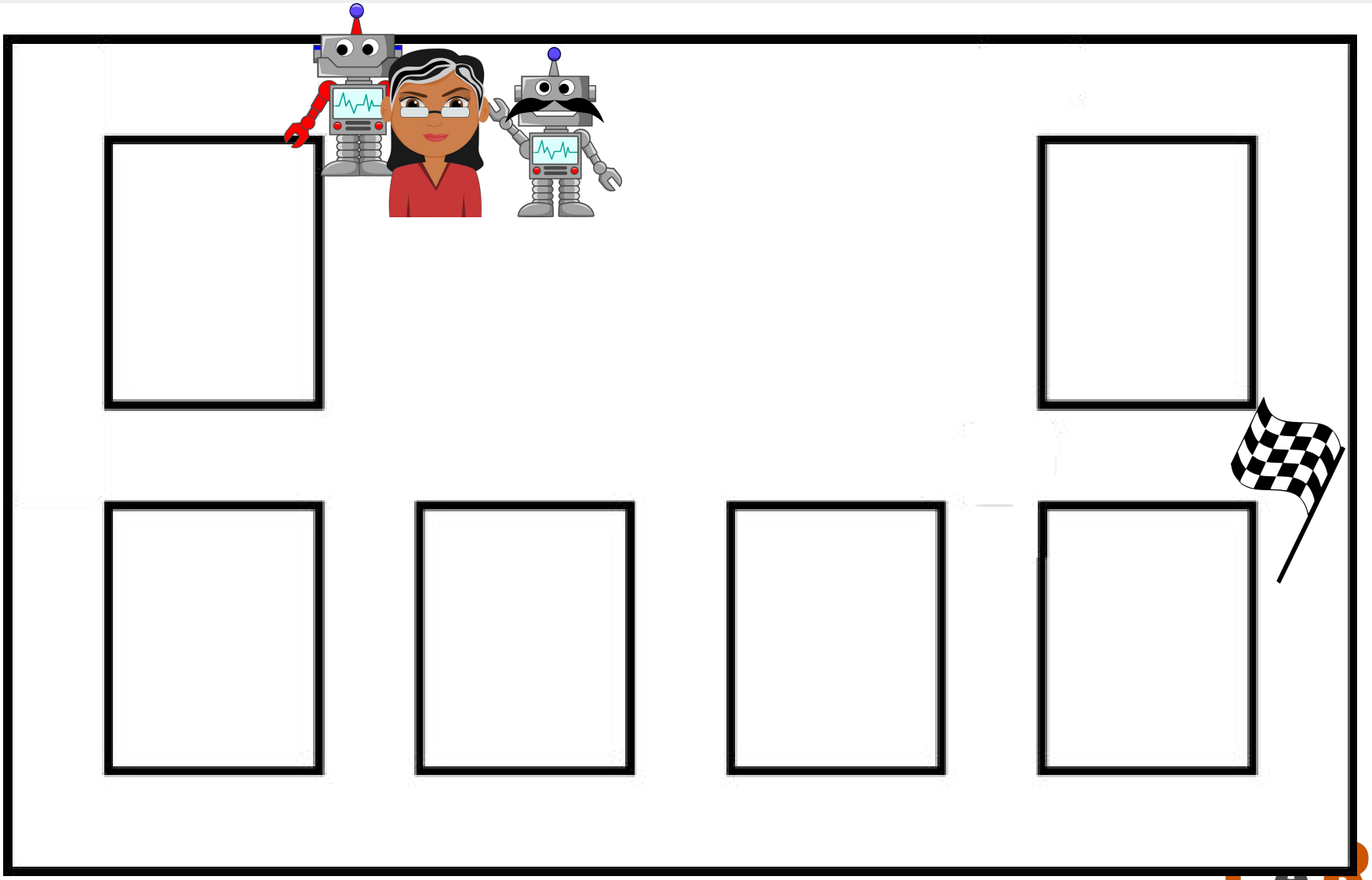


Learning Agents Research Group
The University of Texas at Austin

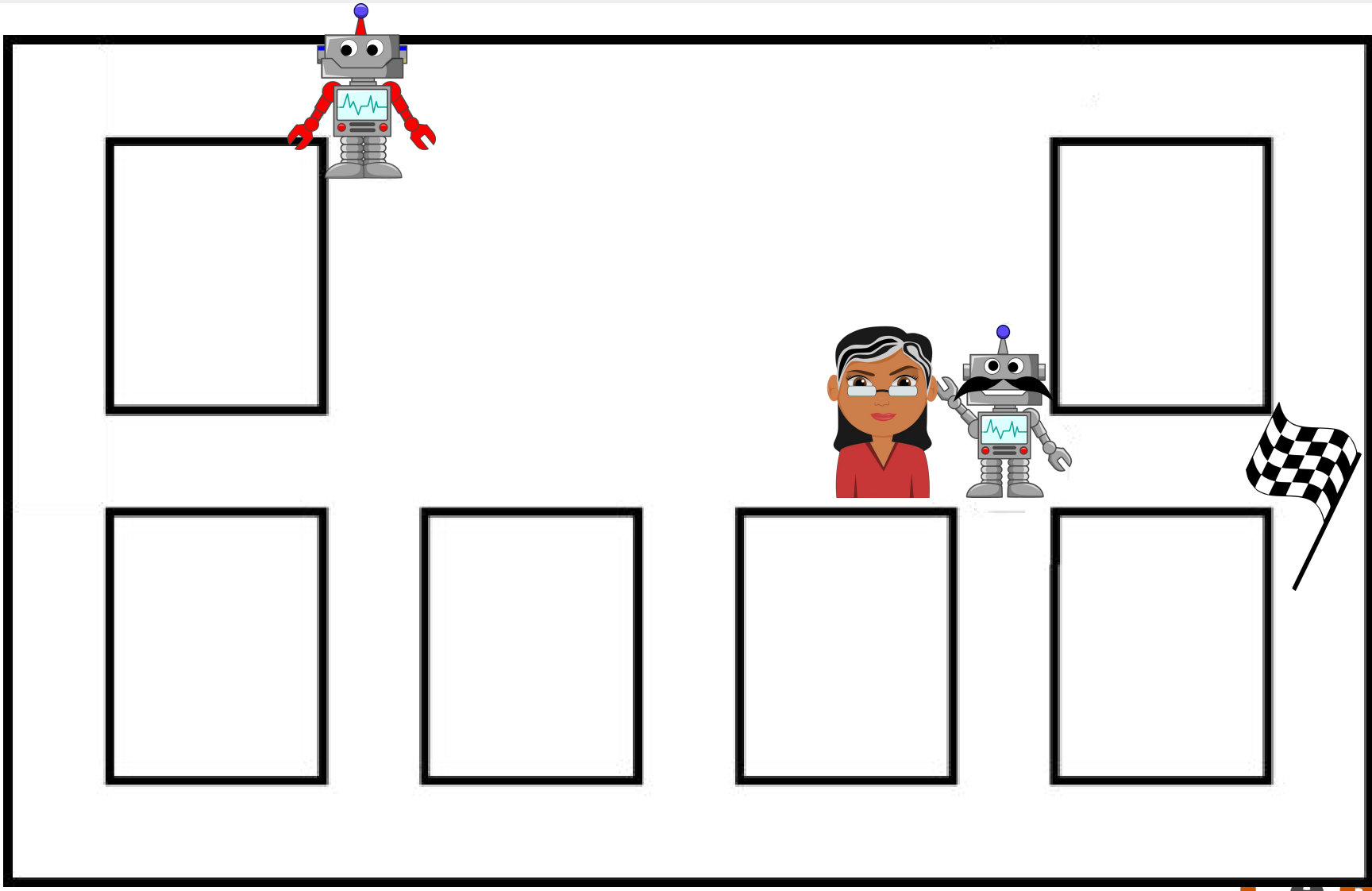
Easy Solution (SingleRobot)



Easy Solution (SingleRobot)

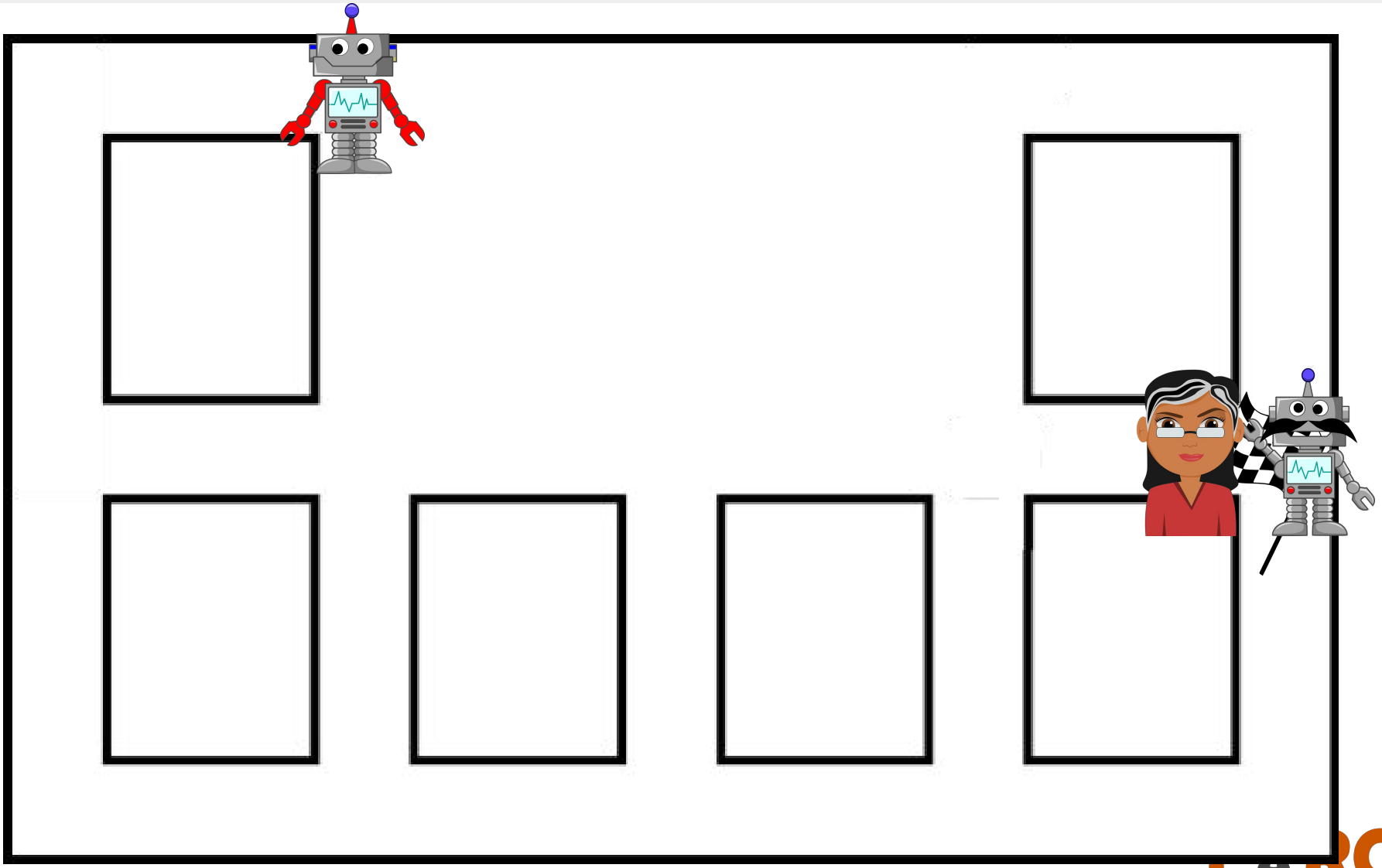


Easy Solution (SingleRobot)

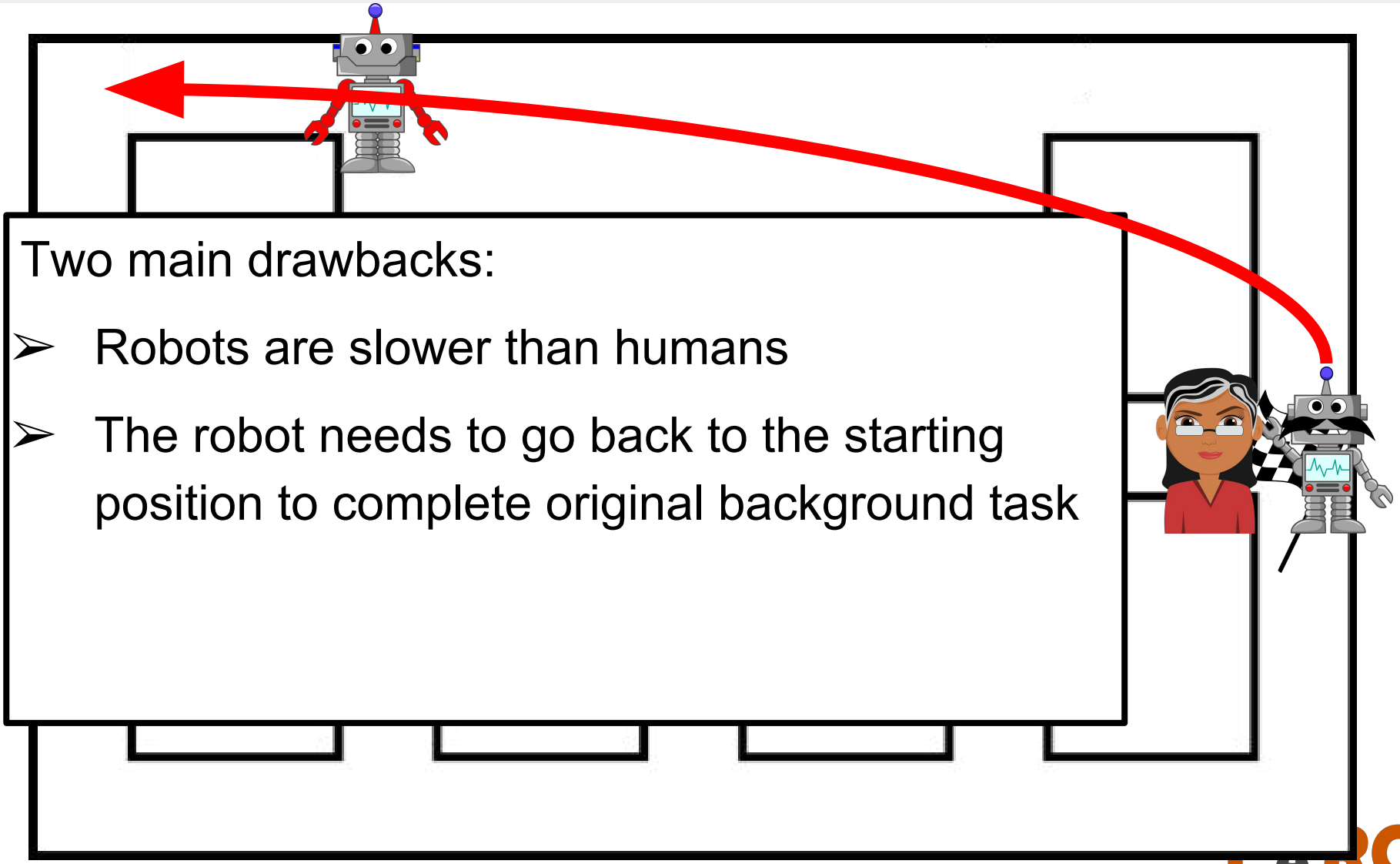


Learning Agents Research Group
The University of Texas at Austin

Easy Solution (SingleRobot)



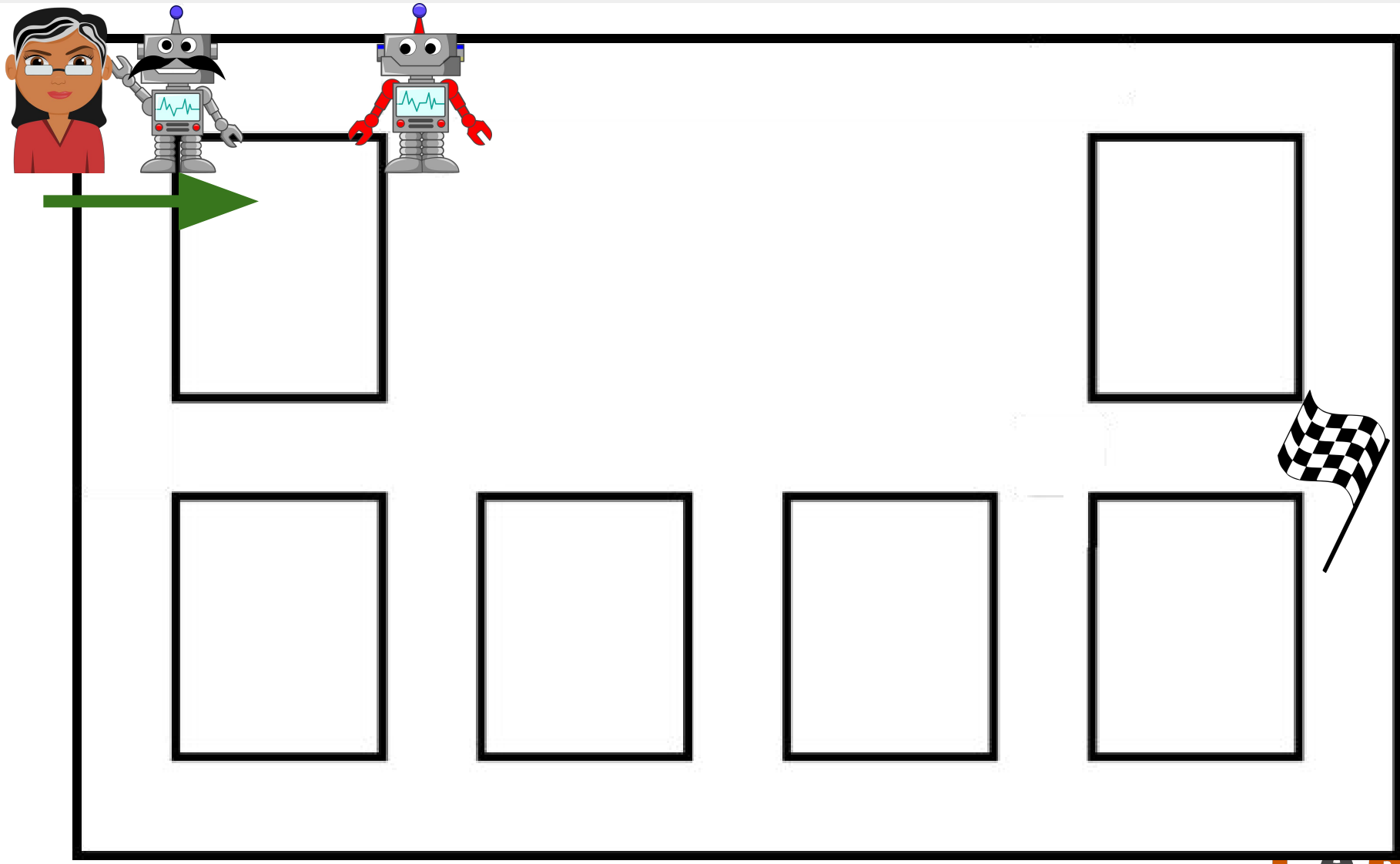
Easy Solution (SingleRobot)



Two main drawbacks:

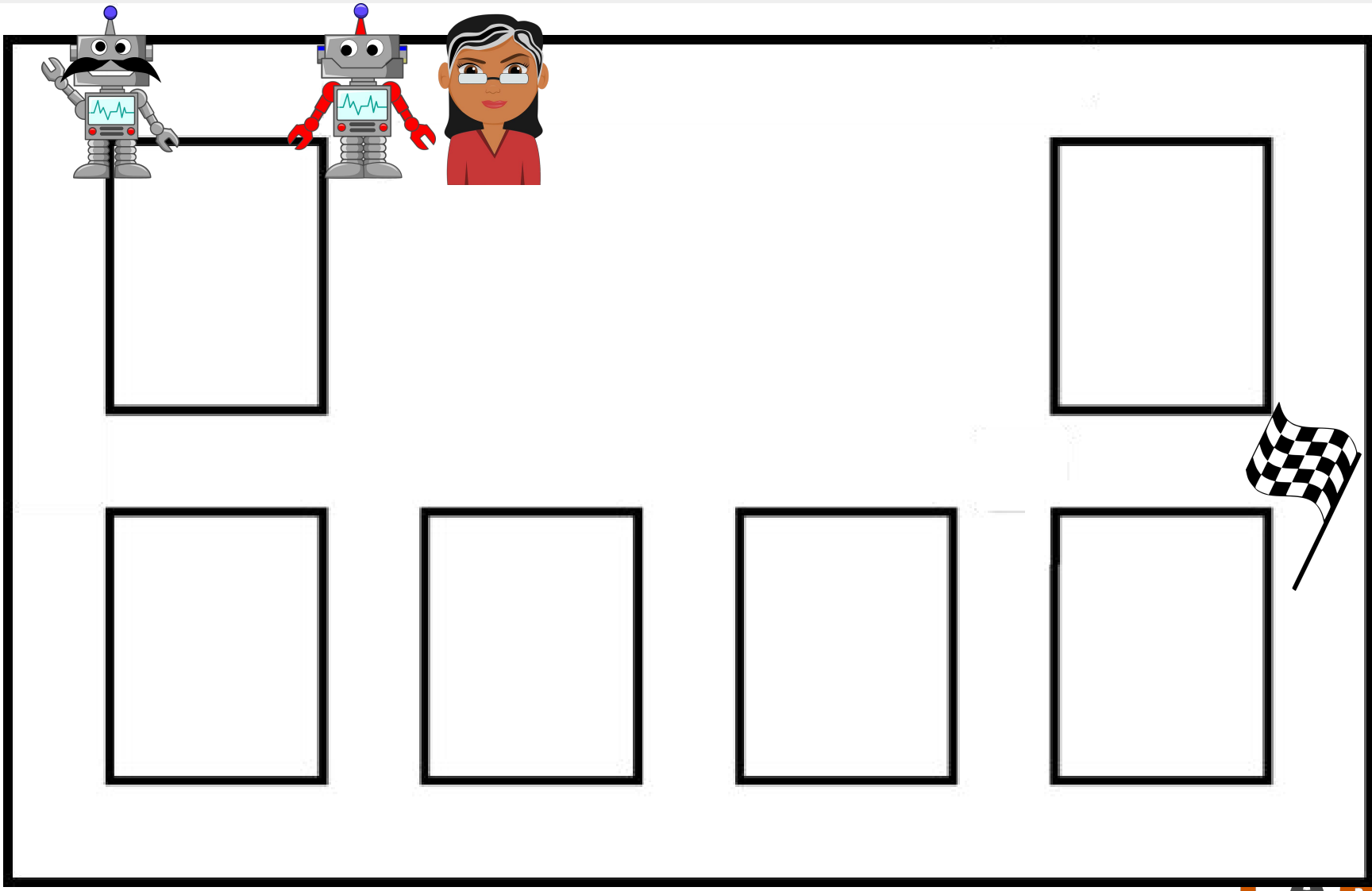
- Robots are slower than humans
- The robot needs to go back to the starting position to complete original background task

Desired Solution

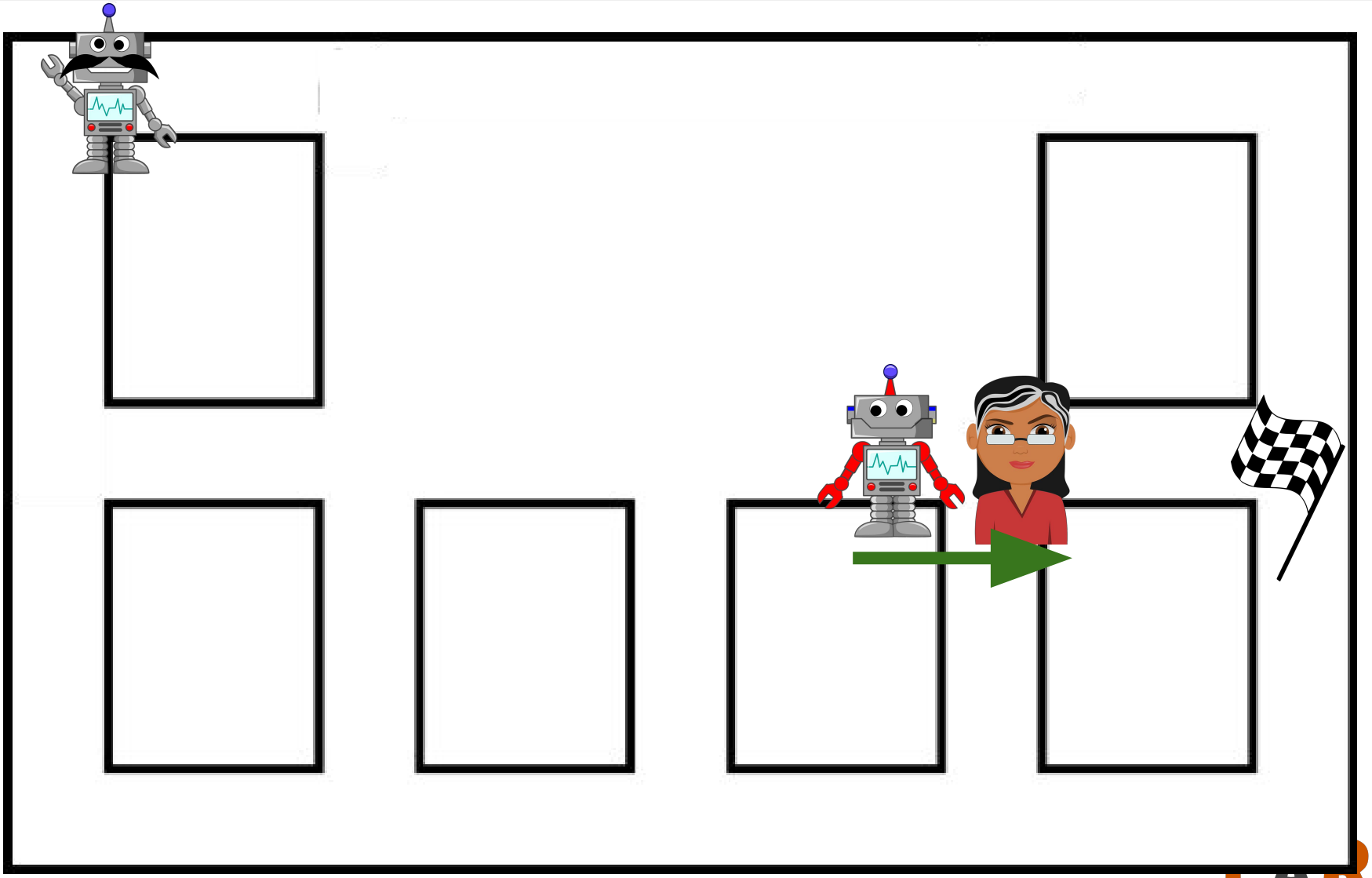


Learning Agents Research Group
The University of Texas at Austin

Desired Solution

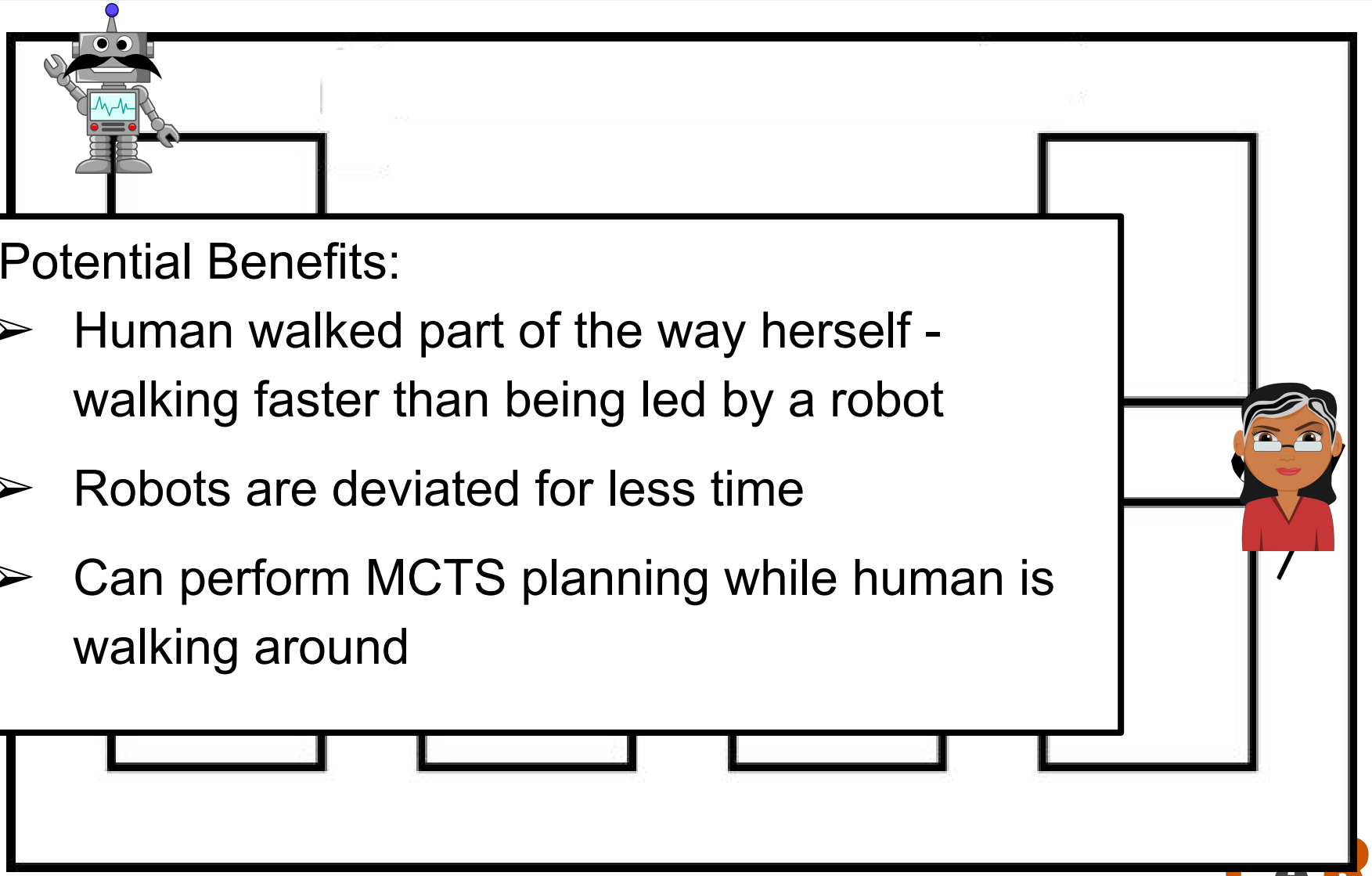


Desired Solution



Learning Agents Research Group
The University of Texas at Austin

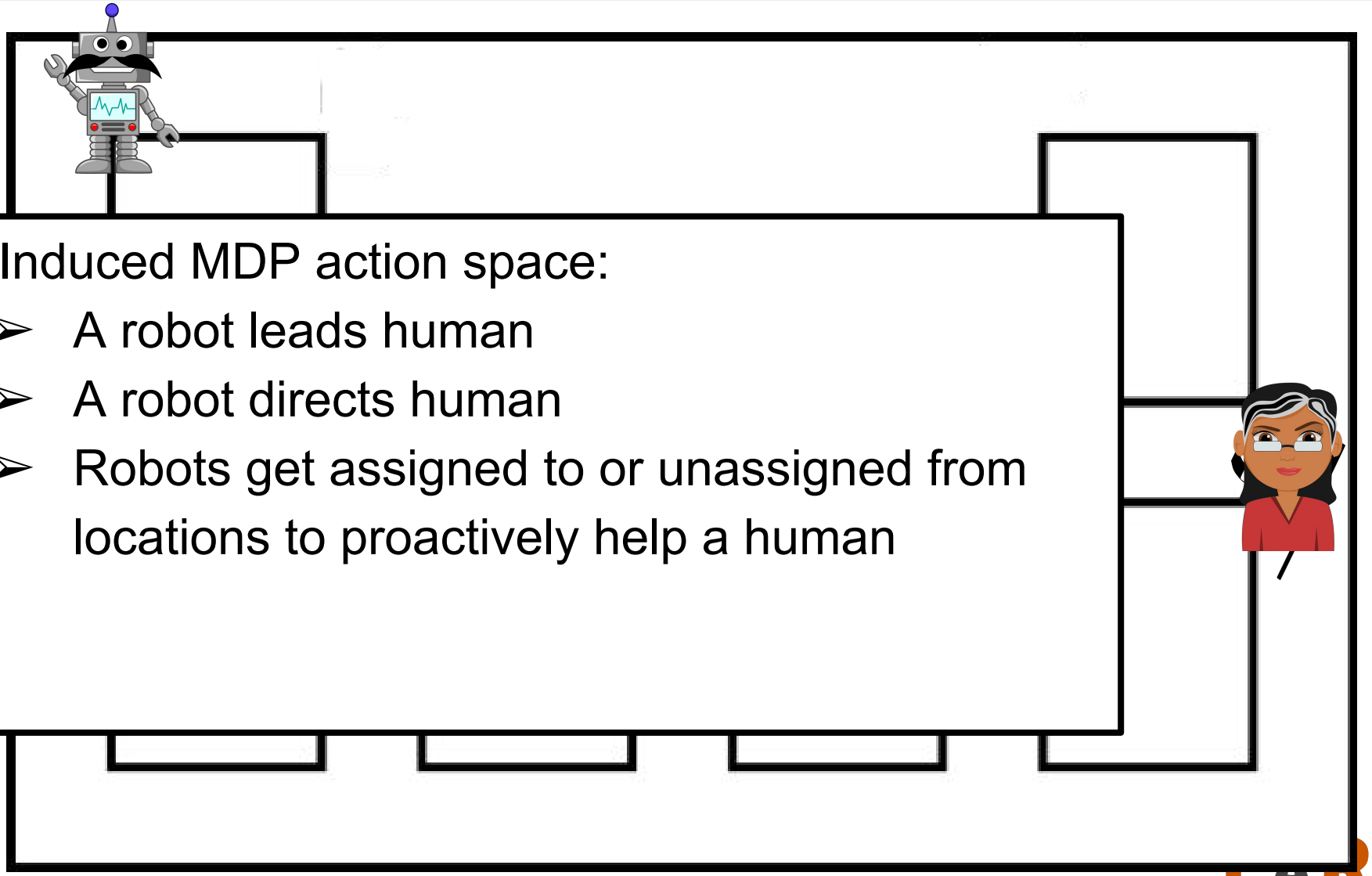
Desired Solution



Potential Benefits:

- Human walked part of the way herself - walking faster than being led by a robot
- Robots are deviated for less time
- Can perform MCTS planning while human is walking around

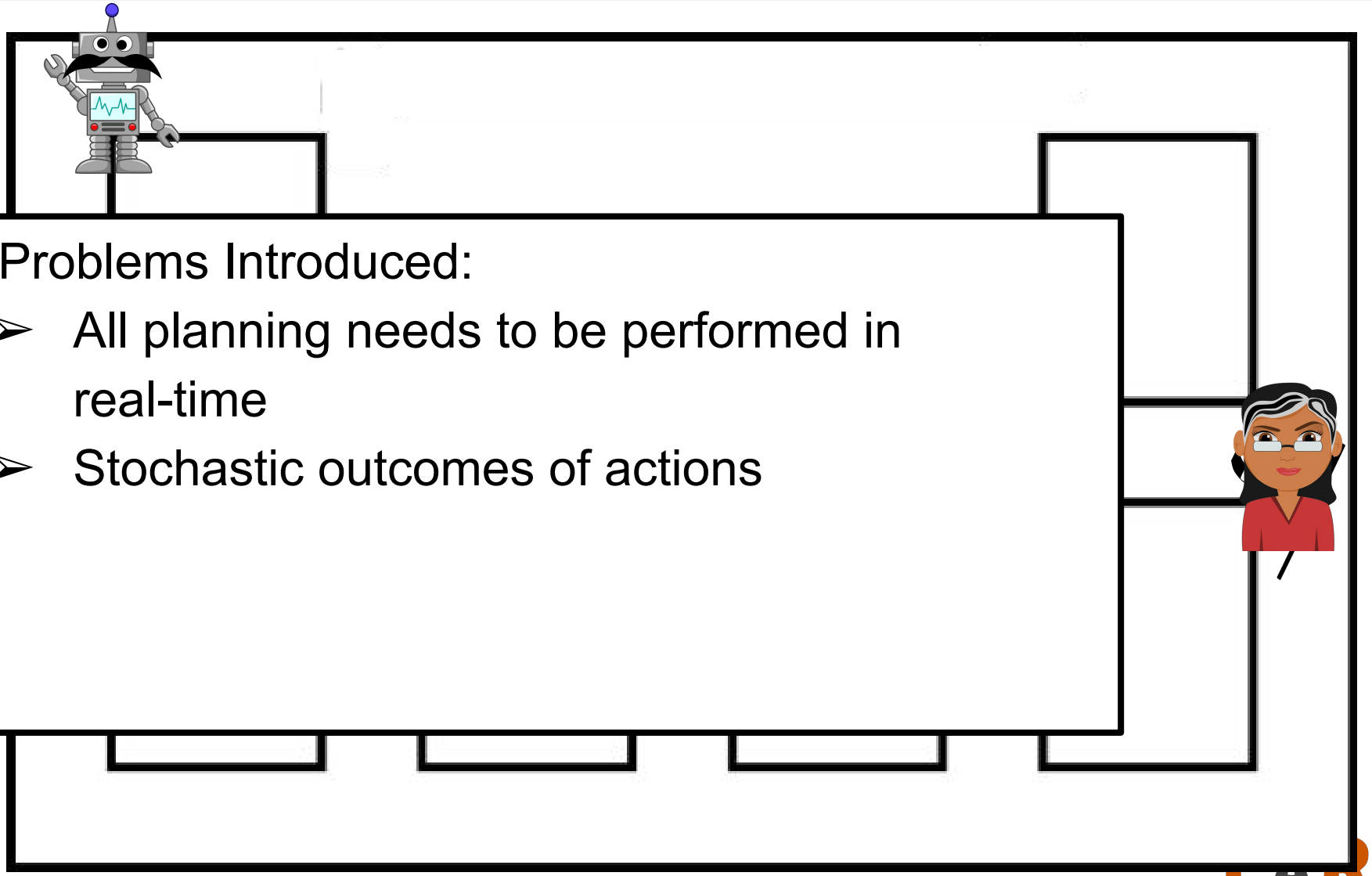
Desired Solution



Induced MDP action space:

- A robot leads human
- A robot directs human
- Robots get assigned to or unassigned from locations to proactively help a human

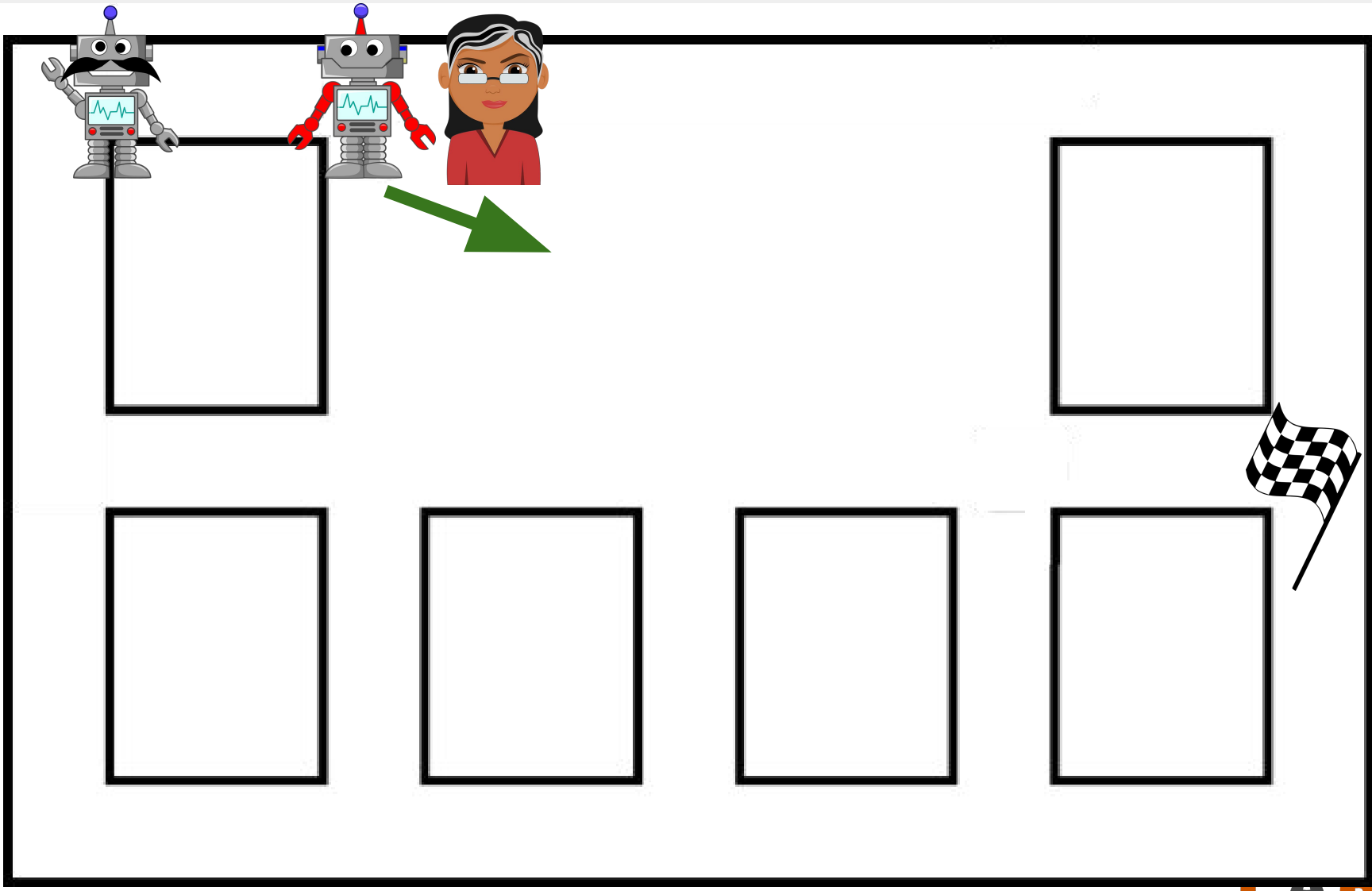
Desired Solution



Problems Introduced:

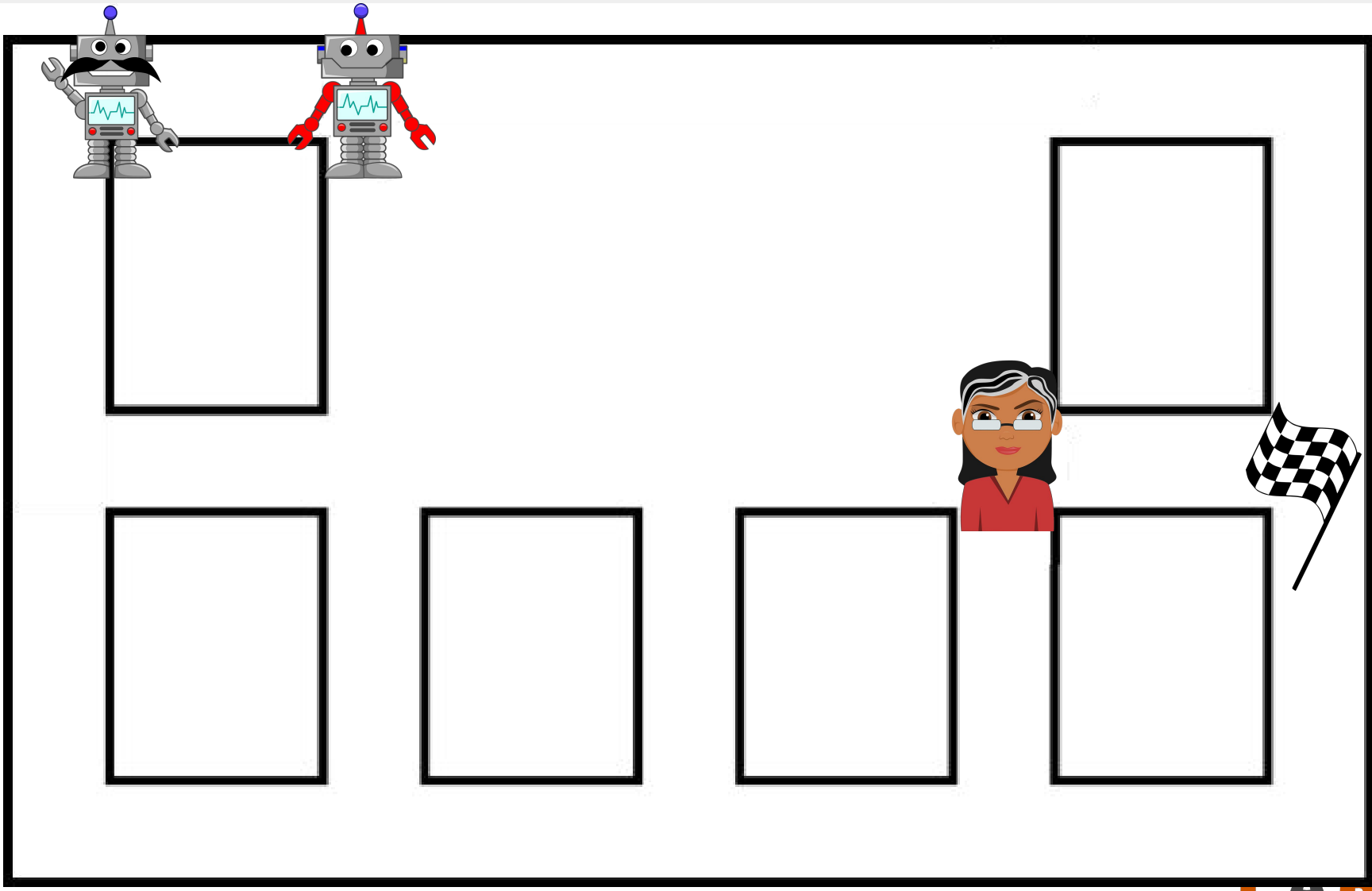
- All planning needs to be performed in real-time
- Stochastic outcomes of actions

Stochastic Outcomes



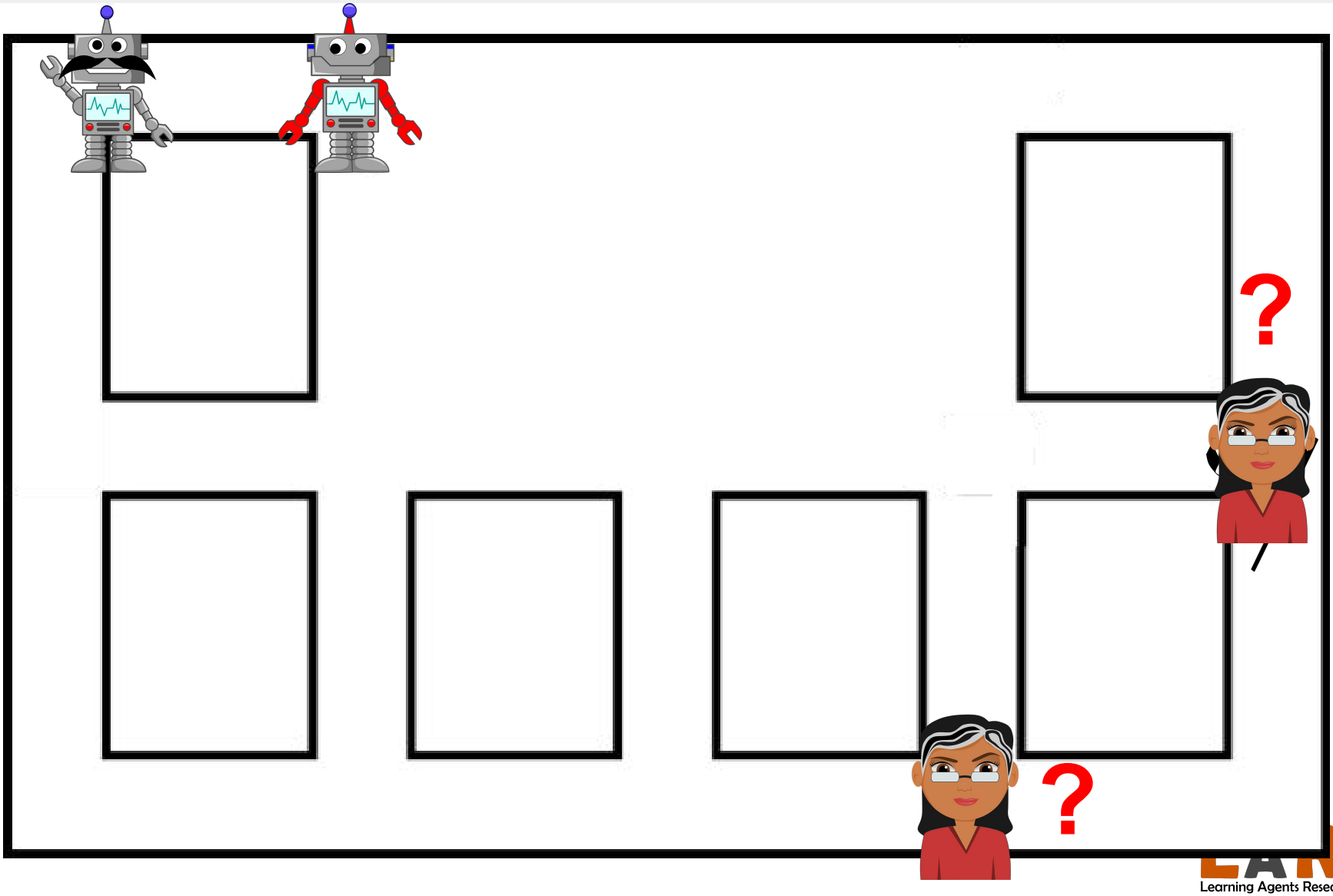
Learning Agents Research Group
The University of Texas at Austin

Stochastic Outcomes



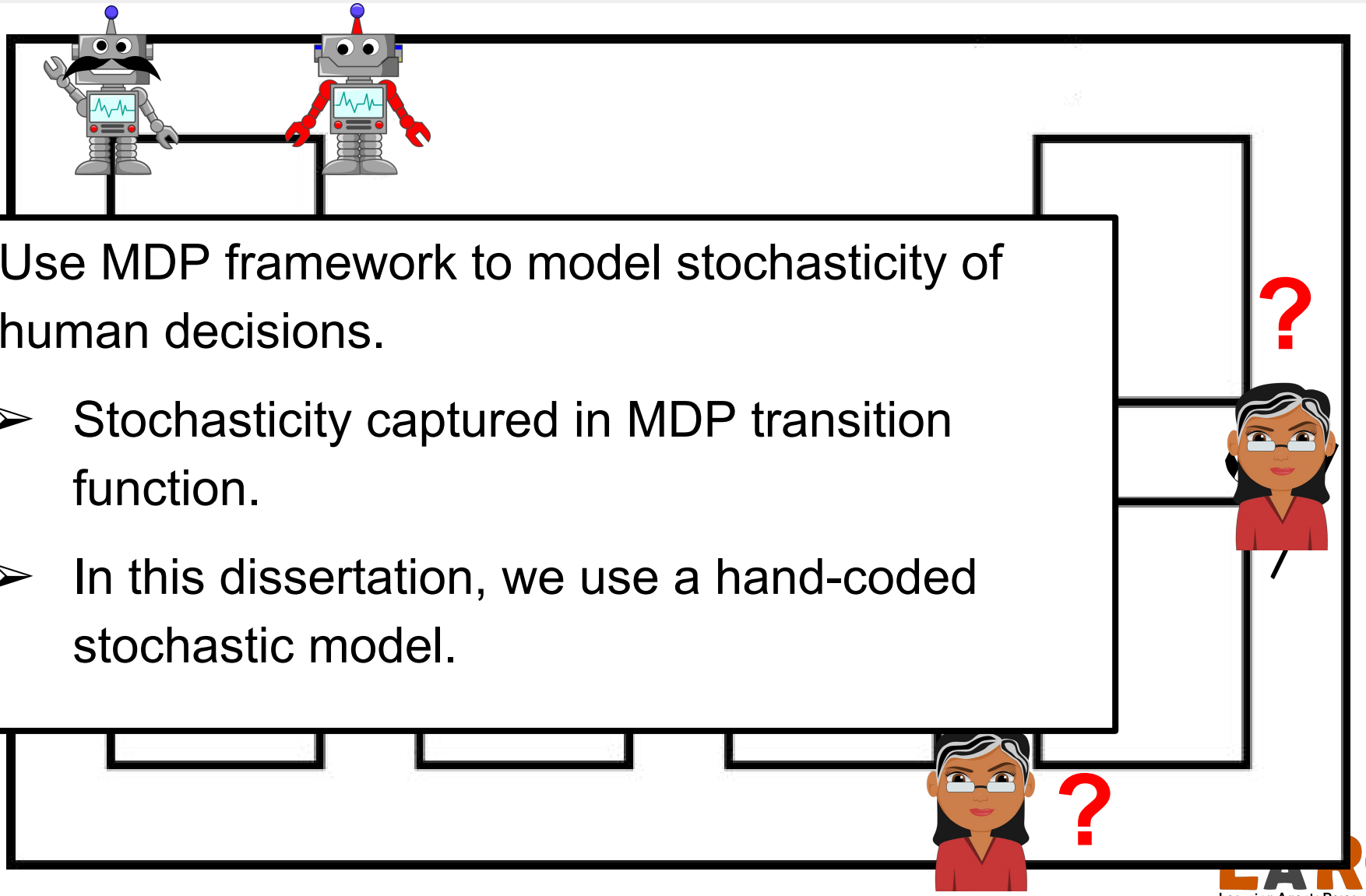
Learning Agents Research Group
The University of Texas at Austin

Stochastic Outcomes

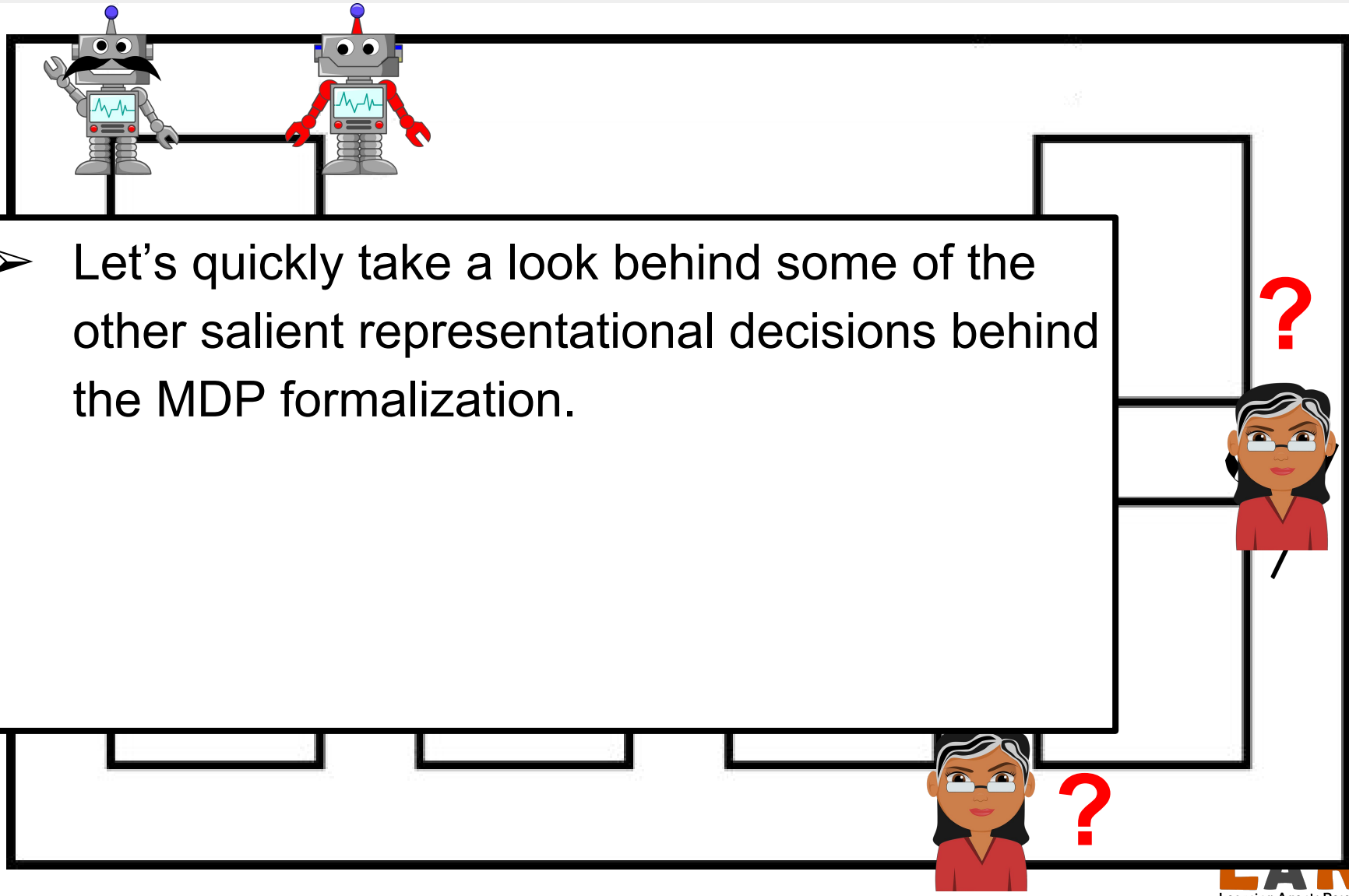


Learning Agents Research Group
The University of Texas at Austin

Stochastic Outcomes

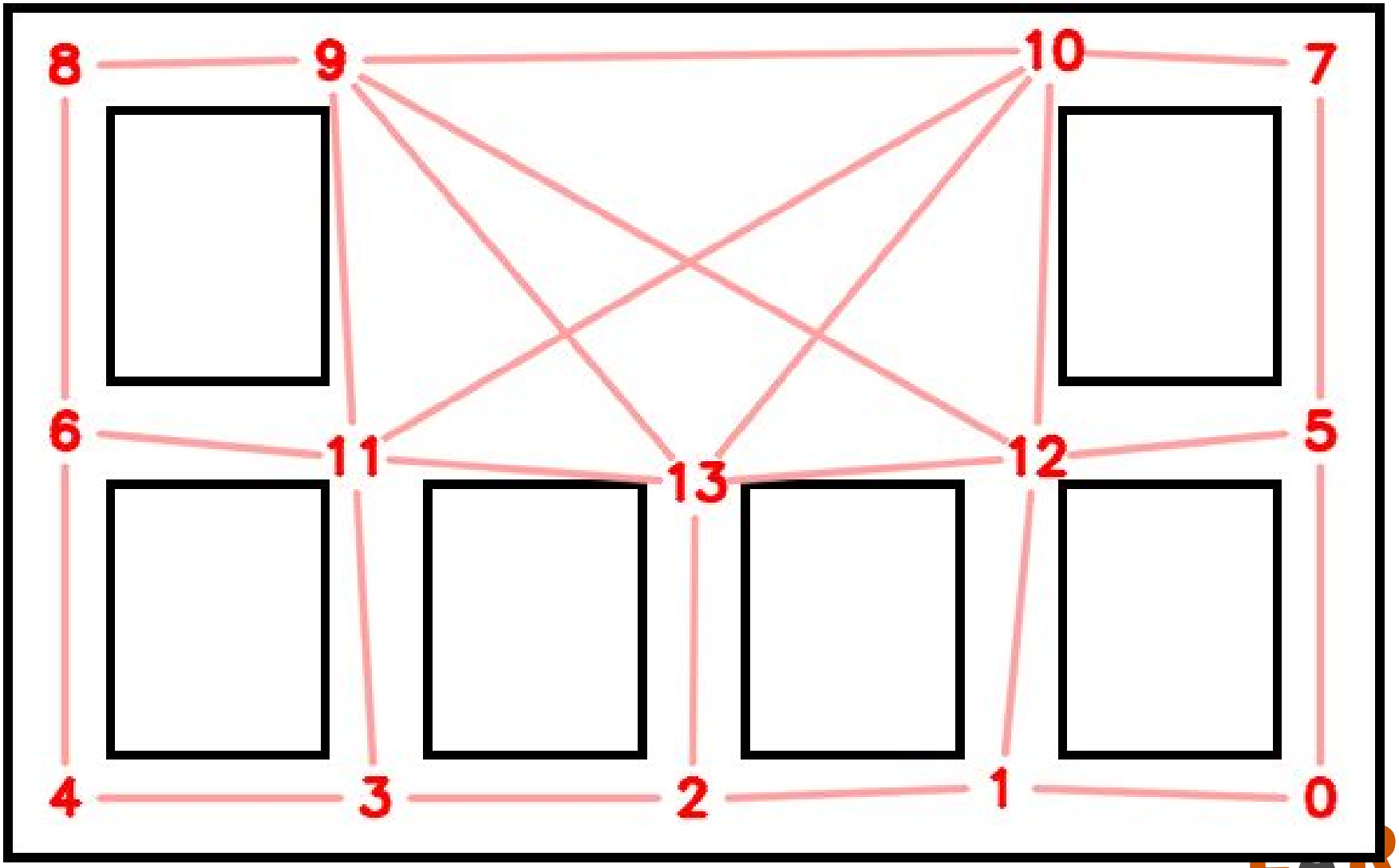


Salient Representational Decisions

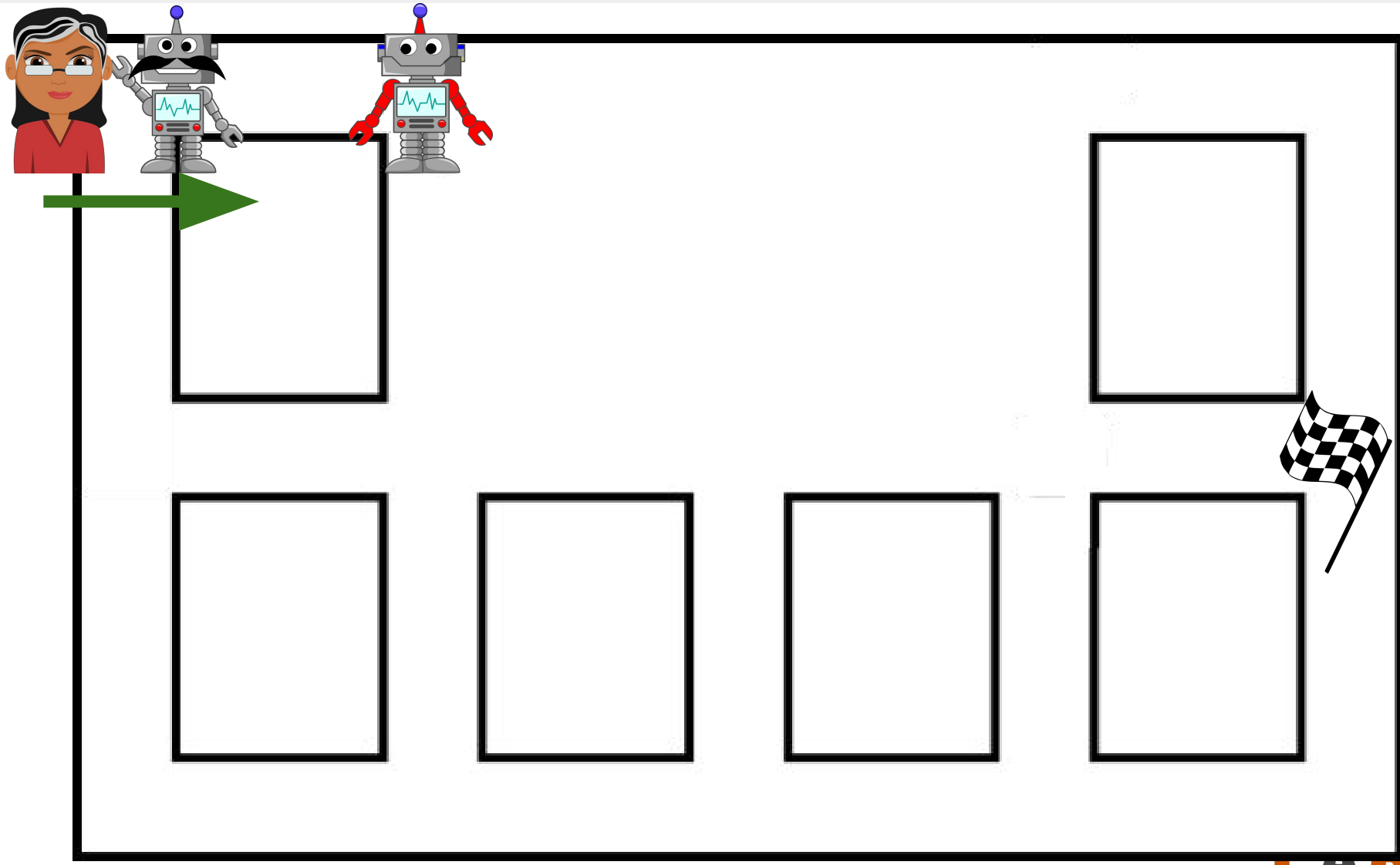


- Let's quickly take a look behind some of the other salient representational decisions behind the MDP formalization.

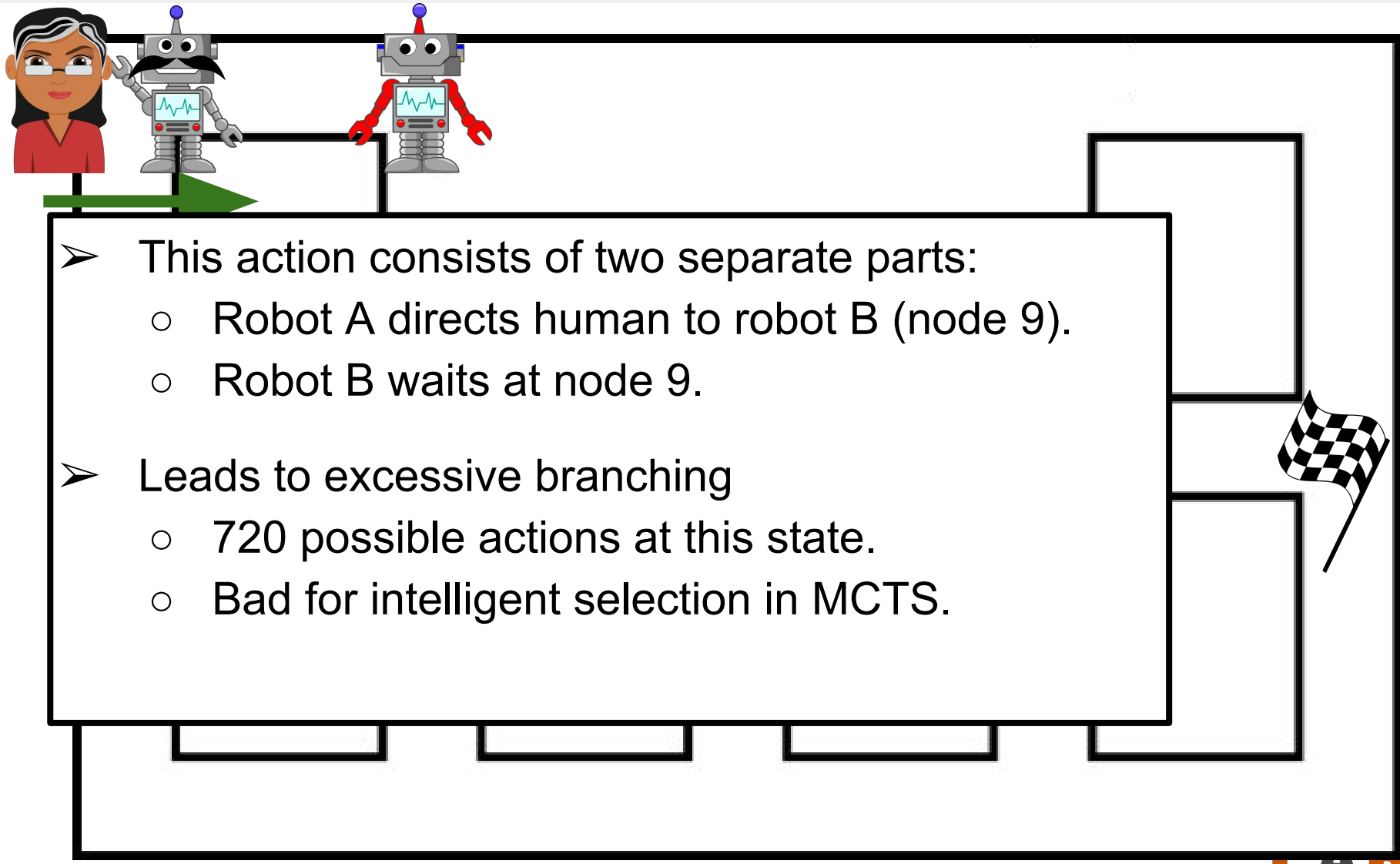
Topological Representation



Action Decomposition

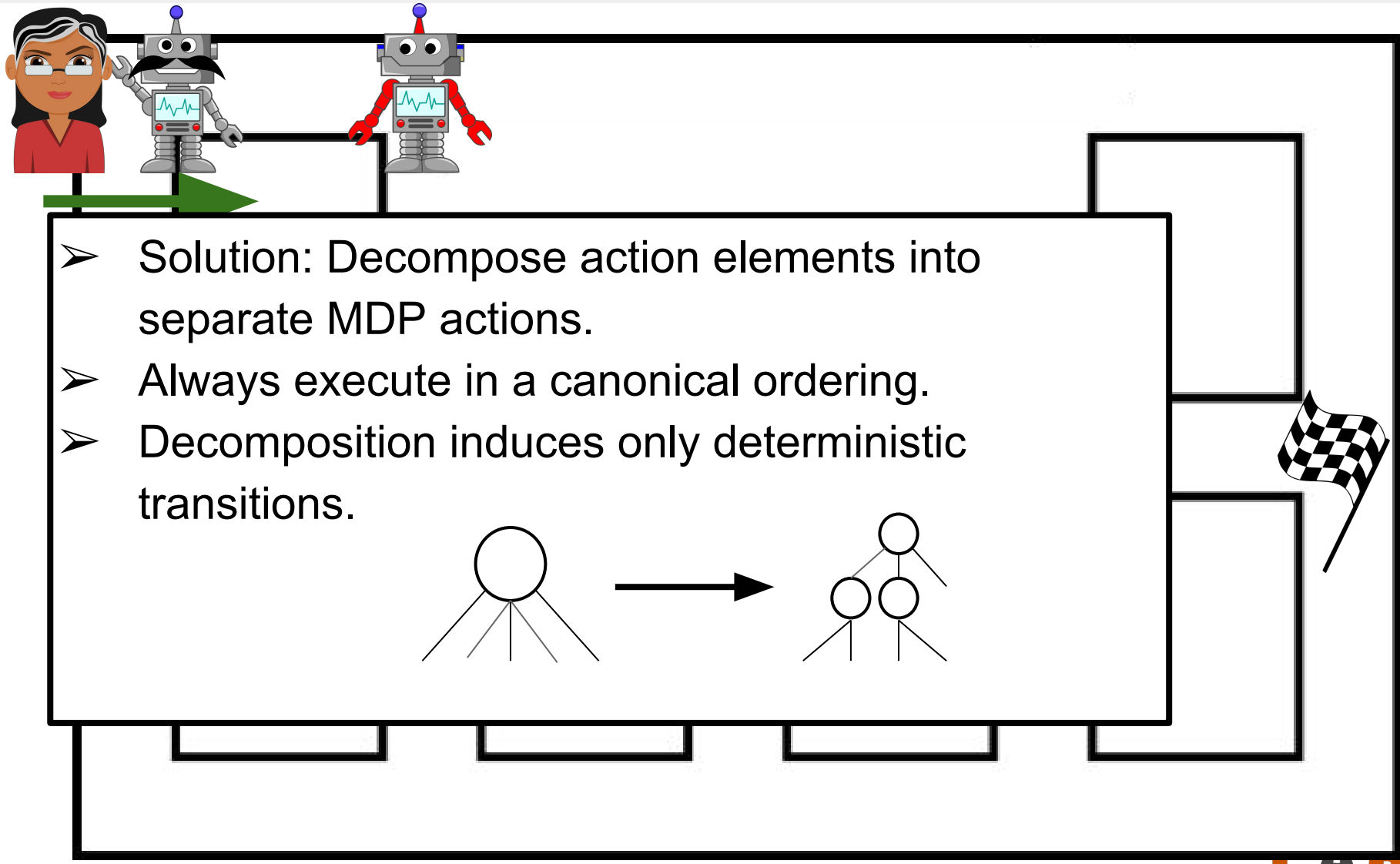


Action Decomposition



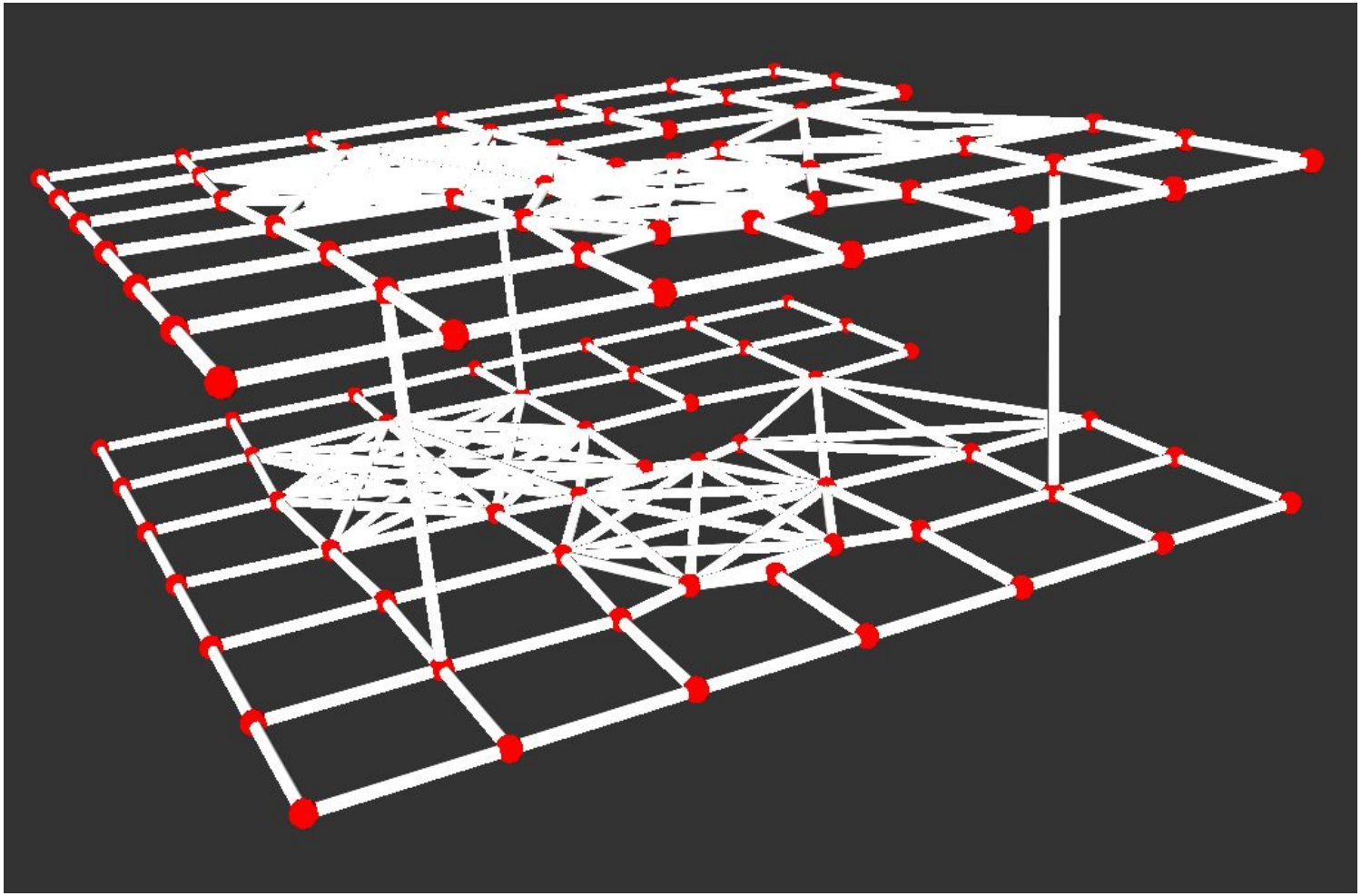
- This action consists of two separate parts:
  - Robot A directs human to robot B (node 9).
  - Robot B waits at node 9.
- Leads to excessive branching
  - 720 possible actions at this state.
  - Bad for intelligent selection in MCTS.

Action Decomposition



- Solution: Decompose action elements into separate MDP actions.
- Always execute in a canonical ordering.
- Decomposition induces only deterministic transitions.

Multiple floors and concurrent requests



MDP Reward - Linear Weighted Combination

- Reward function needs to balance multiple concurrent requests with background tasks of the robots:

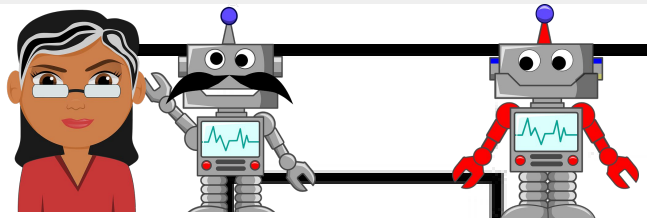
$$R_{ss'} = \underbrace{-\sum_{i=0}^{n} u_i \Delta t}_{\text{on-demand time loss}} - \overbrace{\sum_{r} U^r_{ss'}}^{\text{background utility loss}}$$

- The background utility loss can be calculated using a model:

$$U^r_{ss'} = \bar{\tau}_u(timeToDest(r_{s'}, r_s.\tau_d) + \Delta t - timeToDest(r_s, r_s.\tau_d))$$

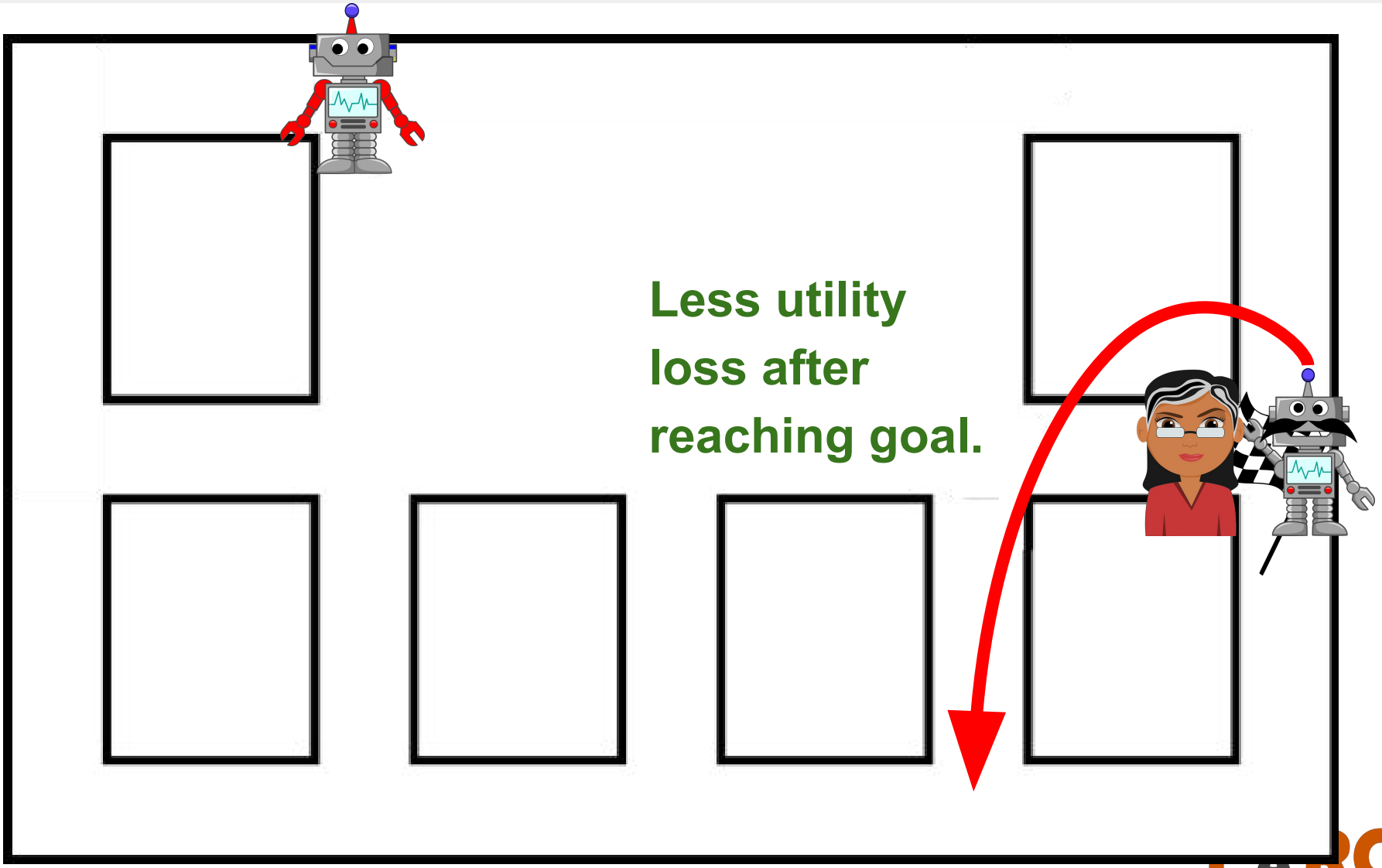
- $\bar{\tau}_u$ is the average background task utility

SingleRobotW Baseline

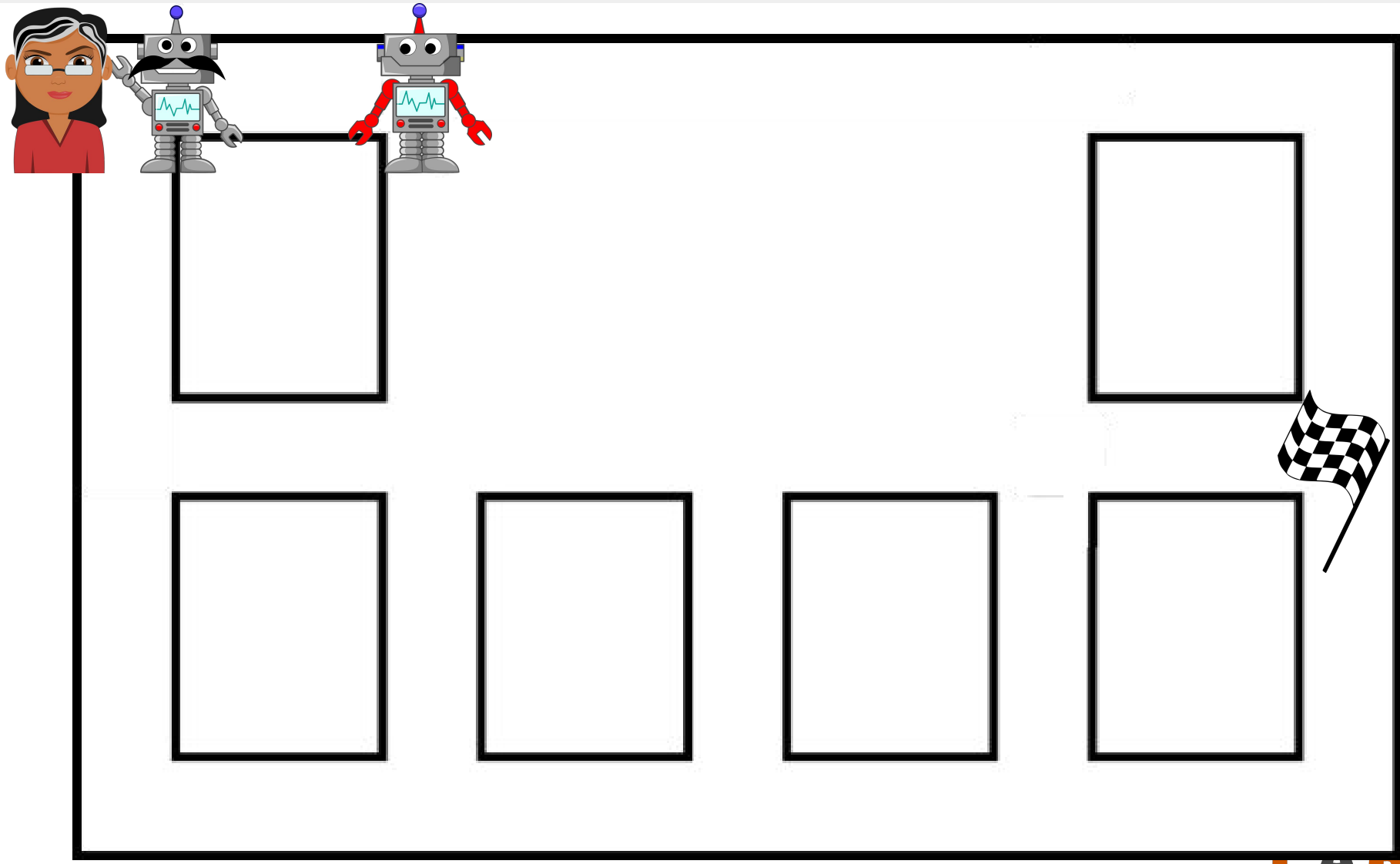


Complete background task first

SingleRobotW Baseline

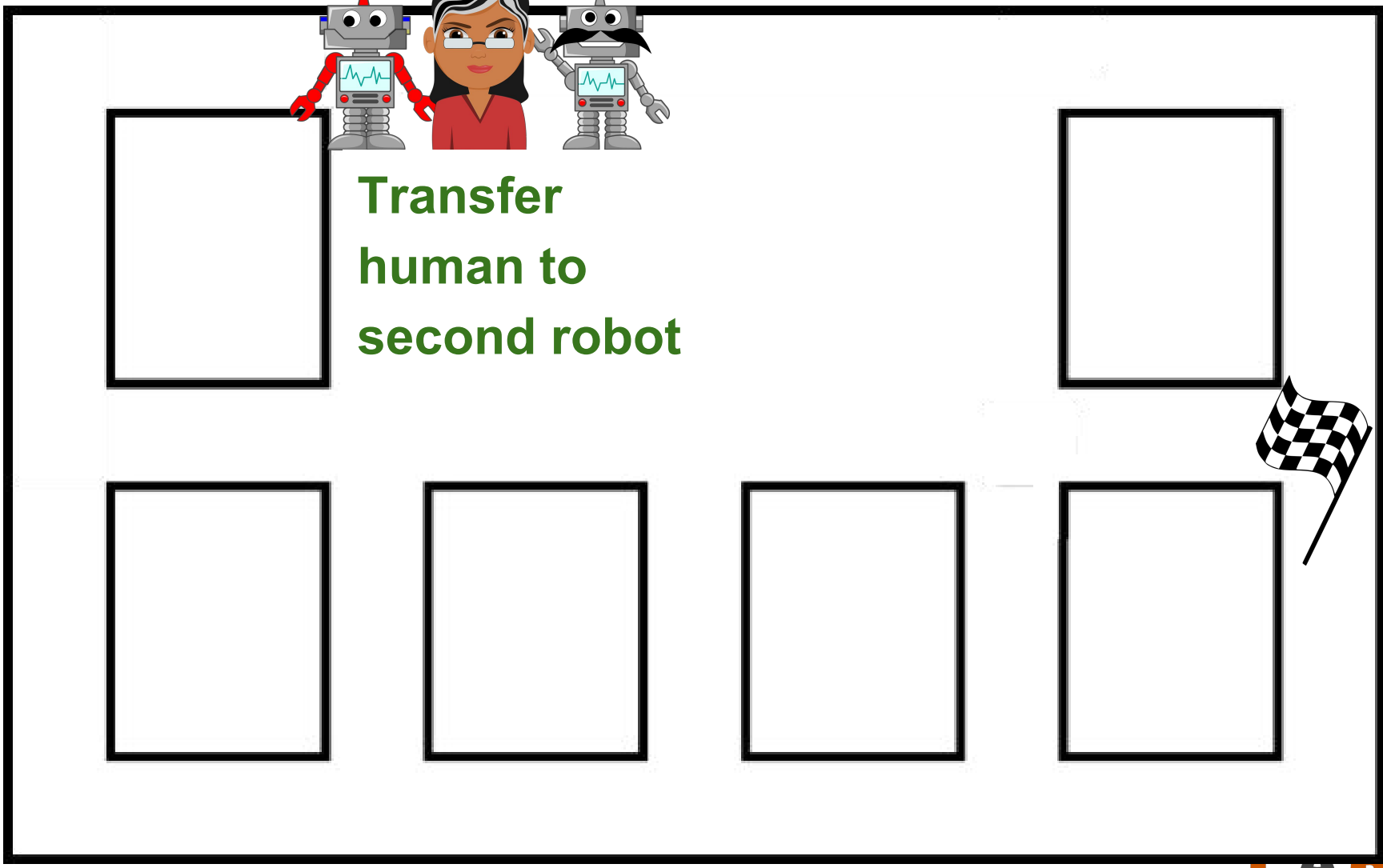


PDP-T [Coltin et al. 2015] baseline

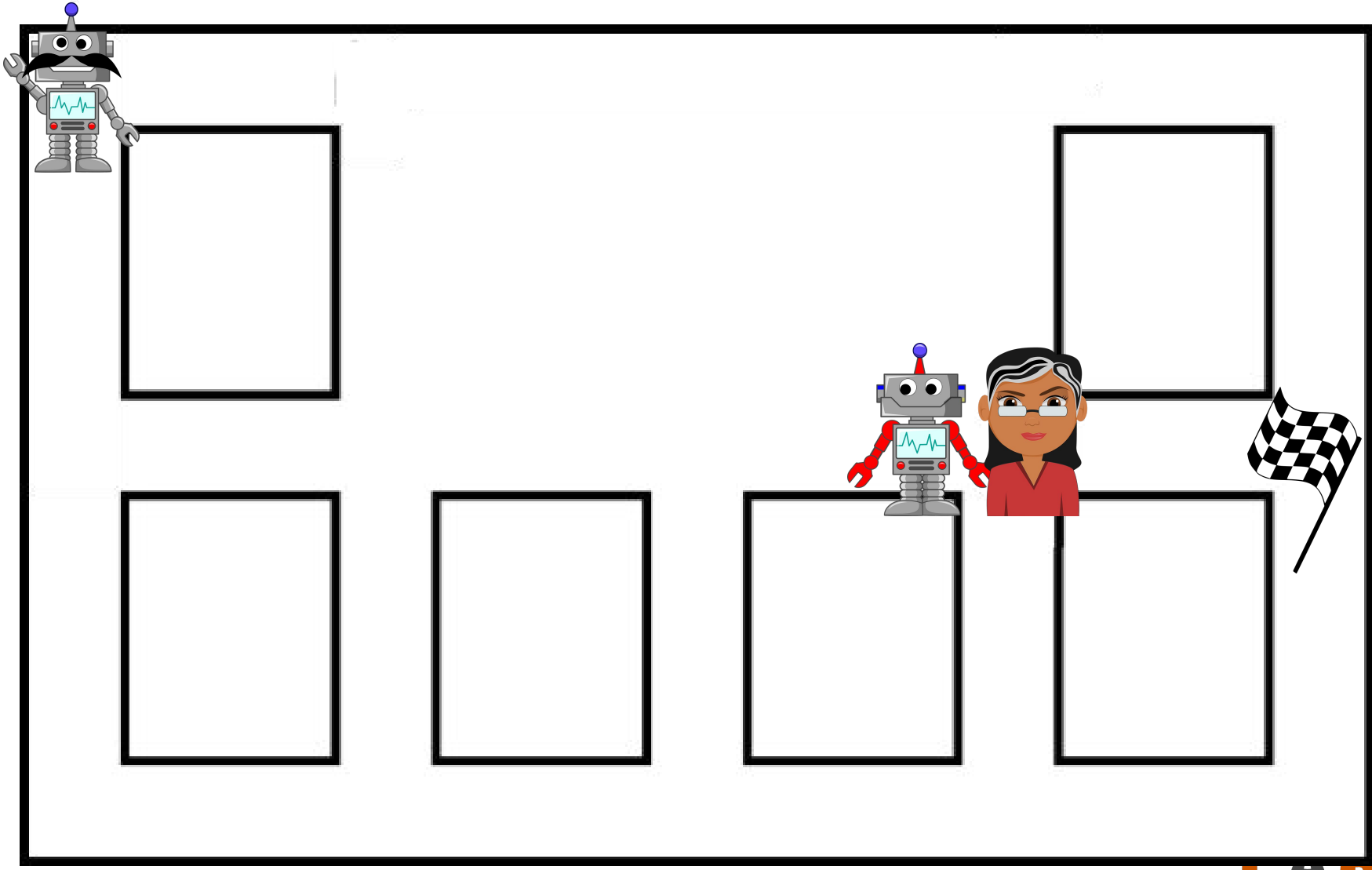


Learning Agents Research Group
The University of Texas at Austin

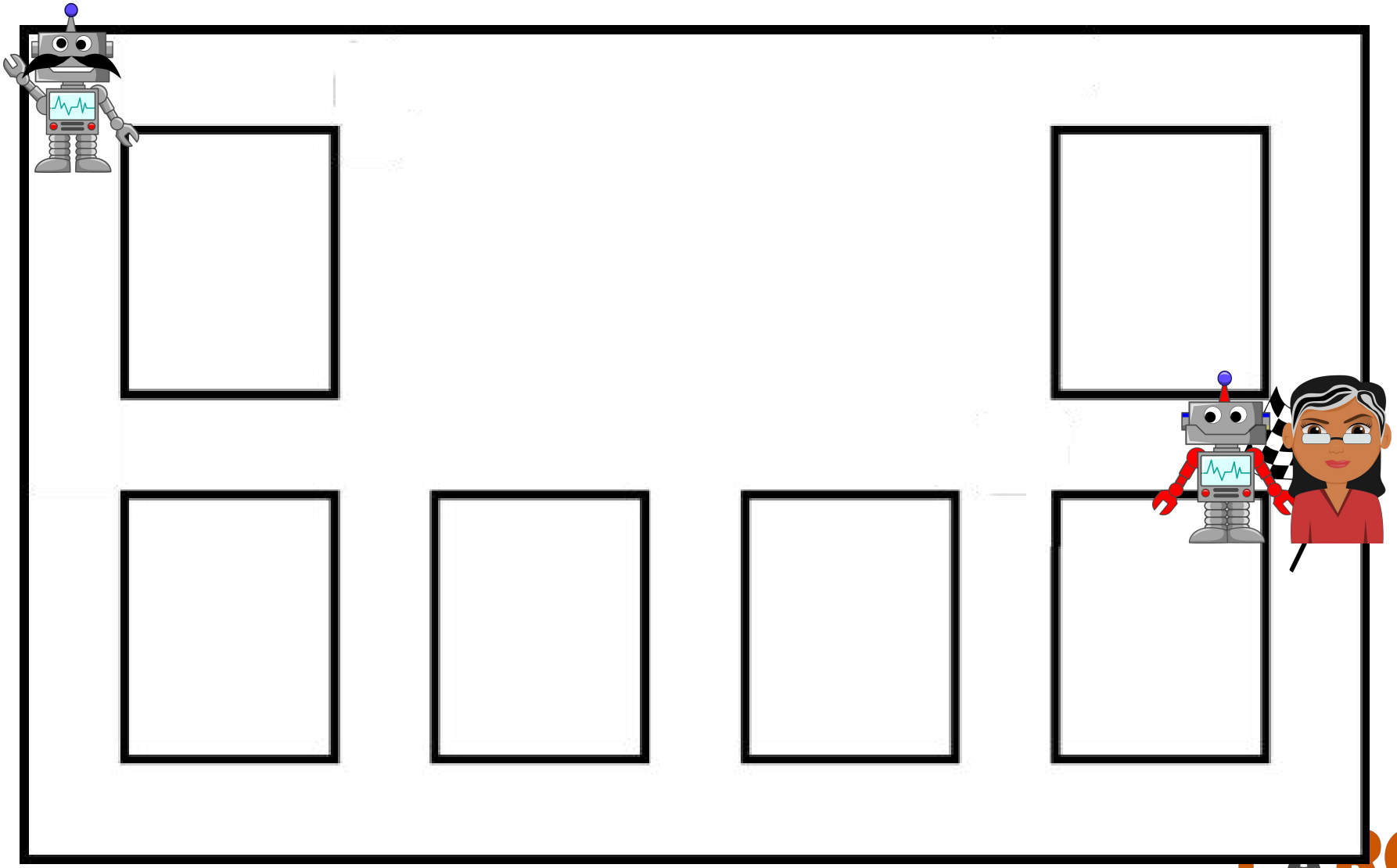
PDP-T [Coltin et al. 2015] baseline



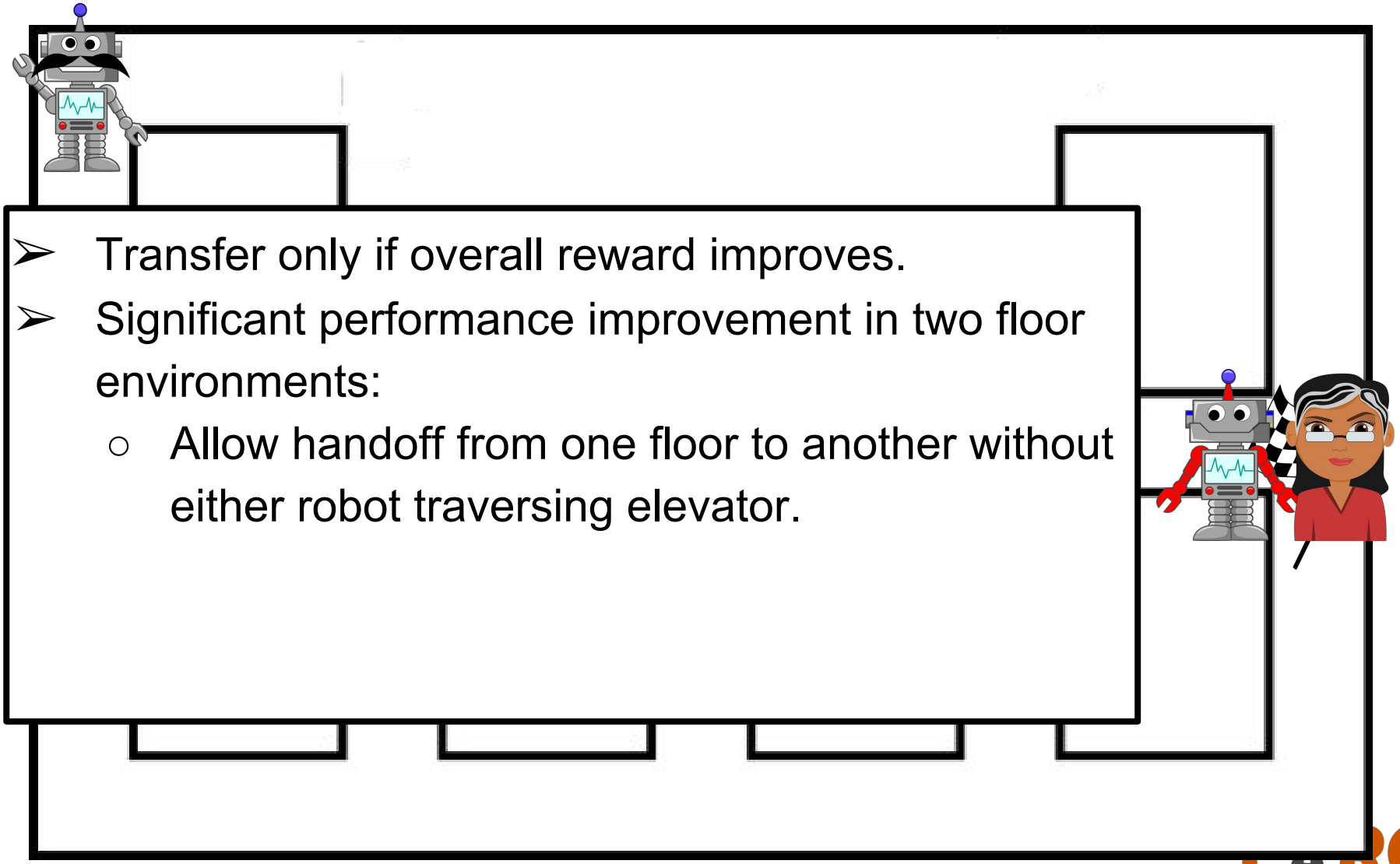
PDP-T [Coltin et al. 2015] baseline



PDP-T [Coltin et al. 2015] baseline

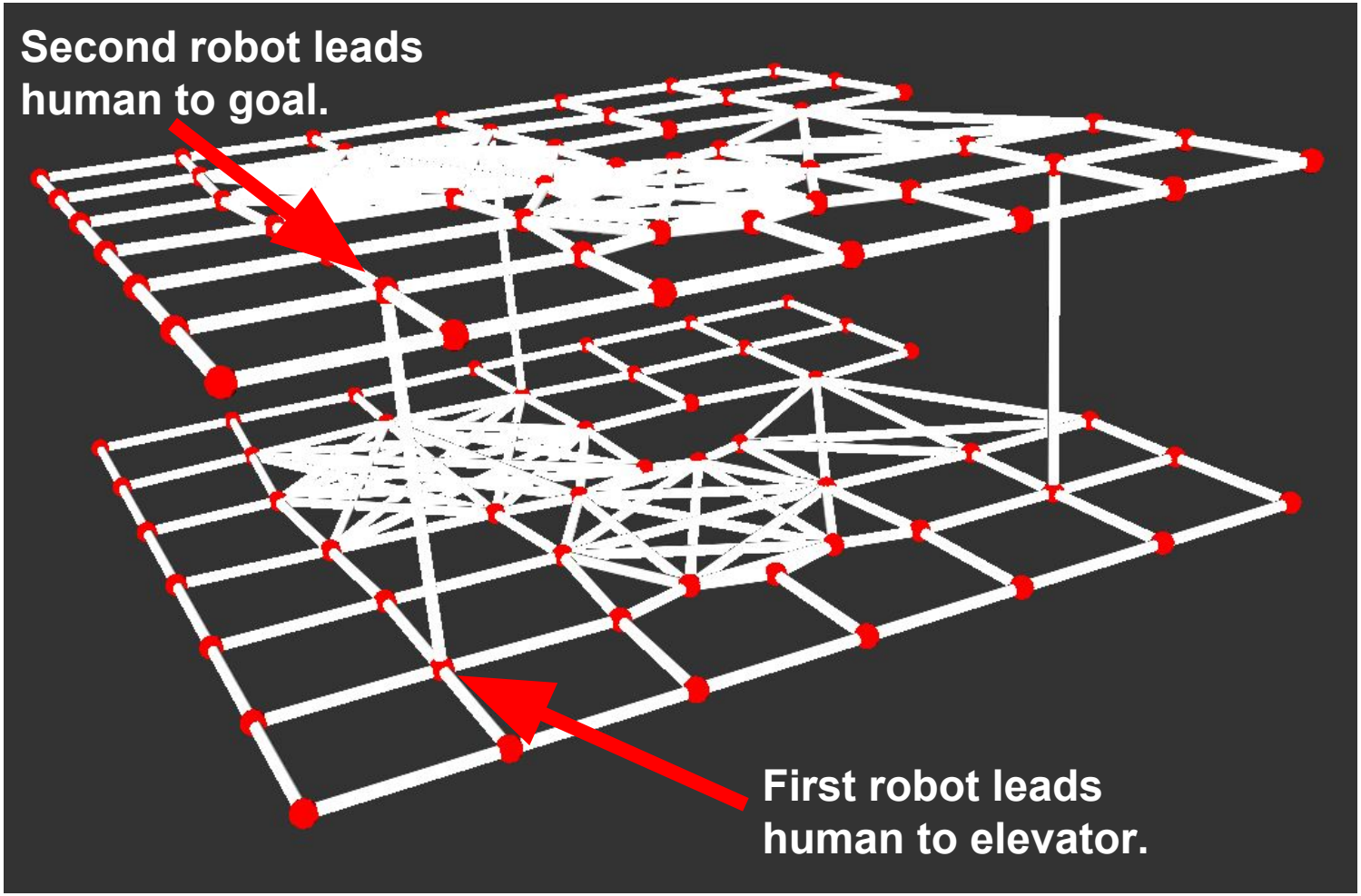


PDP-T [Coltin et al. 2015] baseline



- Transfer only if overall reward improves.
- Significant performance improvement in two floor environments:
  - Allow handoff from one floor to another without either robot traversing elevator.

PDP-T [Coltin et al. 2015] baseline



MCTS-based approaches

- Use MCTS as a general purpose MDP solver.
- To determine how important stochasticity is during planning, we also compare MCTS where a deterministic model of the domain has been used to draw samples.
  - In this model, all action outcomes have been determinized to their most likely outcomes.

MCTS-based approaches

- **When task starts, no time for prior planning….**

MCTS-based approaches

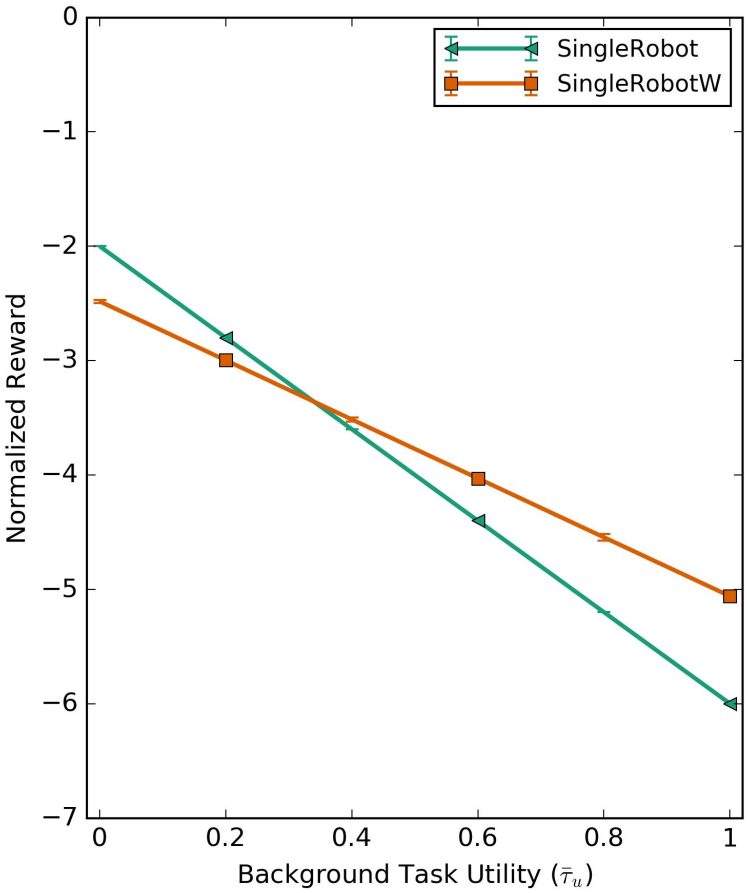
- **When task starts, no time for prior planning….**
  - Ask the human to *wait* so that planning can be done
  - *Lead* human to goal; use MCTS planning subsequently

Evaluation

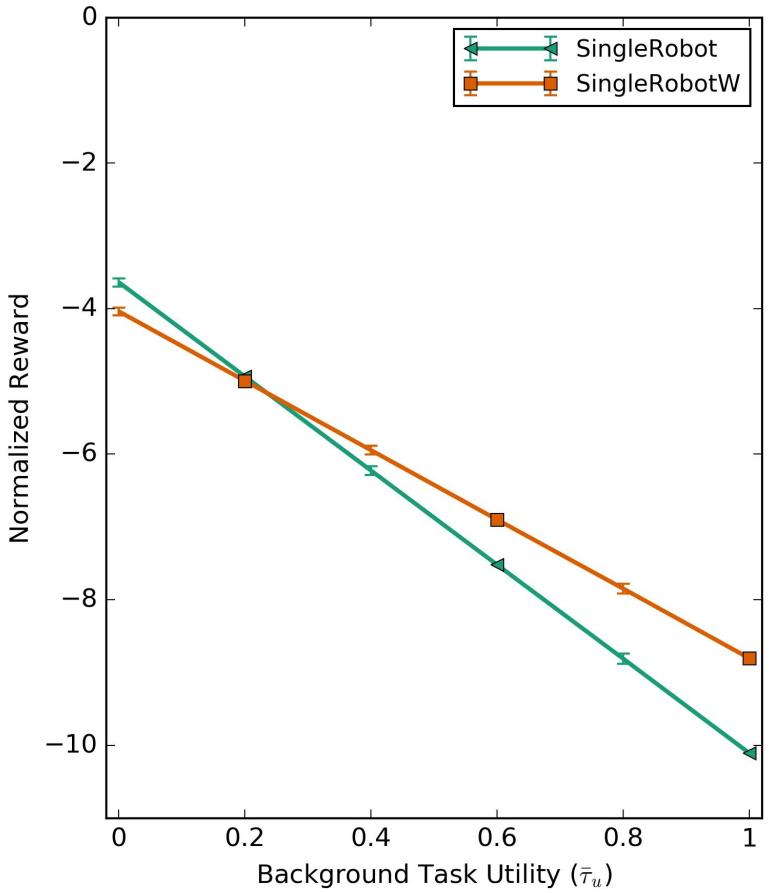
- 3 Heuristic Baselines:
  - SingleRobot
  - SingleRobotW
  - PDP-T
- 4 MCTS variants:
  - MCTS(Wait)
  - MCTS(Lead)
  - MCTS-D(Wait)
  - MCTS-D(Lead)

- We perform evaluation (1000 trials) by generating samples using the hand-coded human decision model.

Evaluation - SingleRobot



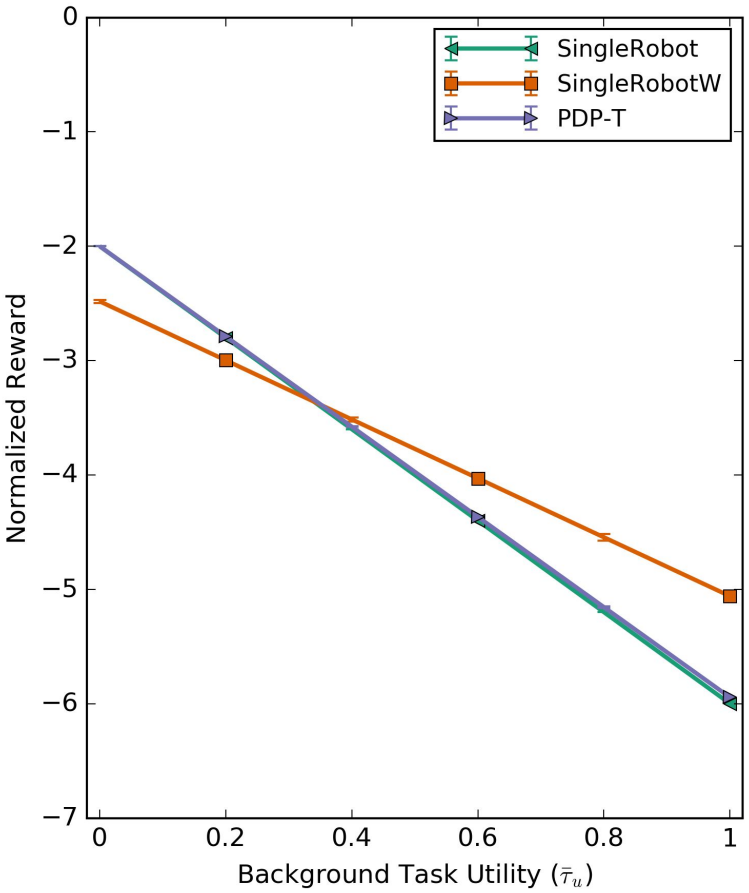
1 floor, 5 robots, 1 request



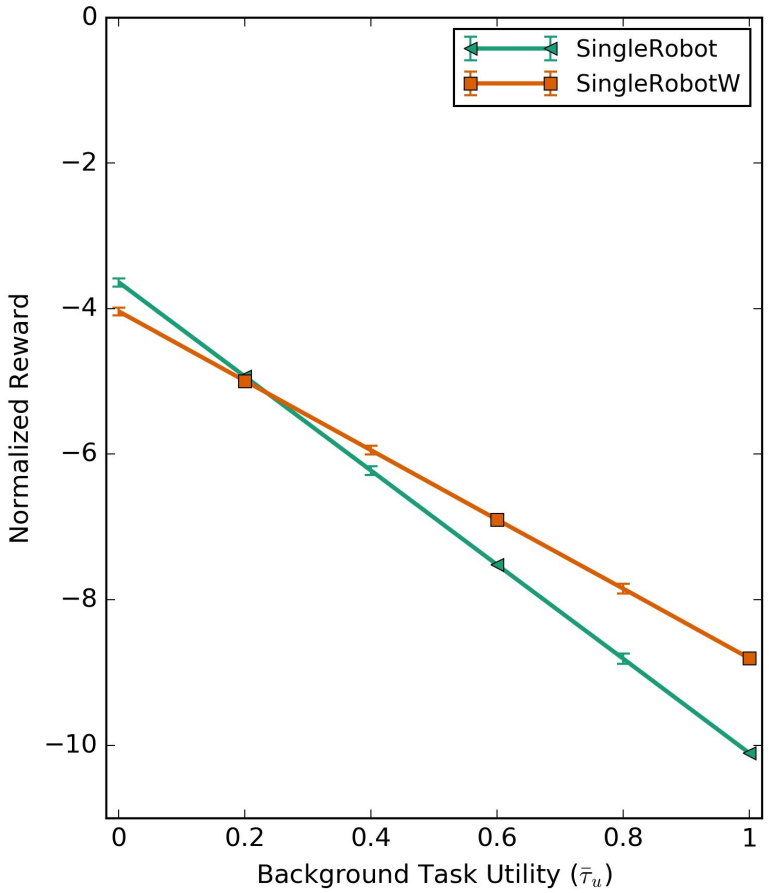
2 floors, 10 robots, 2 requests

Learning Agents Research Group
The University of Texas at Austin

Evaluation - PDP-T



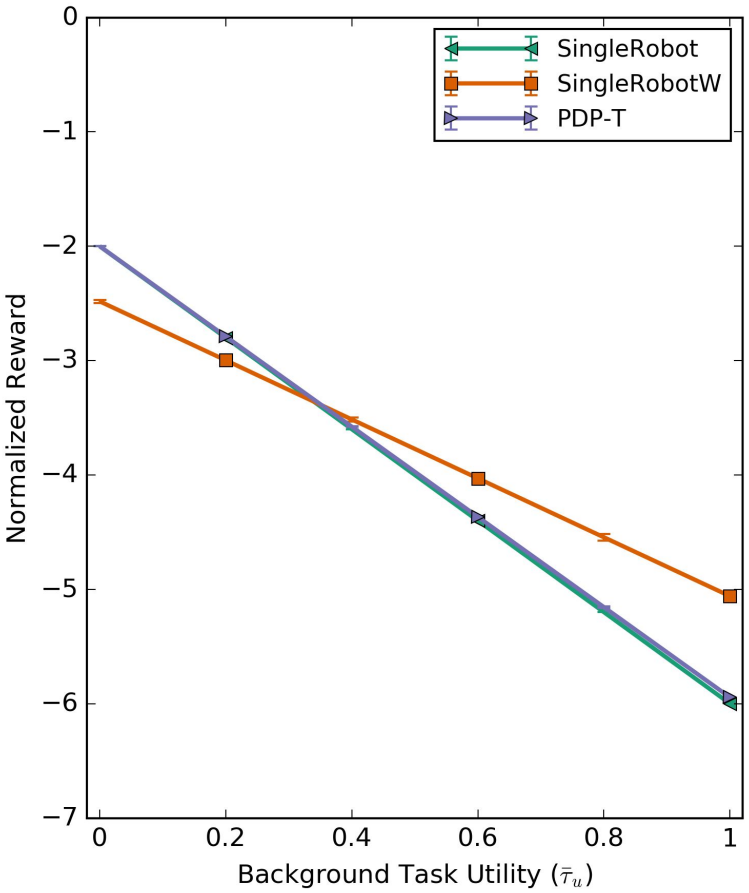
1 floor, 5 robots, 1 request



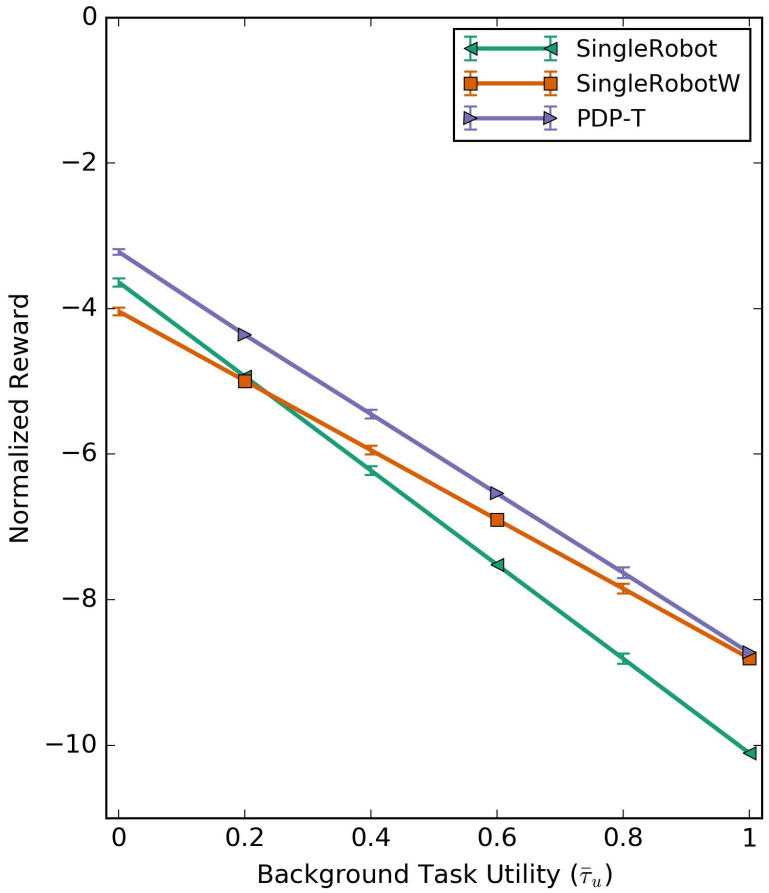
2 floors, 10 robots, 2 requests

Learning Agents Research Group
The University of Texas at Austin

Evaluation - PDP-T



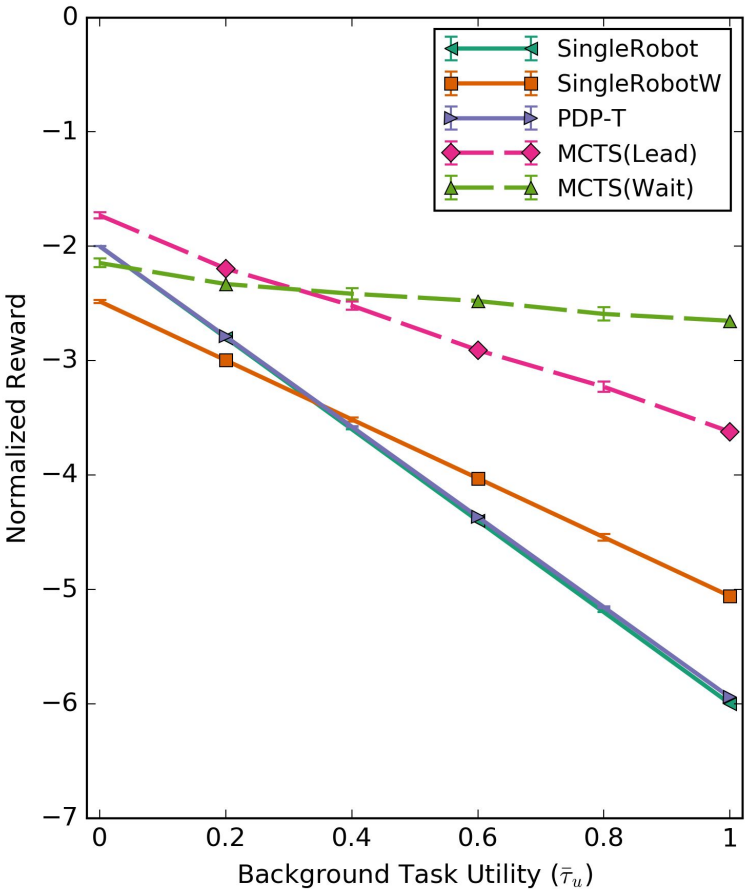
1 floor, 5 robots, 1 request



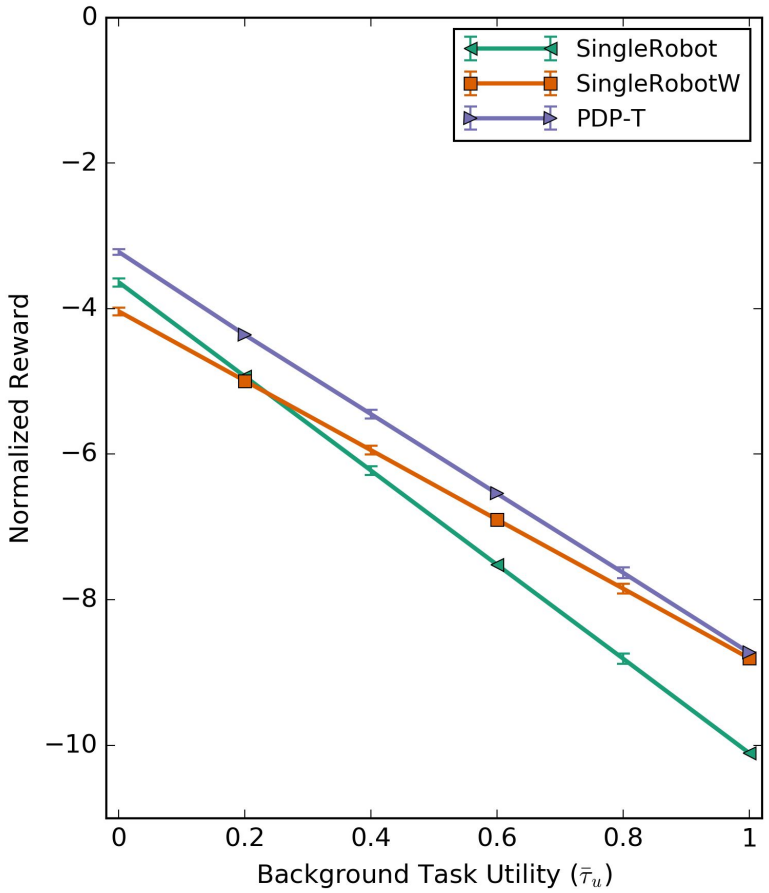
2 floors, 10 robots, 2 requests

Learning Agents Research Group
The University of Texas at Austin

Evaluation - MCTS



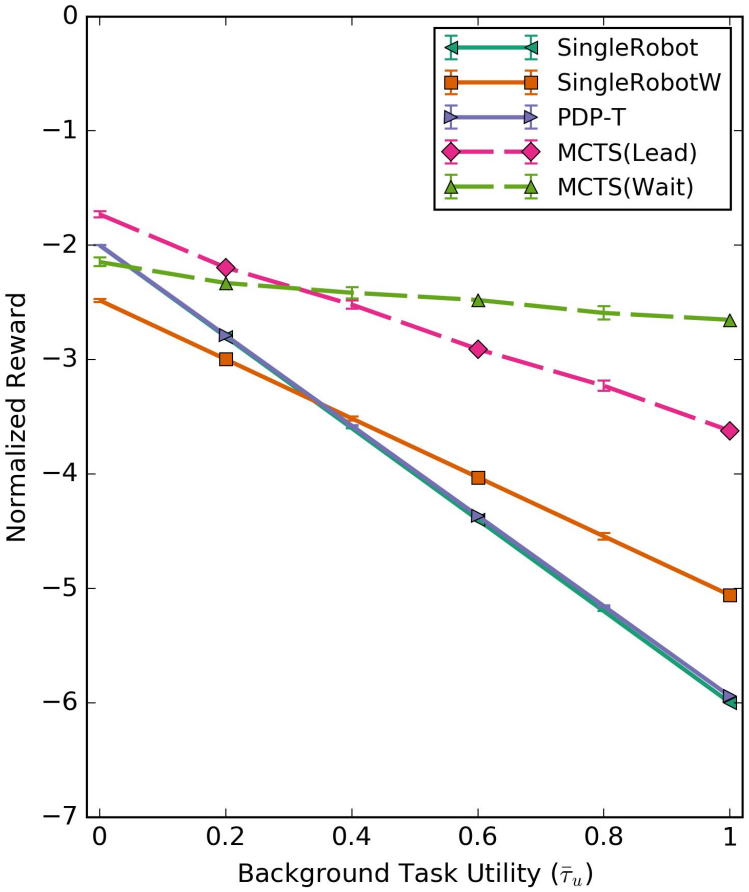
1 floor, 5 robots, 1 request



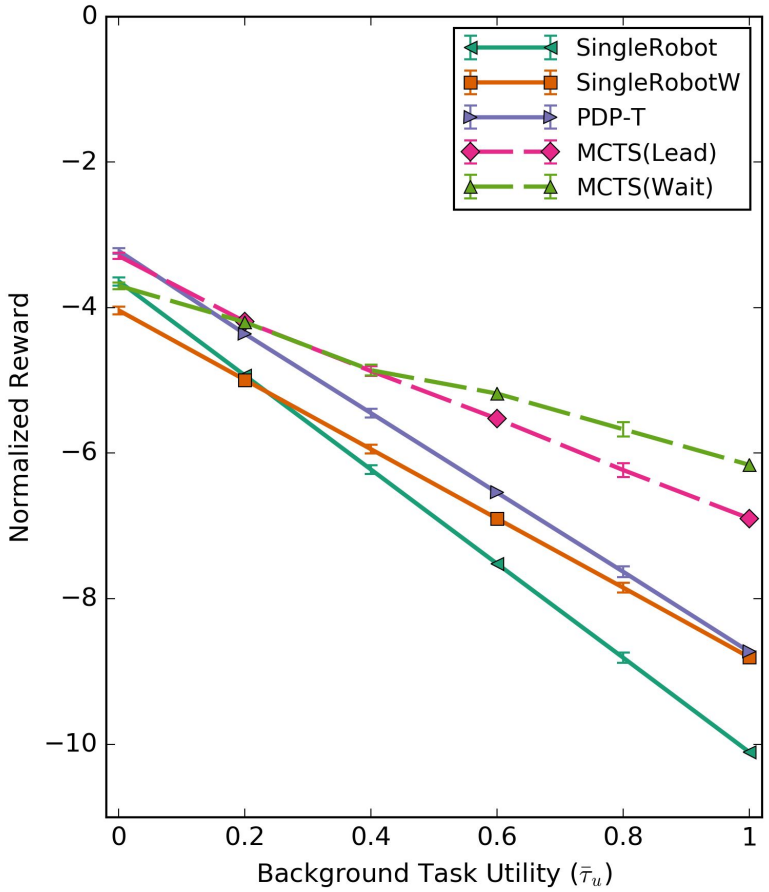
2 floors, 10 robots, 2 requests

Learning Agents Research Group
The University of Texas at Austin

Evaluation - MCTS



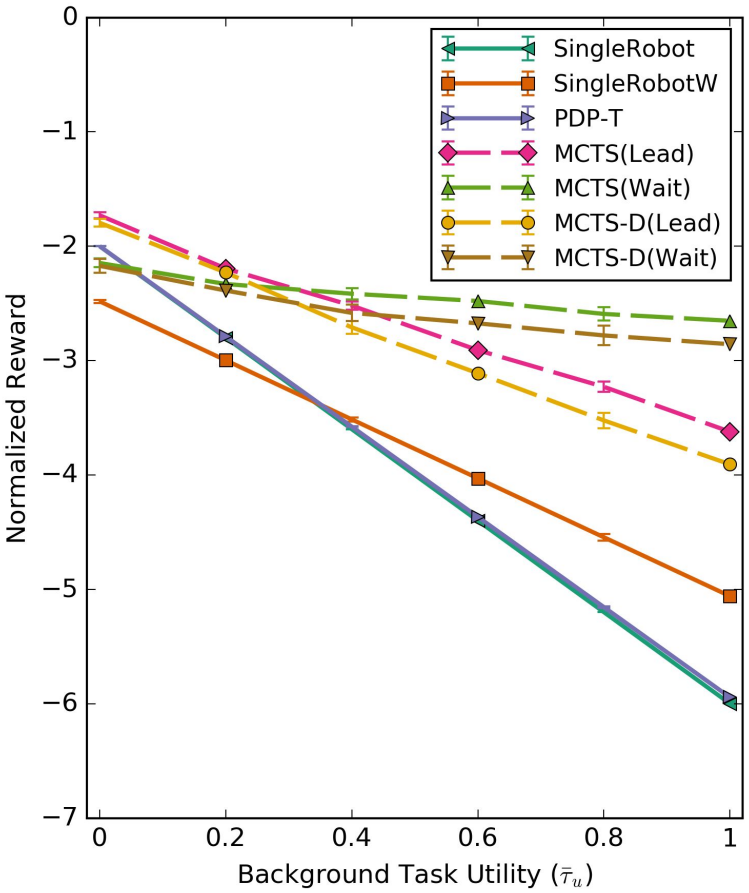
1 floor, 5 robots, 1 request



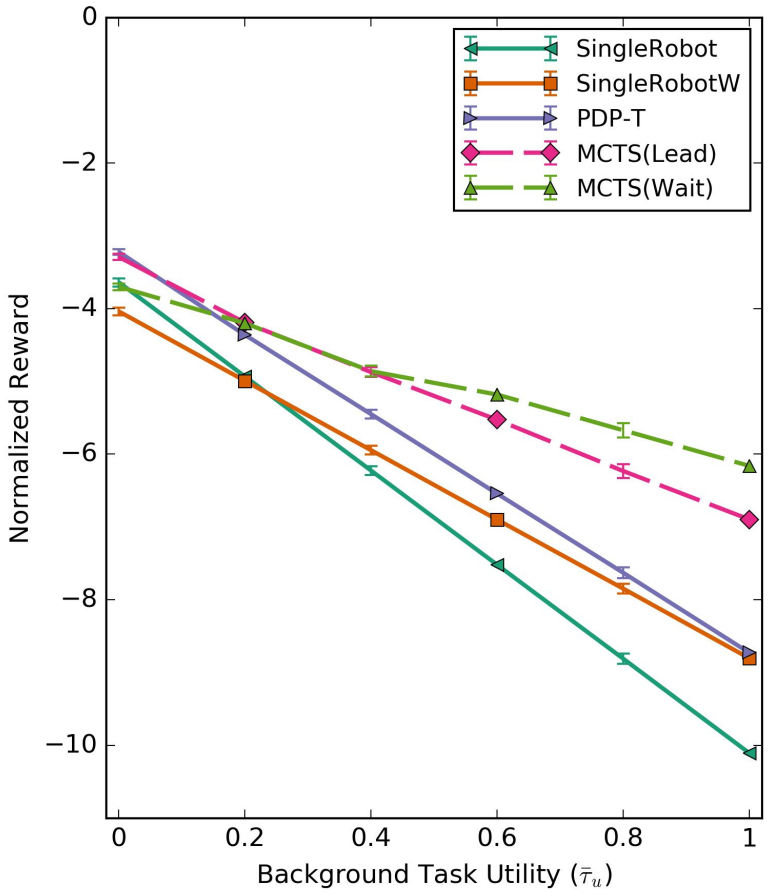
2 floors, 10 robots, 2 requests

Learning Agents Research Group
The University of Texas at Austin

Evaluation - MCTS-D



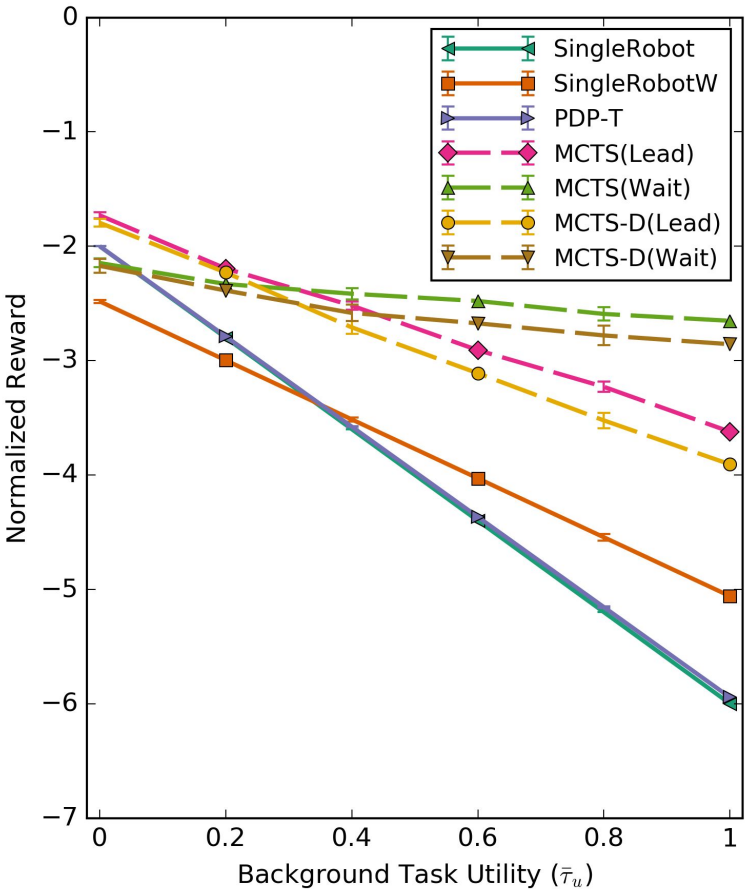
1 floor, 5 robots, 1 request



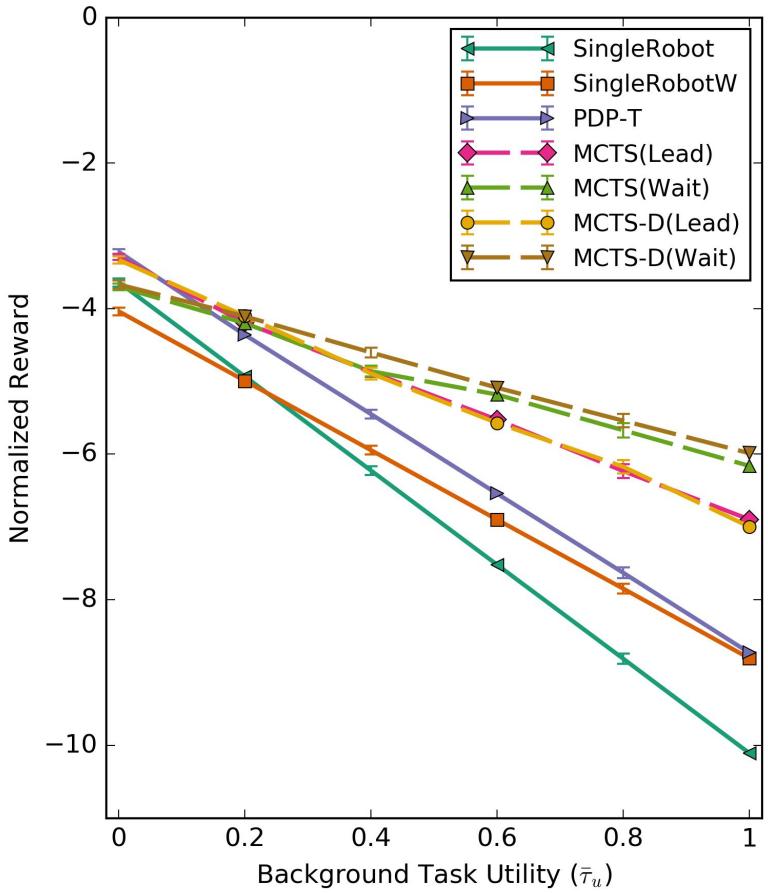
2 floors, 10 robots, 2 requests

Learning Agents Research Group
The University of Texas at Austin

Evaluation - MCTS-D



1 floor, 5 robots, 1 request



2 floors, 10 robots, 2 requests

Learning Agents Research Group
The University of Texas at Austin

Additional Results

- As robot speed increases relative to human speed, the relative performance improvement MCTS has over heuristics decreases.
- MCTS based approaches are robust to some inaccuracies in the model used for planning.

Section Summary

- In this section, we have presented some of the key design decisions behind the MDP formalization of the problem.
- We've also demonstrated that MCTS based search can outperform heuristic baselines in most cases.

- ***However, MCTS planning did not work out of the box!***

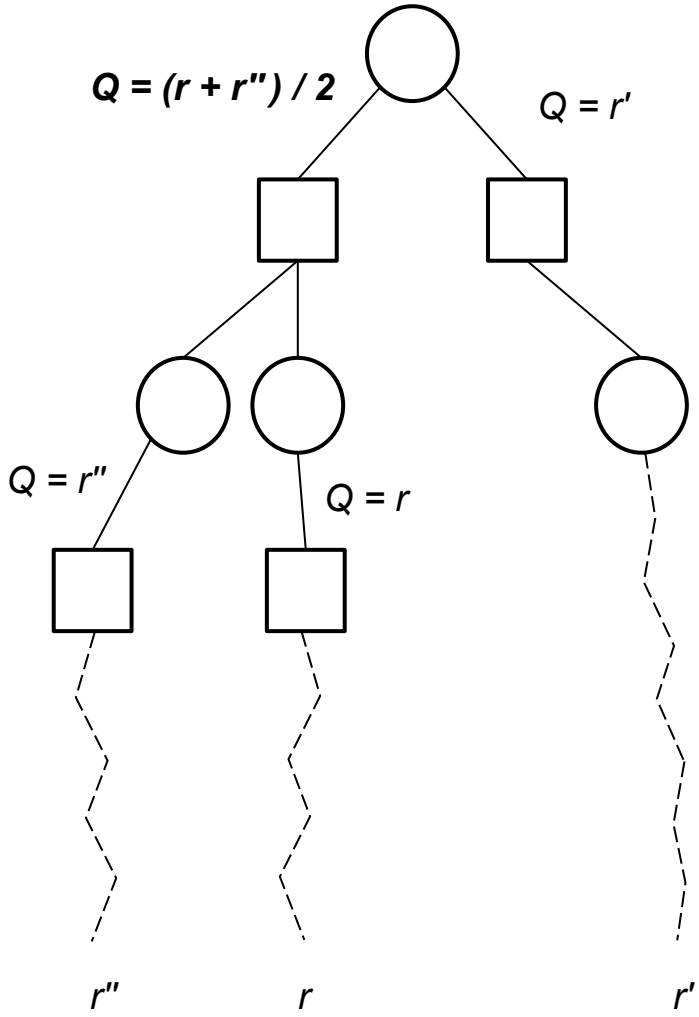
Learning Agents Research Group
The University of Texas at Austin

Overview

- Background
  - Markov Decision Processes
  - Monte Carlo Tree Search
- MDP Formalization of On-Demand Multi-Robot Coordination Problem
- Biased Backup in Monte Carlo Tree Search
- BWIBot Multi-Robot System + Conclusion

Learning Agents Research Group
The University of Texas at Austin

Let's Revisit Backups



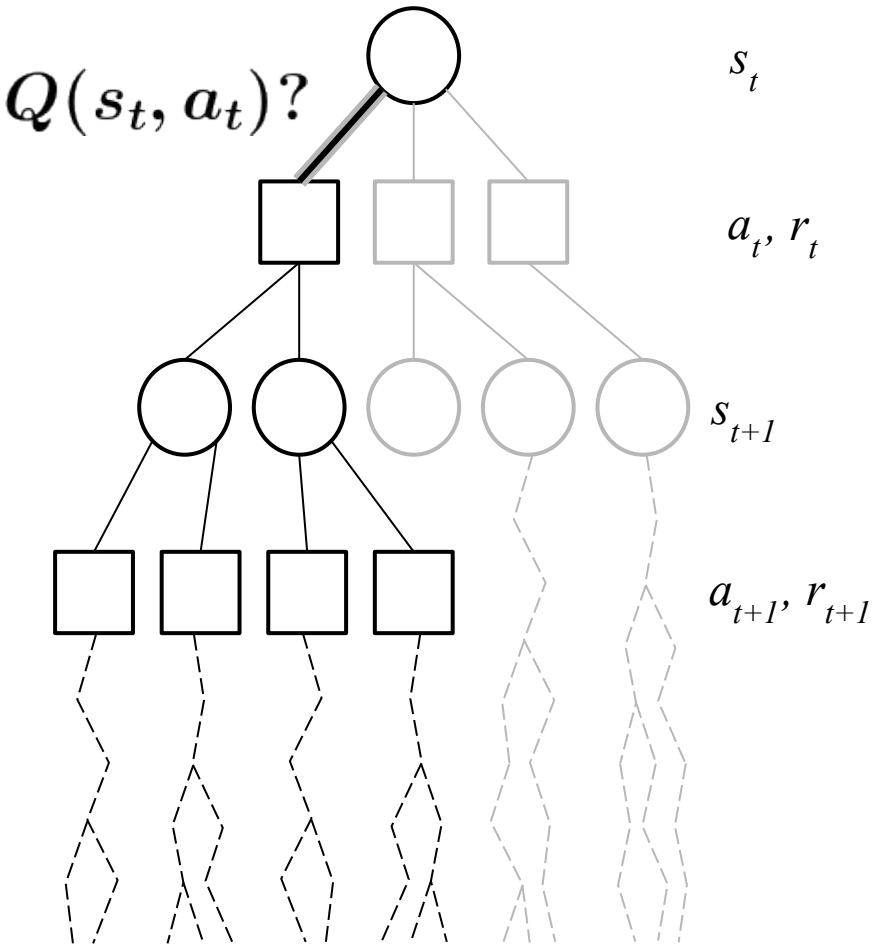
- 4 stages in MCTS:
  - Selection
  - Expansion
  - Simulation
  - ***Backup***

As part of this dissertation we have analyzed different backup techniques. [3]

[3] Khandelwal et al. *On the Analysis of Complex Backup Strategies in Monte Carlo Tree Search*. ICML 2016

Learning Agents Research Group
The University of Texas at Austin

MCTS - Backup (Motivation)



Monte Carlo backup for single trajectory:

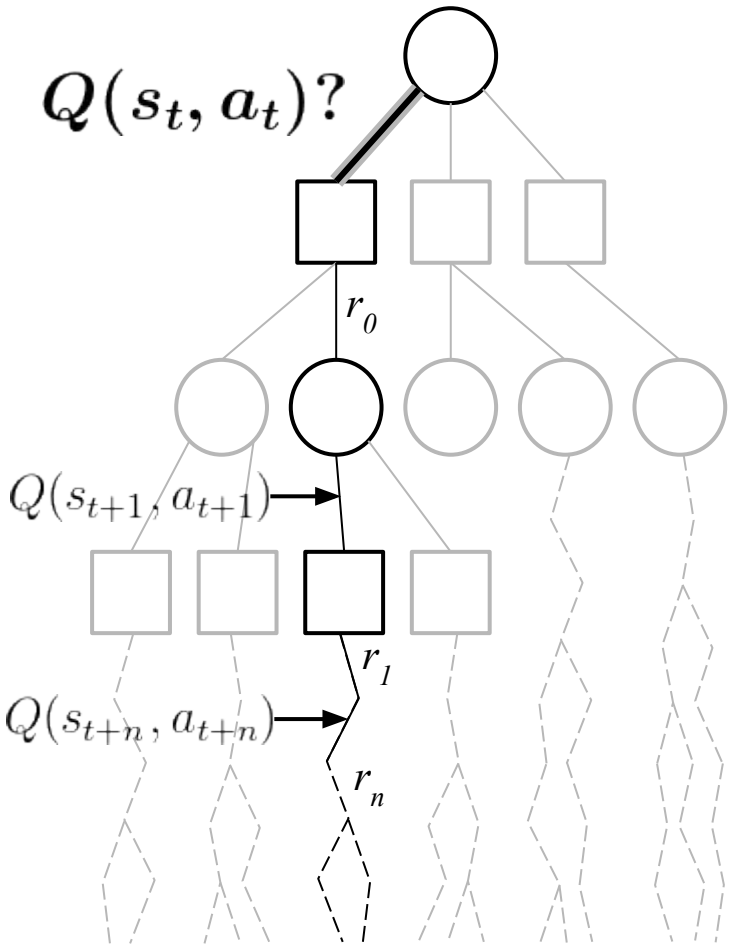
$$R = \sum_{i=0}^{L-1} \gamma^i r_{t+i}$$

Across all trajectories:

$$Q(s_t, a_t) = \mathbb{E}\left[\sum_{i=0}^{L-1} \gamma^i r_{t+i}\right]$$

Can we do better?

n-step return (bias-variance tradeoff)



We can compute the return sample in many different ways!

1-step:

$$R^{(1)} = r_t + \gamma Q(s_{t+1}, a_{t+1}),$$

n-step:

$$R^{(n)} = \left[\sum_{i=0}^{n-1} \gamma^i r_{t+i}\right] + \gamma^n Q(s_{t+n}, a_{t+n})$$

Monte Carlo:

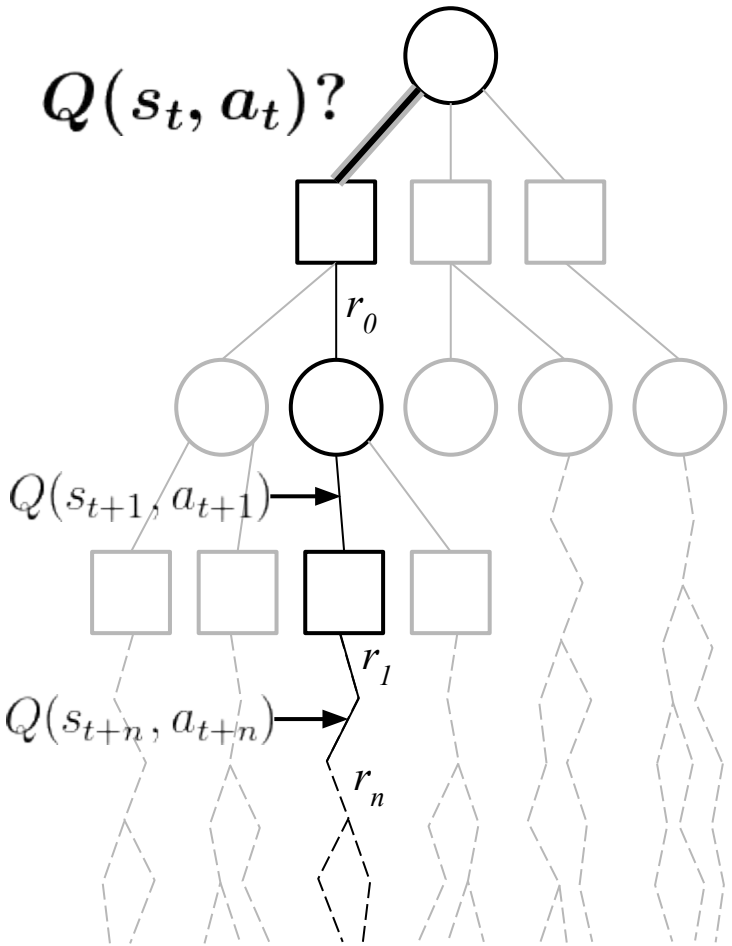
$$R = \sum_{i=0}^{L-1} \gamma^i r_{t+i}$$

More Bias

More Variance

Learning Agents Research Group
The University of Texas at Austin

Complex returns



Complex return: $$R^C = \sum_{i=1}^{L} \left[w_{n,L} \cdot R^{(n)} \right]$$

λ-return/eligibility [Rummery 1995]:

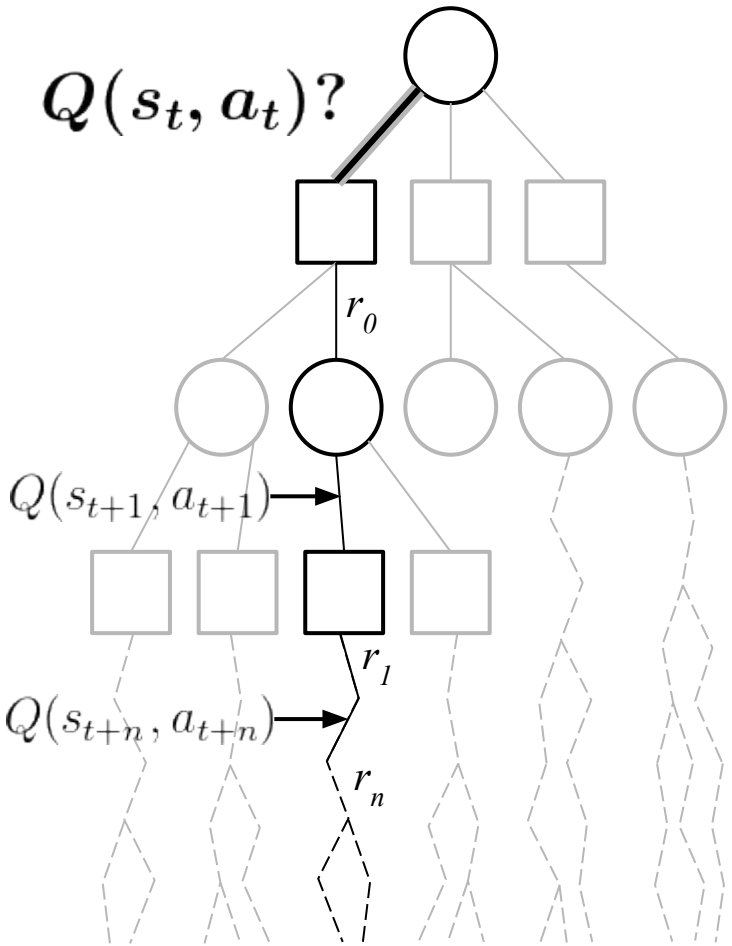
$$w^{\lambda}_{n,L} = \begin{cases} (1-\lambda)\lambda^{n-1} & 1 \leq n < L \\ \lambda^{L} & n = L \end{cases}$$

γ-return weights [Konidaris et al. 2011]:

$$w^{\gamma}_{n,L} = \frac{\left(\sum_{i=1}^{n} \gamma^{2(i-1)}\right)^{-1}}{\sum_{n=1}^{L}\left(\sum_{i=1}^{n} \gamma^{2(i-1)}\right)^{-1}}$$

LARG Learning Agents Research Group The University of Texas at Austin

MCTS - Novel Variants with Biased Backups!



Complex return: $$R^C = \sum_{i=1}^{L} \left[w_{n,L} \cdot R^{(n)} \right]$$

λ-return/eligibility [Rummery 1995]:

➡ **MCTS(λ)**

$$w^{\lambda}_{n,L} = \begin{cases} (1-\lambda)\lambda^{n-1} & 1 \leq n < L \\ \lambda^{L} & n = L \end{cases}$$

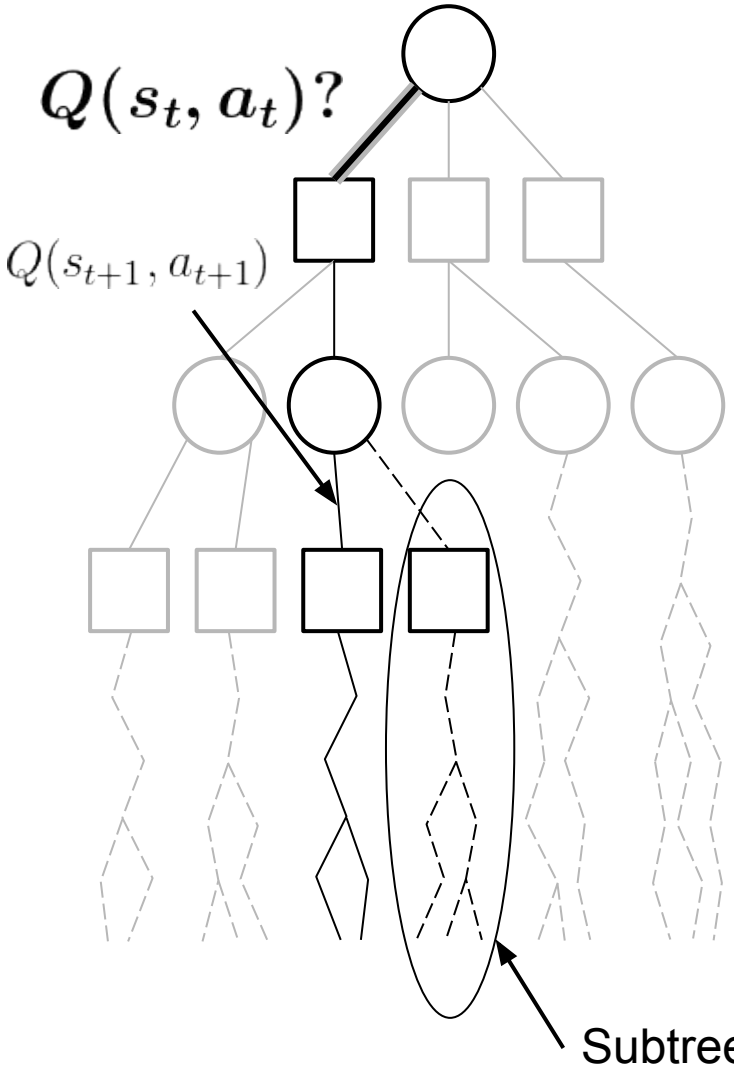
γ-return weights [Konidaris et al. 2011]:

➡ **MCTSγ**

$$w^{\gamma}_{n,L} = \frac{\left(\sum_{i=1}^{n} \gamma^{2(i-1)}\right)^{-1}}{\sum_{n=1}^{L}\left(\sum_{i=1}^{n} \gamma^{2(i-1)}\right)^{-1}}$$

LARG Learning Agents Research Group The University of Texas at Austin

MaxMCTS - Off-Policy Returns



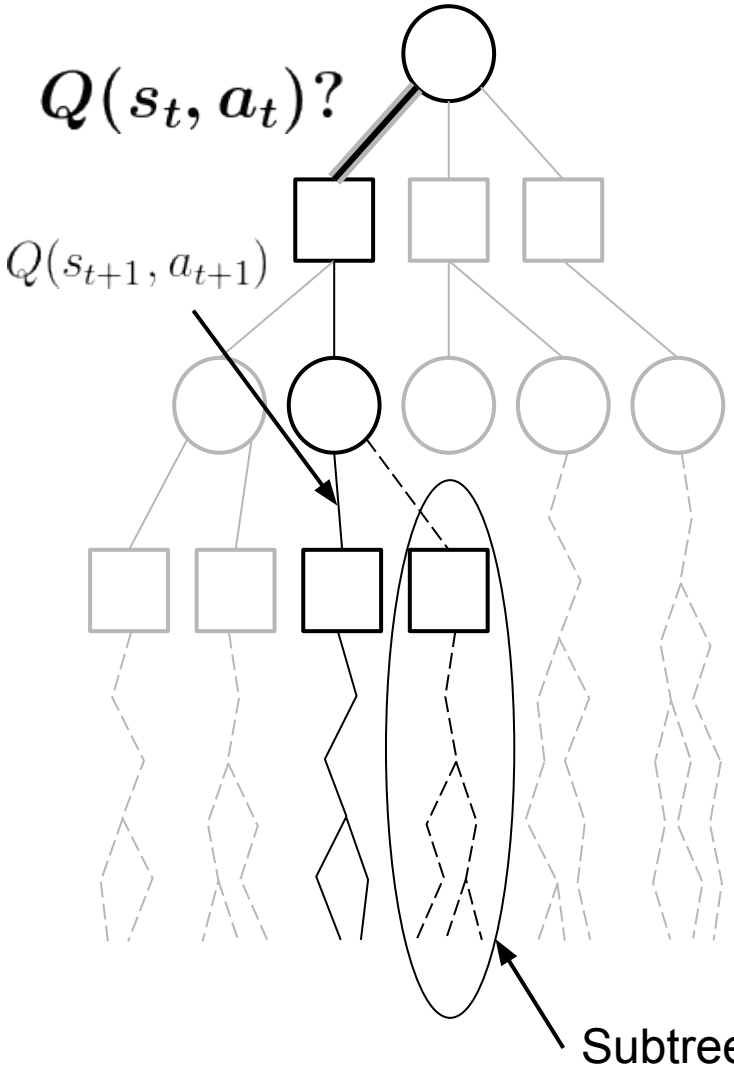
Backup using best known action:

$$R^{(1)} = r_t + \gamma \max_a Q(s_{t+1}, a)$$

$$R^{(n)} = \sum_{i=0}^{n-1} \gamma^i r_{t+i} + \gamma^n \max_a Q(s_{t+n}, a)$$

Learning Agents Research Group
The University of Texas at Austin

MaxMCTS - Off-Policy Returns



Backup using best known action:

$$R^{(1)} = r_t + \gamma \max_a Q(s_{t+1}, a)$$

$$R^{(n)} = \sum_{i=0}^{n-1} \gamma^i r_{t+i} + \gamma^n \max_a Q(s_{t+n}, a)$$

Intuition:

- Don't penalize exploratory actions.
- Reinforce previously seen better trajectories instead.

Equivalent to Peng's Q(λ) style updates.

MaxMCTS(λ) and **MaxMCTSγ**

LARG Learning Agents Research Group The University of Texas at Austin

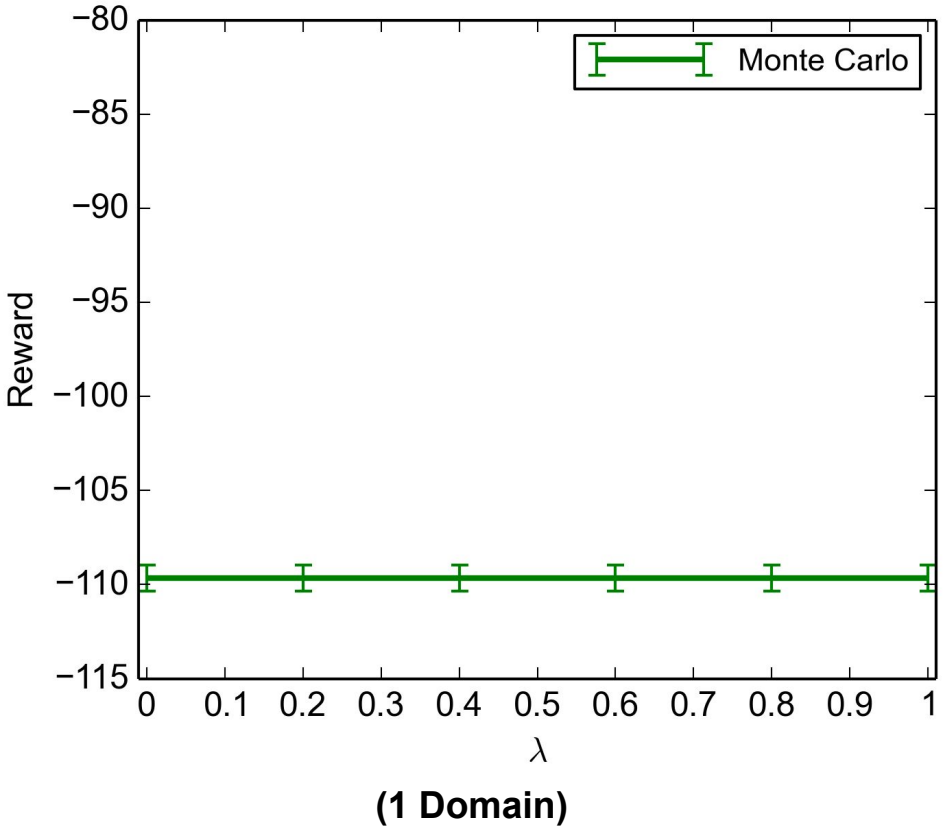
Evaluation

- We have proposed 4 novel variants:
  - On-policy: MCTS(λ) and MCTS$_\gamma$
  - Off-policy: MaxMCTS(λ) and MaxMCTS$_\gamma$

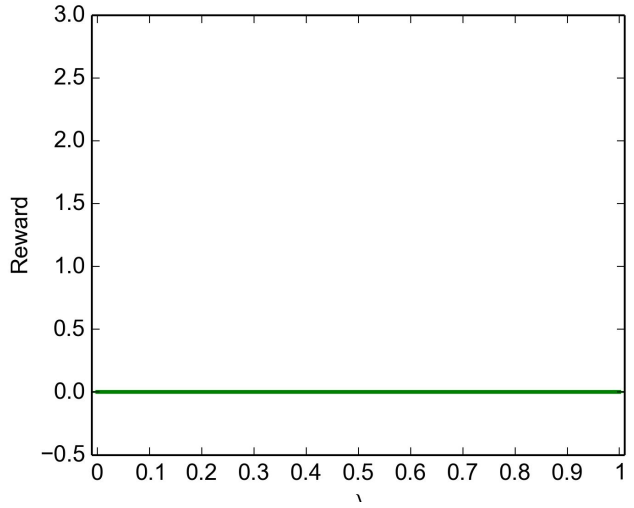
- We only show the performance of MaxMCTS variants with random action selection during planning.

- Test performance in 12 different IPC domains
  - Limited planning time (10,000 rollouts per step).

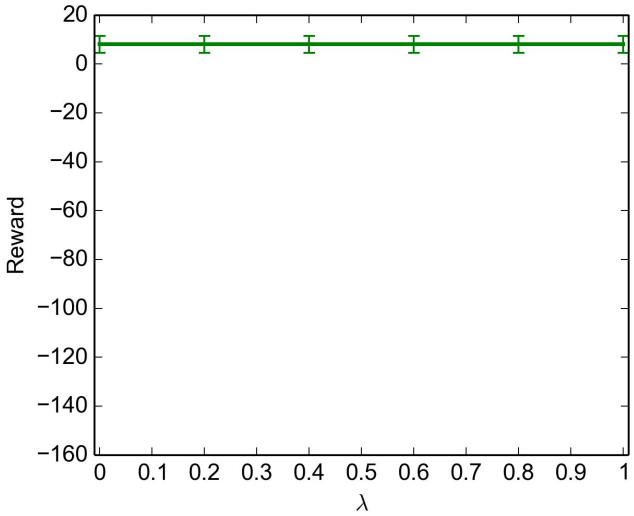
IPC - Random action selection



(1 Domain)



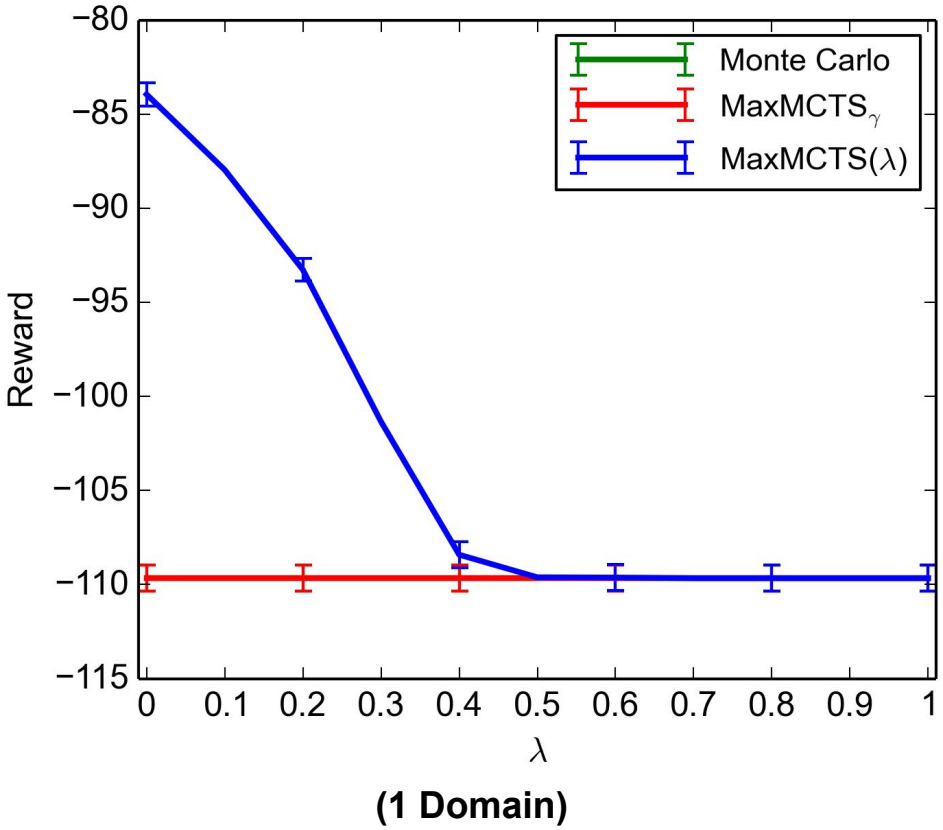
(3 Domains)



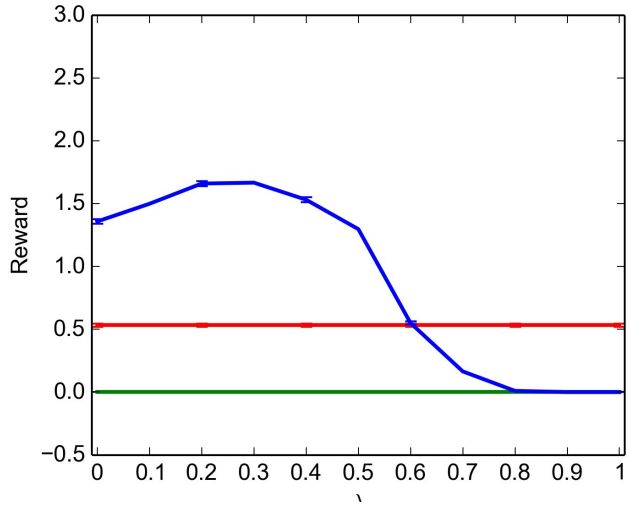
(7 Domains)

Learning Agents Research Group
The University of Texas at Austin

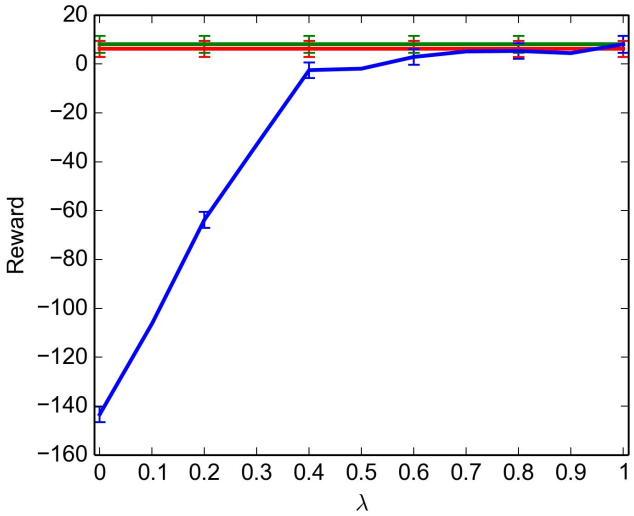
IPC - Random action selection



(1 Domain)



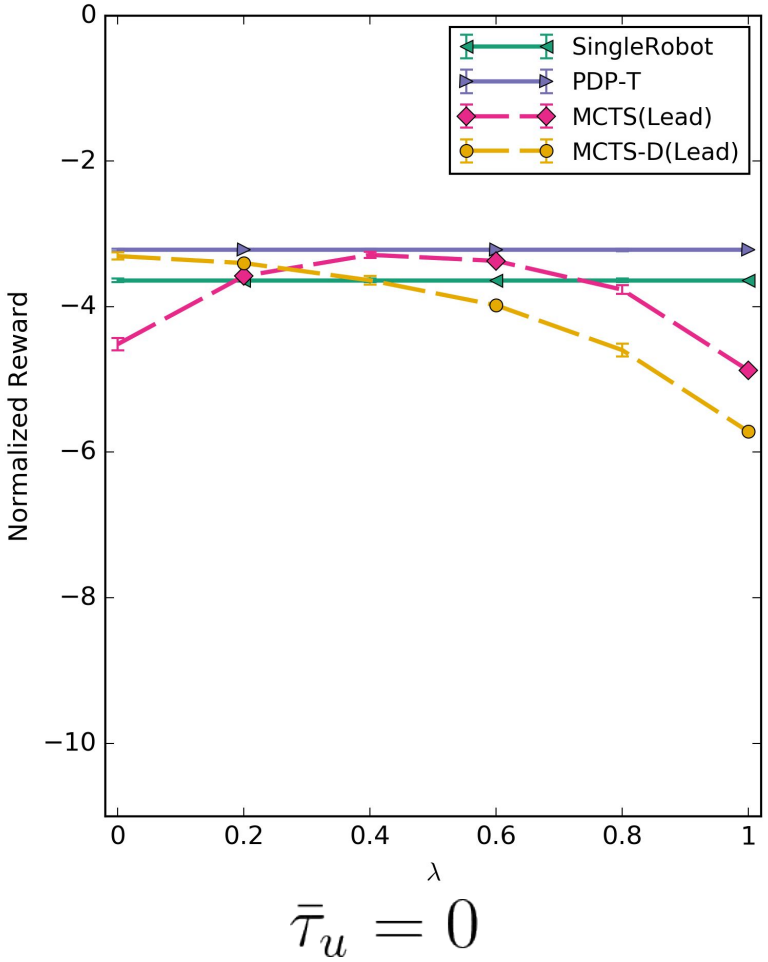
(3 Domains)



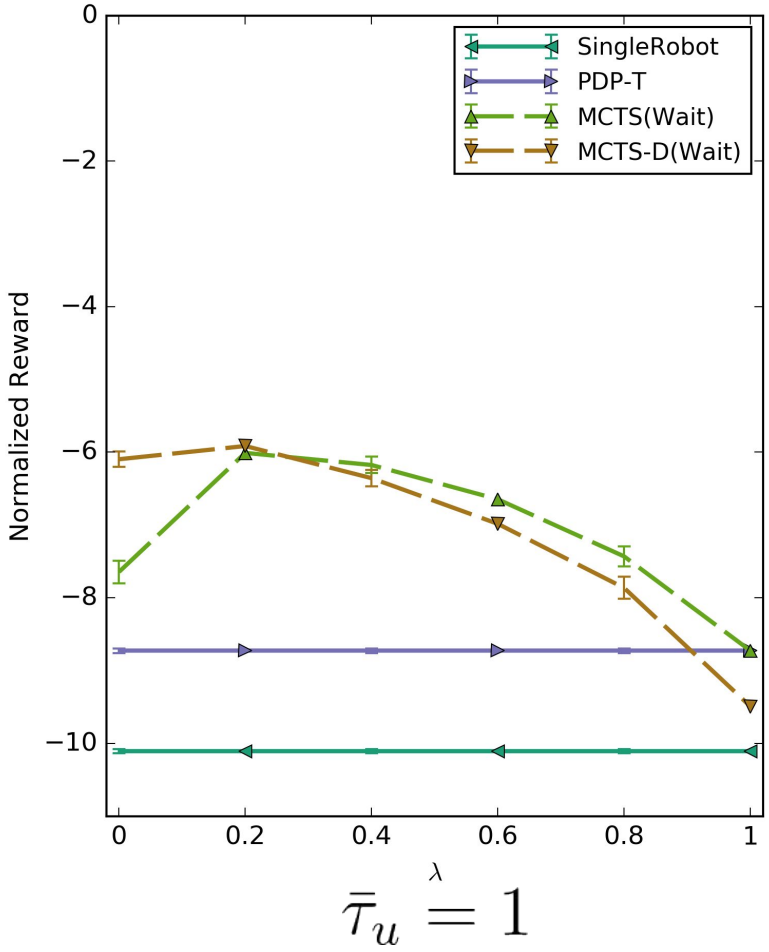
(7 Domains)

Learning Agents Research Group
The University of Texas at Austin

We used MaxMCTS(λ) in previous results!



$\bar{\tau}_u = 0$



$\bar{\tau}_u = 1$

2 floors, 10 robots, 2 requests

LARG Learning Agents Research Group The University of Texas at Austin

Additional Results

- Different backup techniques can be implemented efficiently, and typically do not change overall planning time by 10%.

- We also demonstrated dependence between domain structure and and bias by using a parametrized grid-world domain.

Section Summary

- We have introduced and analyzed some principled and parametrized approaches for inducing bias during backup in MCTS.

- In some domains, selecting the right complex backup strategy is important. *On-demand Multi-Robot Coordination is one such domain!*

- Applicable to many different domains:
  - Music Recommendation System [Liebman et al. 2017].

Overview

- Background
  - Markov Decision Processes
  - Monte Carlo Tree Search
- Biased Backup in Monte Carlo Tree Search
- MDP Formalization of On-Demand Multi-Robot Coordination Problem
- BWIBot Multi-Robot System + Conclusion

BWIBot Multi-Robot System

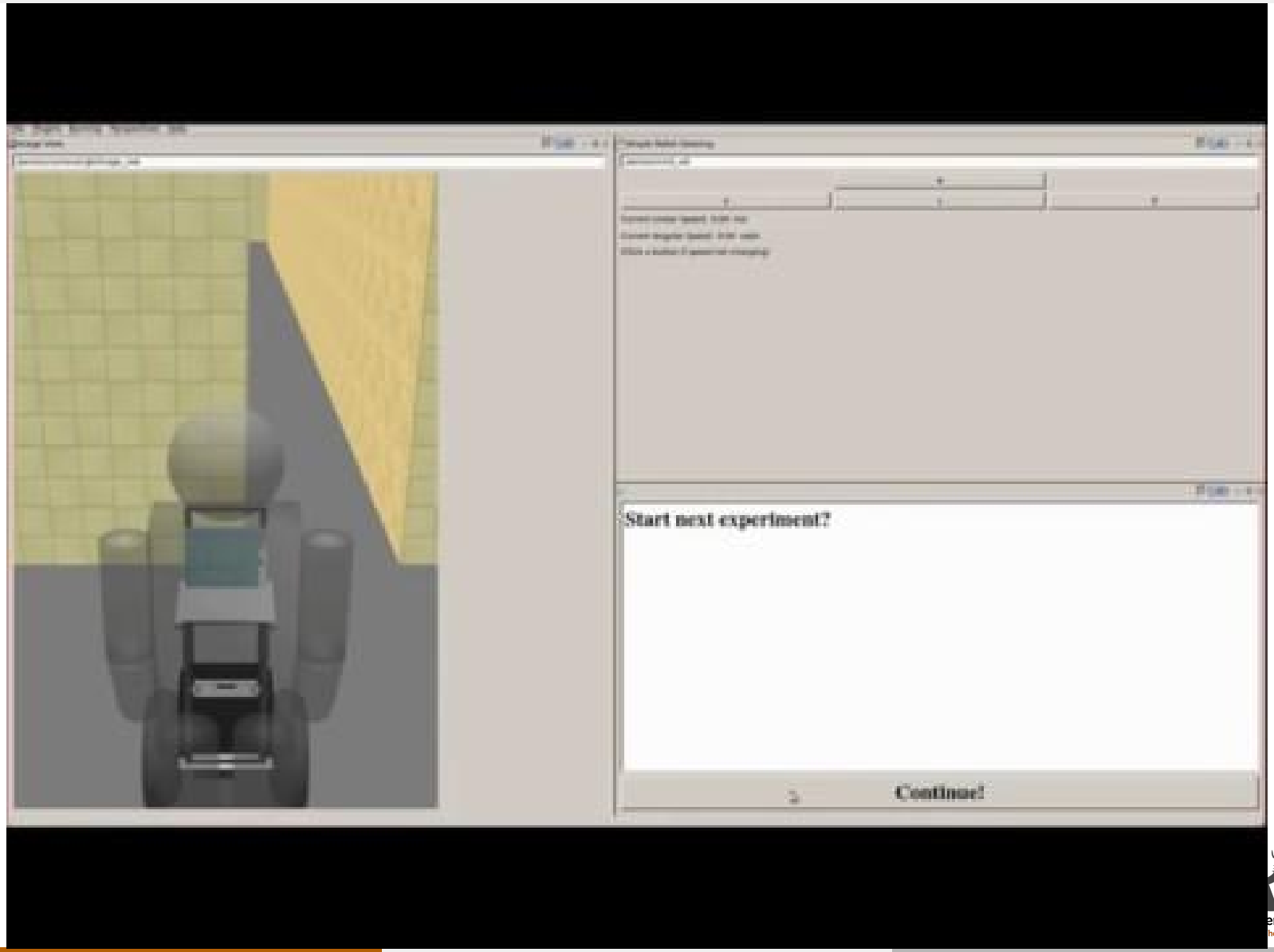
- 3rd iteration of BWIBot platform [4]
- Over 600km of recorded distance traveled.
- Realistic 3D simulation using ROS and Gazebo.

- Implement MDP framework on real robots and simulated user study.

[4] Khandelwal et al. *BWIBots: A platform for bridging the gap between AI and human–robot interaction research*. IJRR 2017

LARG
Learning Agents Research Group
The University of Texas at Austin

User Study Interface



Dissertation - Key Contributions

- Designed MDP formalization of on-demand multi-robot coordination problem.

Dissertation - Key Contributions

- Designed MDP formalization of on-demand multi-robot coordination problem.
- Key representational decisions applicable to other formalizations of multi-robot coordination problems as well.
  - Topological representation, action decomposition, and variable duration of actions.

Dissertation - Key Contributions

- Designed MDP formalization of on-demand multi-robot coordination problem.
- Key representational decisions applicable to other formalizations of multi-robot coordination problems as well.
  - Topological representation, action decomposition, and variable duration of actions.
- Introduced and analyzed some principled and parameterized approaches for inducing bias during backup in MCTS.

Dissertation - Key Contributions

- Designed MDP formalization of on-demand multi-robot coordination problem.
- Key representational decisions applicable to other formalizations of multi-robot coordination problems as well.
  - Topological representation, action decomposition, and variable duration of actions.
- Introduced and analyzed some principled and parameterized approaches for inducing bias during backup in MCTS.
- Developed the BWIBot multi-robot system and proof-of-concept implementation of the MDP framework.

LARG
Learning Agents Research Group
The University of Texas at Austin

Directions for Future Work

- Human decision model learning; quickly selecting an appropriate model.
- Generalization of value estimates in MCTS.
- Alternate planning approaches (RTDP, Deterministic planning with limited stochasticity).
- Alternate on-demand tasks.

LARG
Learning Agents Research Group
The University of Texas at Austin

Related Work (MCTS Backup)

- λ-return has been applied previously for planning:
  - TEXPLORE used a slightly different version of MaxMCTS(λ) [Hester 2012].

- Other backup strategies:
  - MaxMCTS(λ=0) is equivalent to MaxUCT [Keller, Helmert 2012].
  - Coulom analyzed hand-designed backup strategies in 9x9 Computer Go [Coulom 2007].

Related Work (MDP Framework)

- Multi-Robot Task Allocation Strategies [Gerkey et al. 2004].
- Flexible Job-shop Scheduling Problem [Brucker 1990].
- Dec-POMDP and Macro-Actions for multi-robot coordination [Amato et al. 2015].

- Operator Decomposition [Standley 2010].

- MINERVA Tour Guide [Thrun et al. 1998].
- Single Robot Human Guidance [Montemerlo et al. 2002].

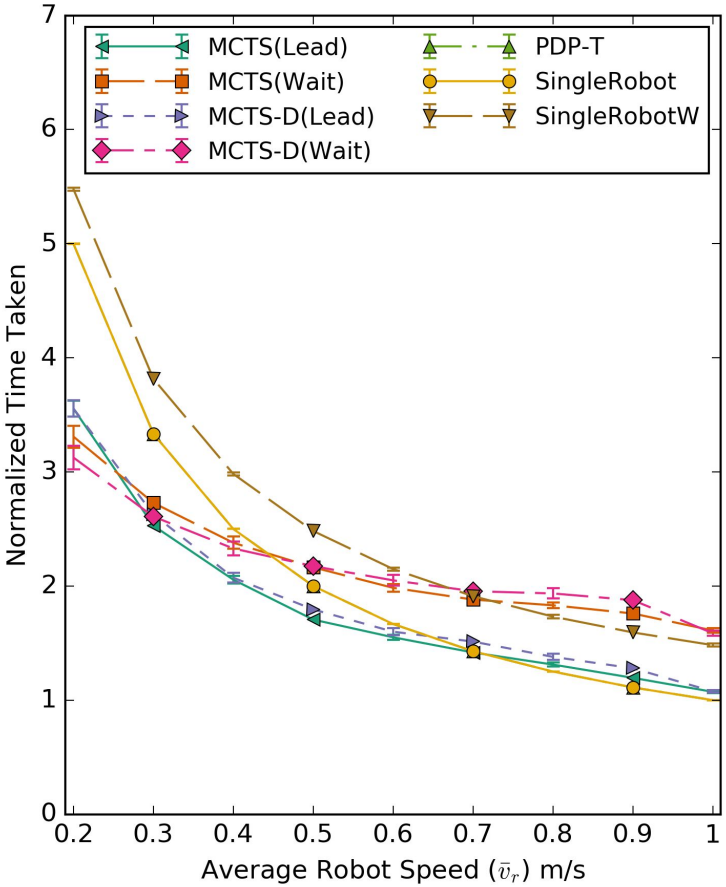
LARG
Learning Agents Research Group
The University of Texas at Austin

Thanks!

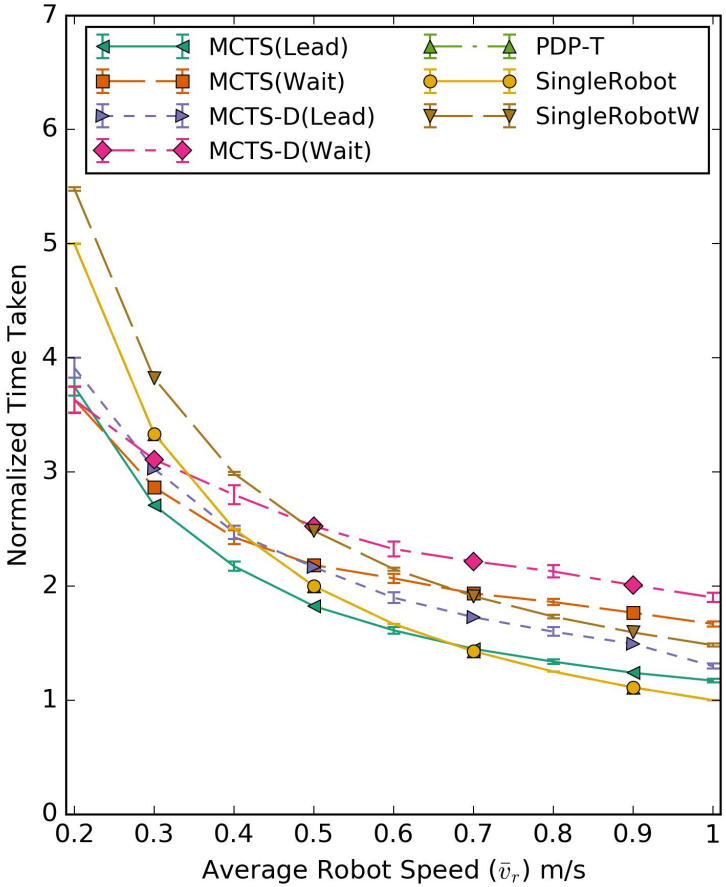
MDP - Variable duration of actions

- MDP decisions are made when human completes a transition to a graph node
- Action decomposition induces another source of variable duration of actions
  - Decomposed actions do not take time in the real world
- Planning approaches such as MCTS need to plan for a sequence of individual actions
- Due to variable duration, the formulation is Semi-Markov

MDP - Increasing Robot Speed



$\bar{\tau}_u = 0$

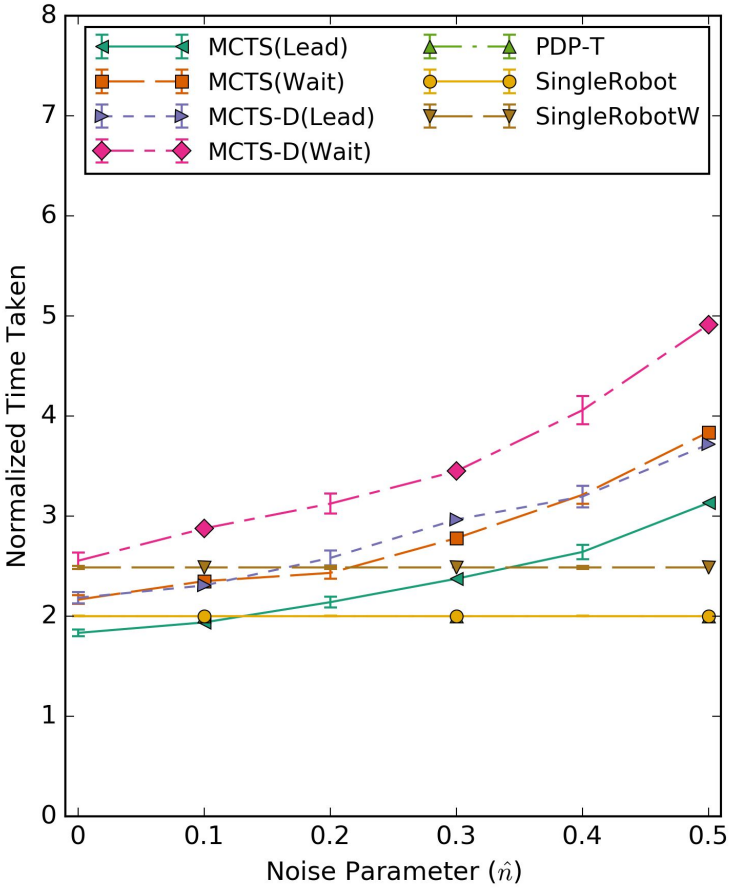


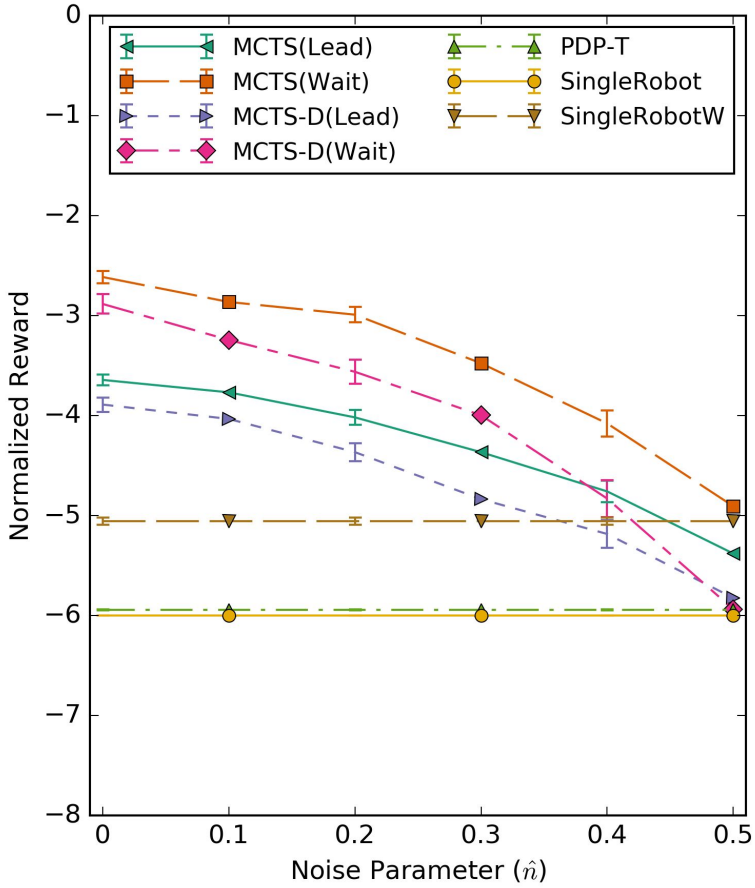
$\bar{\tau}_u = 1$

1 floor, 5 robots

LARG
Learning Agents Research Group
The University of Texas at Austin

MDP - Model Inaccuracy during Planning

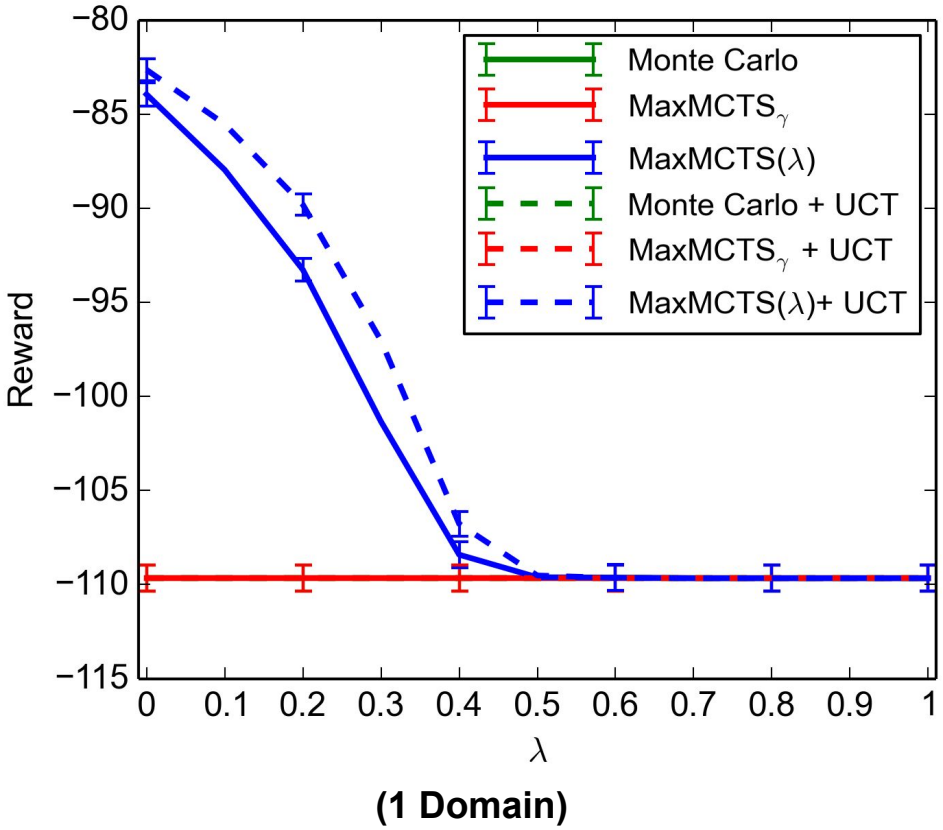




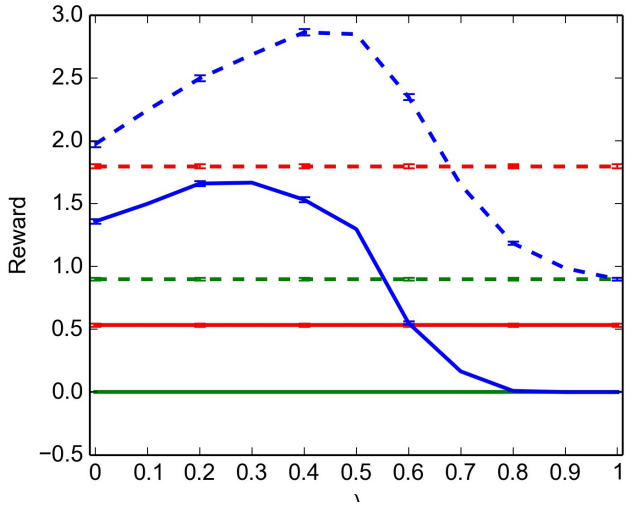
1 floor, 5 robots, $\bar{\tau}_u = 1$

Learning Agents Research Group
The University of Texas at Austin

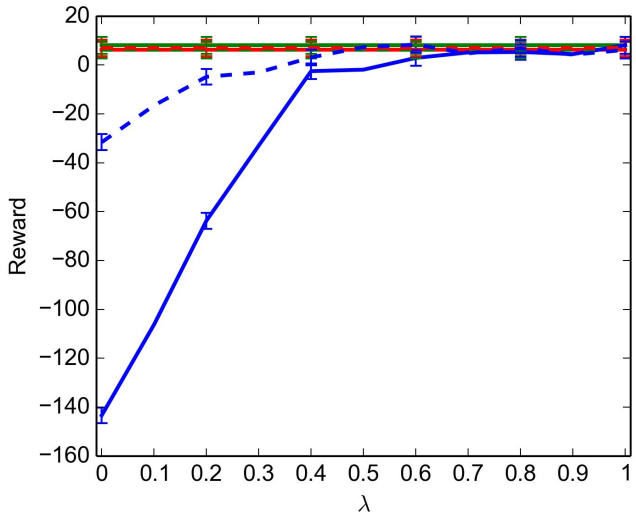
IPC - UCB1 action selection



(1 Domain)



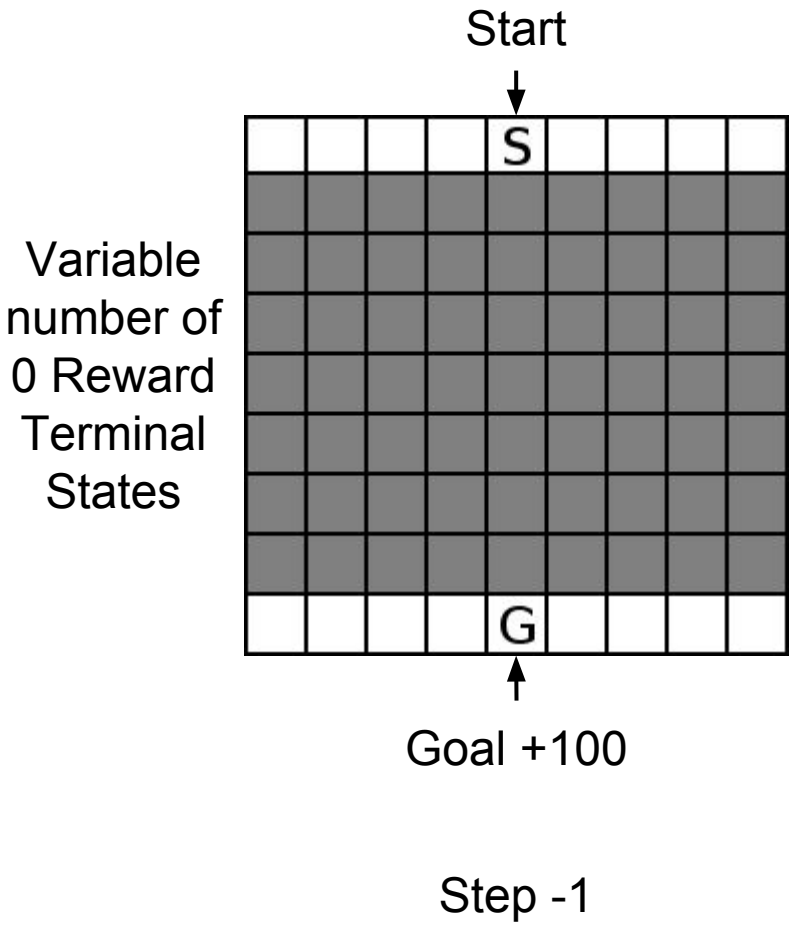
(3 Domains)



(7 Domains)

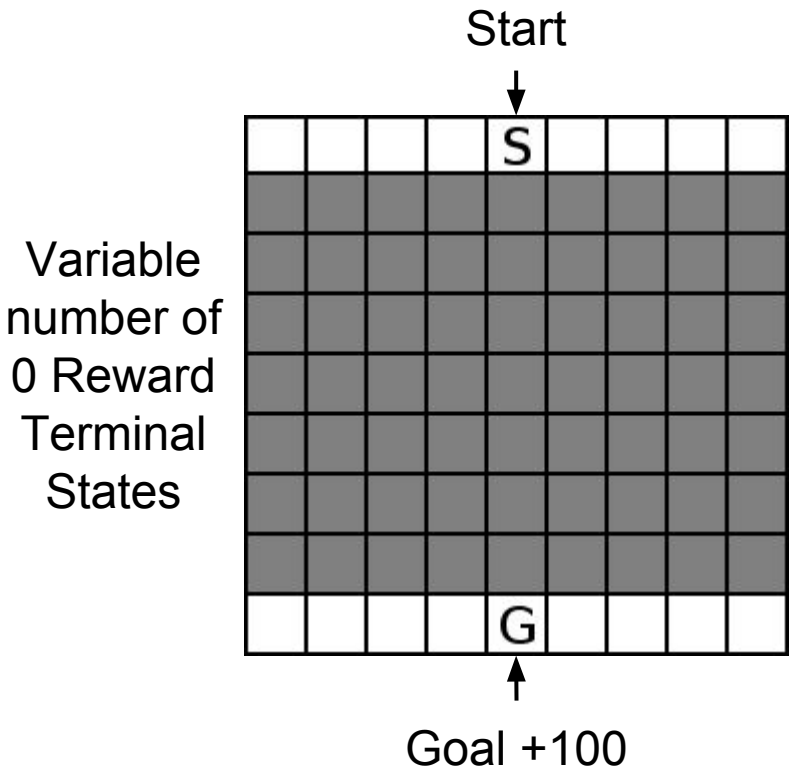
Learning Agents Research Group
The University of Texas at Austin

MCTS - Domain Dependence - Grid World



- 90% chance of moving in intended direction.
- 10% chance of moving to any neighbor randomly.

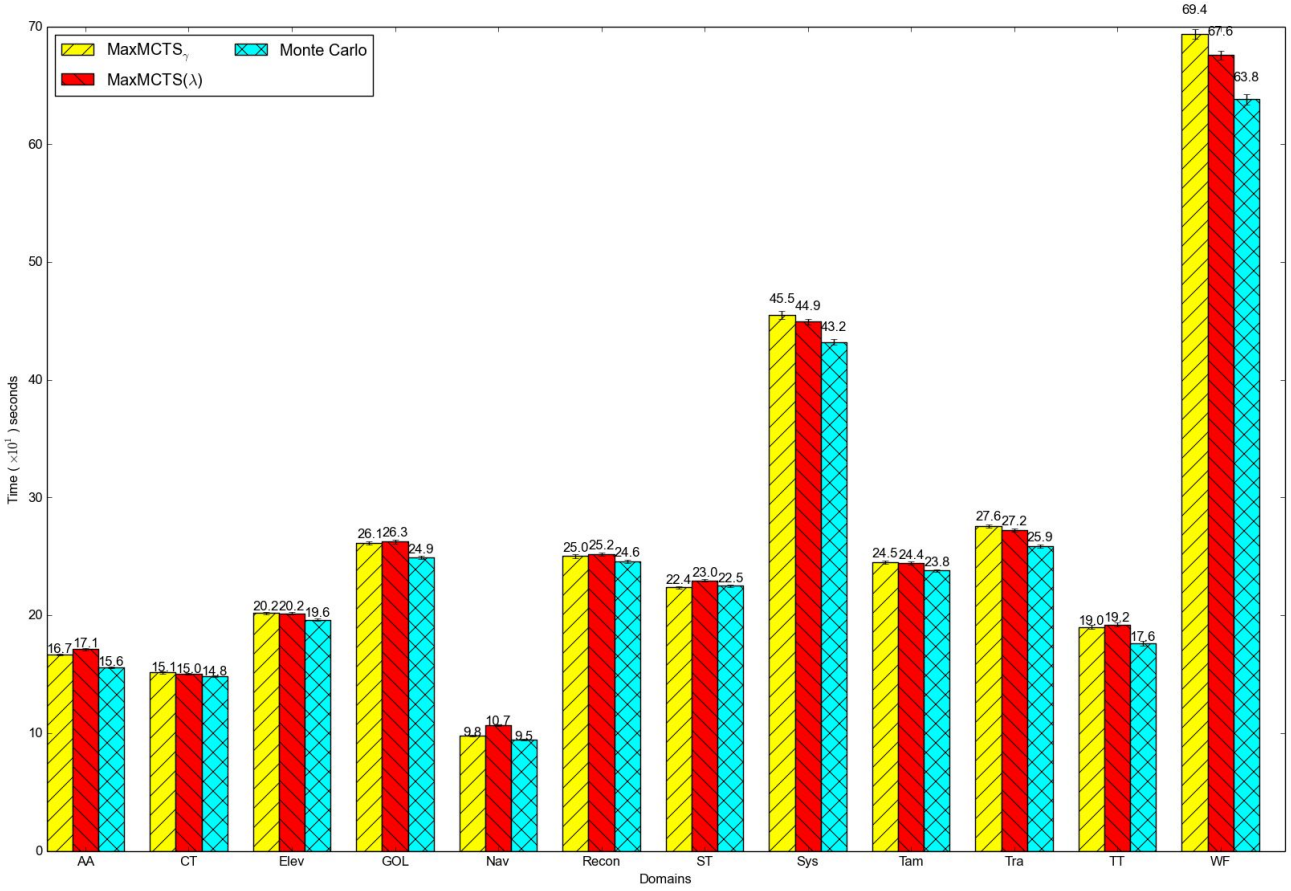
MCTS - Domain Dependence - Grid World



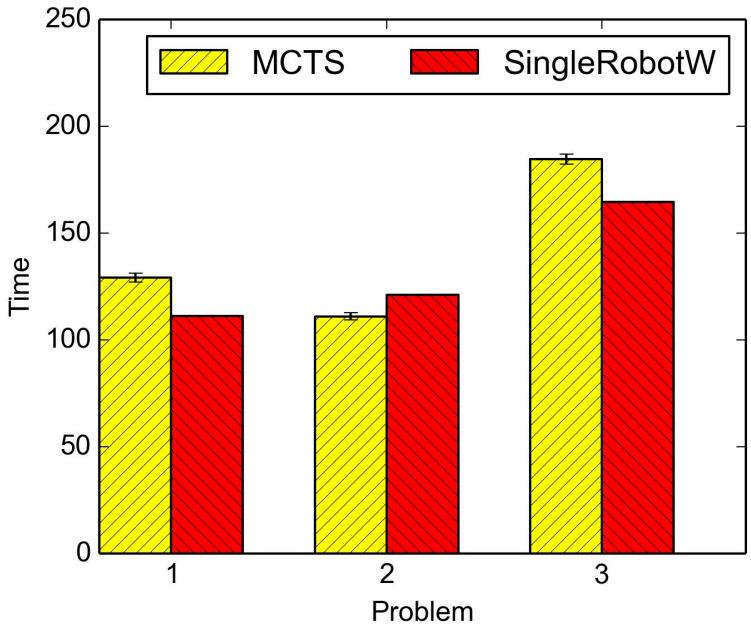
#0-Term	0	3	6	15
$\lambda = 1$	***90.4***	11.3	0.9	-2.2
$\lambda = 0.8$	90.2	28.0	10.7	-1.4
$\lambda = 0.6$	89.5	62.8	45.3	8.5
$\lambda = 0.4$	88.7	***85.1***	77.6	24.1
$\lambda = 0.2$	87.7	82.6	**78.1**	28.4
$\lambda = 0$	84.5	79.8	74.1	**31.8**

LARG
Learning Agents Research Group
The University of Texas at Austin

MCTS - Computational Time Comparison



Simulated User Study Results

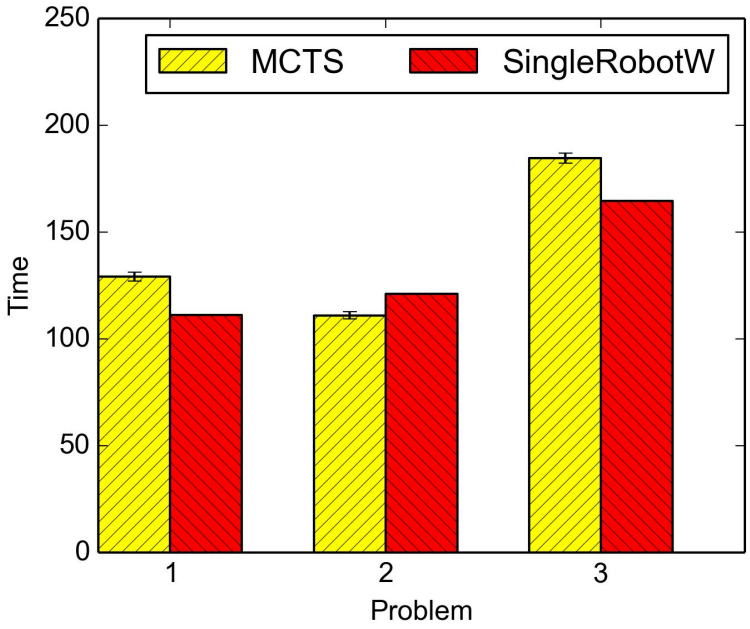


Simulation Results

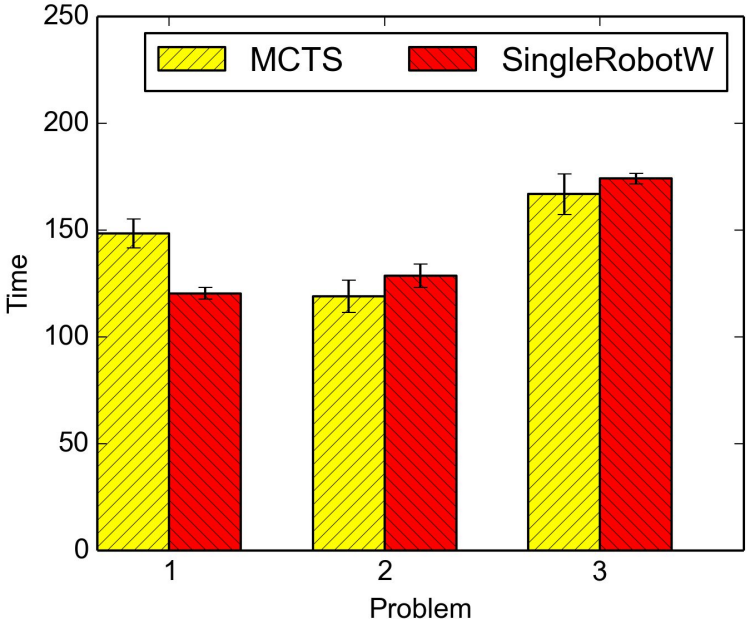
- 3 hand-selected problems.
- Compare time human takes to reach the goal on average against simulation results using hand-coded human decision model.
- 22 participants.

Learning Agents Research Group
The University of Texas at Austin

Simulated User Study Results



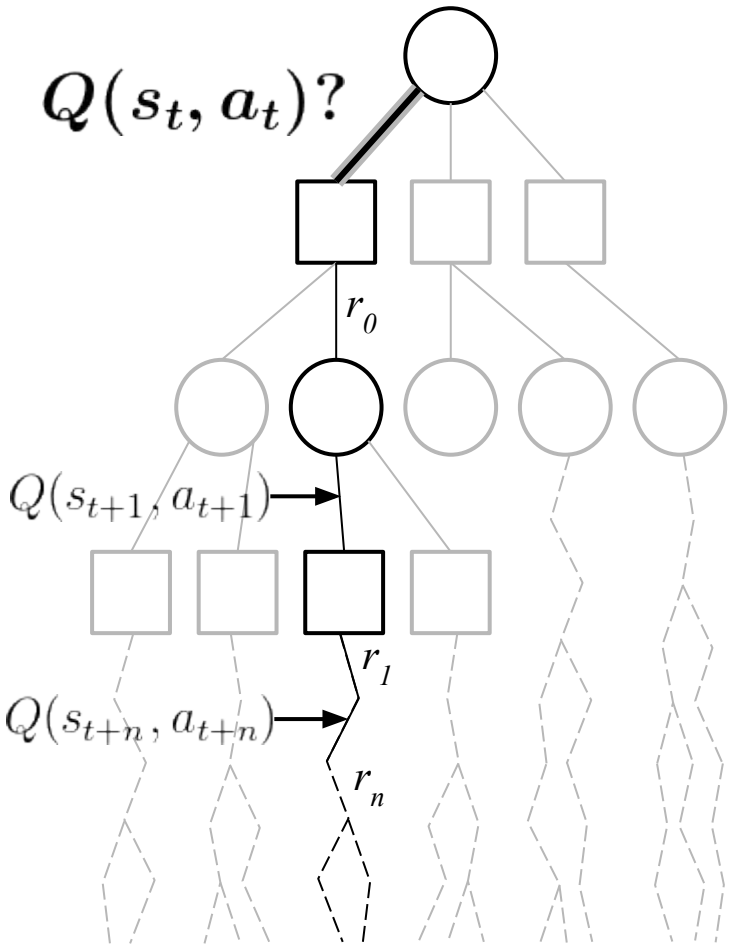
Simulation Results



User Study Results

Learning Agents Research Group
The University of Texas at Austin

MCTS - Novel Variants with Biased Backups!



Complex return: $R^C = \sum_{i=1}^{L} \left[w_{n,L} \cdot R^{(n)} \right]$

λ-return/eligibility [Rummery 1995]:

$$w^{\lambda}_{n,L} = \begin{cases} (1-\lambda)\lambda^{n-1} & 1 \leq n < L \\ \lambda^{L} & n = L \end{cases}$$

➡ **MCTS(λ)**

- ➢ Easier to implement.
- ➢ Assumes n-step return variances increase @ λ^{-1}.

γ-return weights [Konidaris et al. 2011]:

$$w^{\gamma}_{n,L} = \frac{\left(\sum_{i=1}^{n} \gamma^{2(i-1)}\right)^{-1}}{\sum_{n=1}^{L}\left(\sum_{i=1}^{n} \gamma^{2(i-1)}\right)^{-1}}$$

➡ **MCTSγ**

- ➢ Parameter free.
- ➢ Assumes n-step return variances are highly correlated.

LARG Learning Agents Research Group The University of Texas at Austin