Human and Agent Cooperative Learning

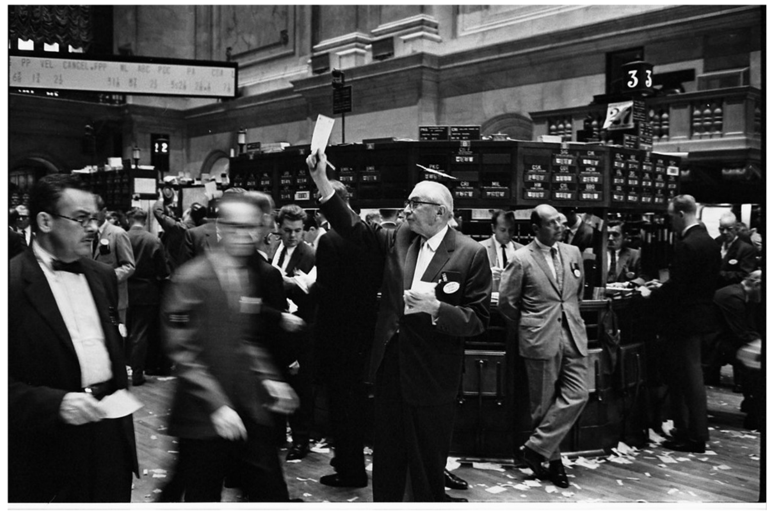
Matt Taylor

University of Alberta: Intelligent Robot Learning Lab (irll.ca)
Fellow-in-Residence: Alberta Machine Intelligence Institute (Amii.ca)
AI Redefined: Research Director (AI-R.com)

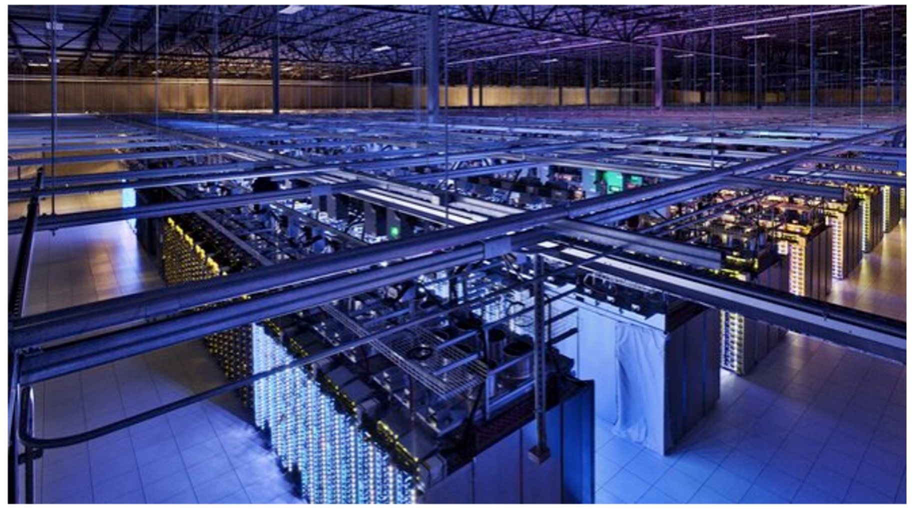
Canada CIFAR AI Chair, Amii

RL Applications

Developing vs. PoC vs. MVP

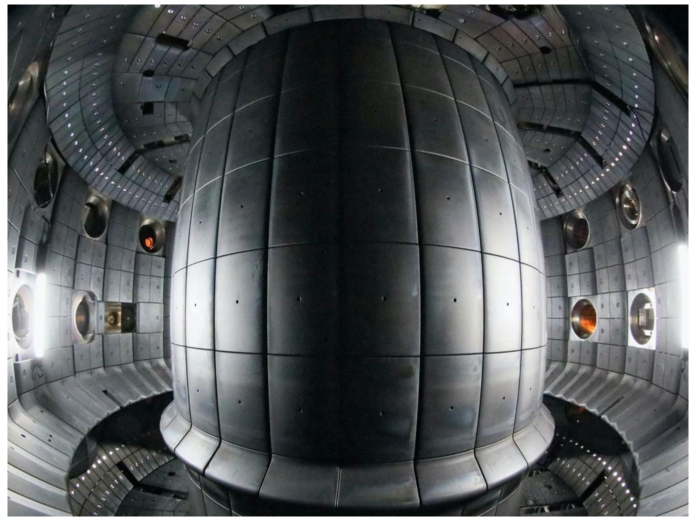


American Options Exercise Policy



Data Center Cooling

Balloon Control



Tokamak Fusion

AlphaGO

Stock Trading

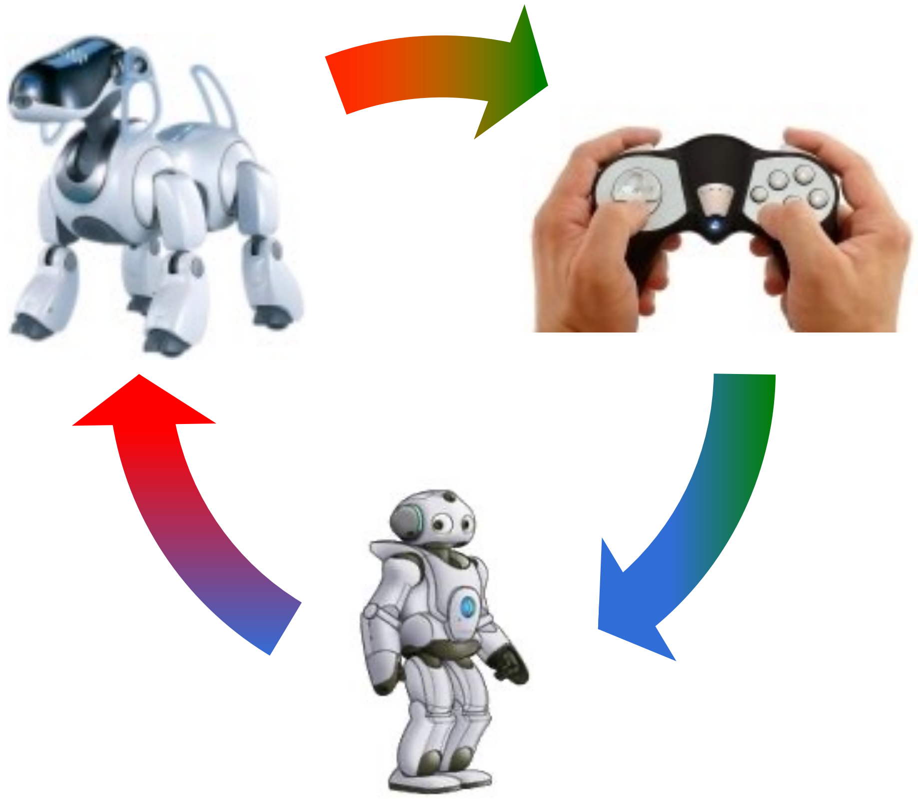
Water Treatment

?

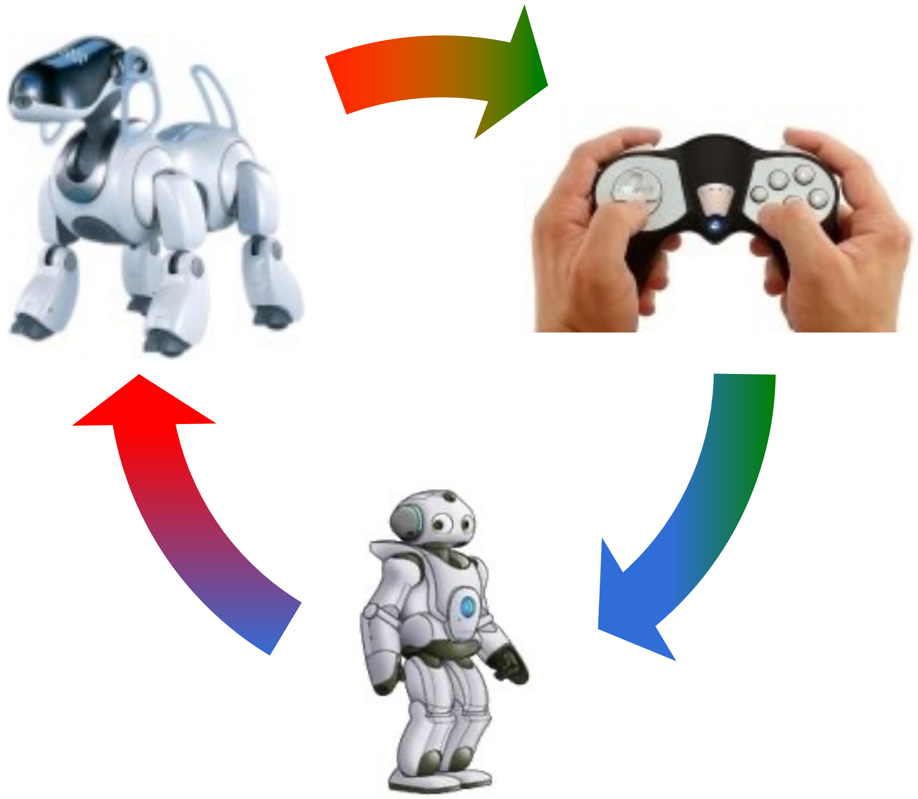
How do we deploy more RL solutions?

- Better understand how to identify, de-risk, and tackle real world problems
  - Challenges of Real-World Reinforcement Learning
- Understand how and when to “cheat” by using external information
  - Existing agents/data
  - Human knowledge

Cooperative Learning



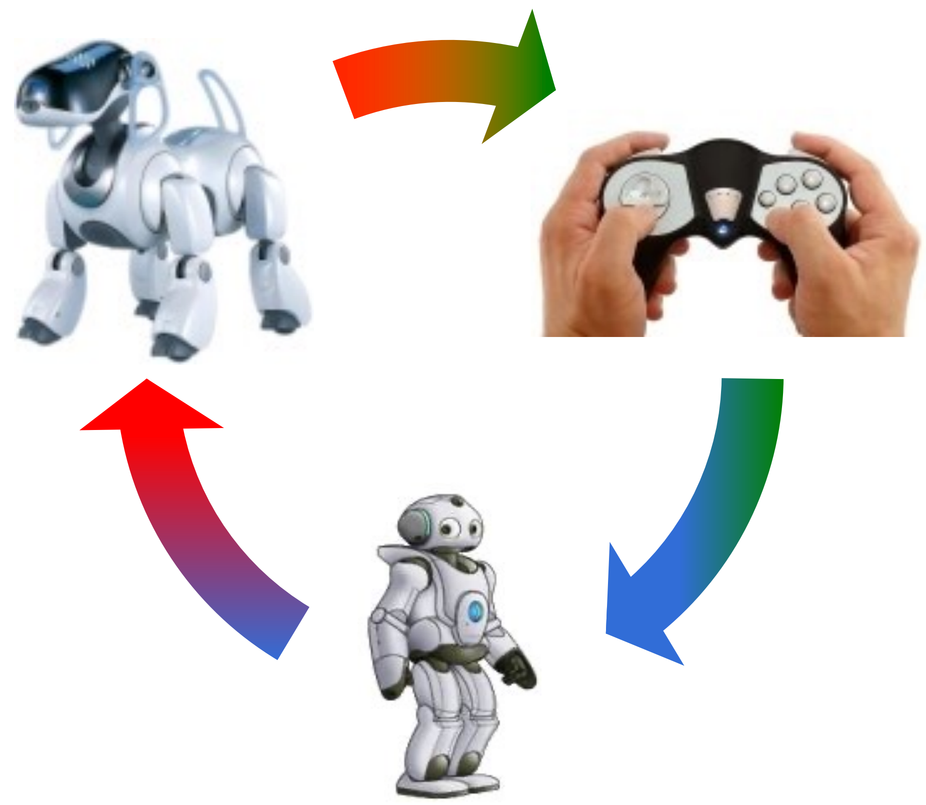
Cooperative Learning



Agent → Agent

- Offline / Batch RL
- Transfer Learning
- Curriculum Learning / Meta RL
- Advice

Cooperative Learning

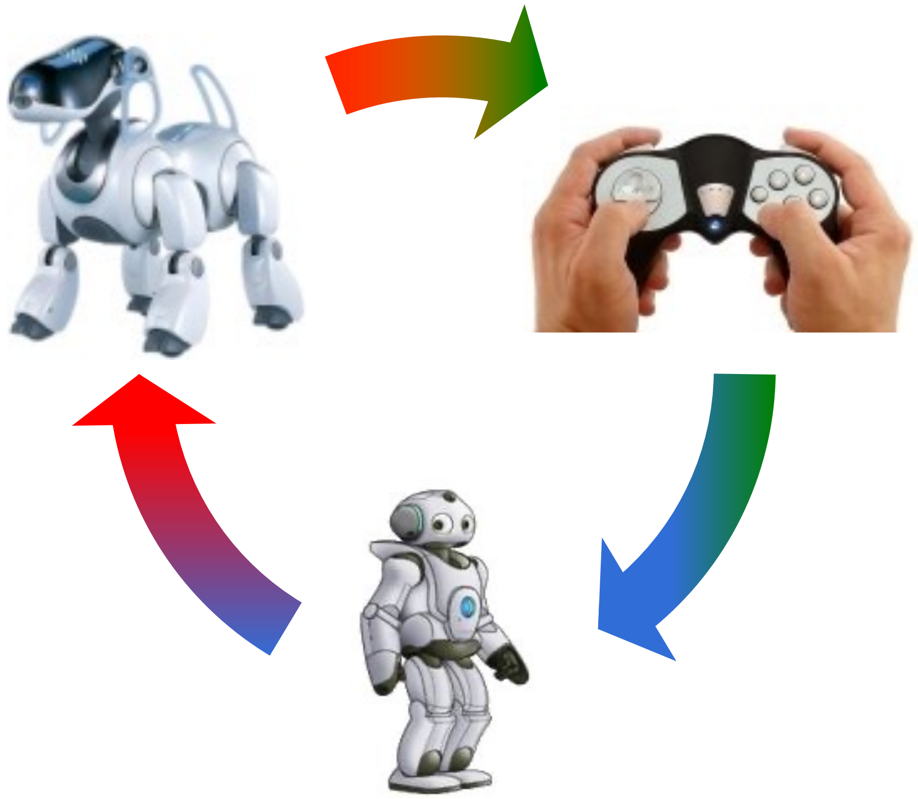


Agent → Agent

Human → Agent

- Offline / Batch RL
- Demonstrations
- Curriculum Learning / Meta RL
- Advice

Cooperative Learning



Agent → Agent

Human → Agent

Agent → Human

- Intelligent Tutoring Systems

Agent → Agent Bootstrapping

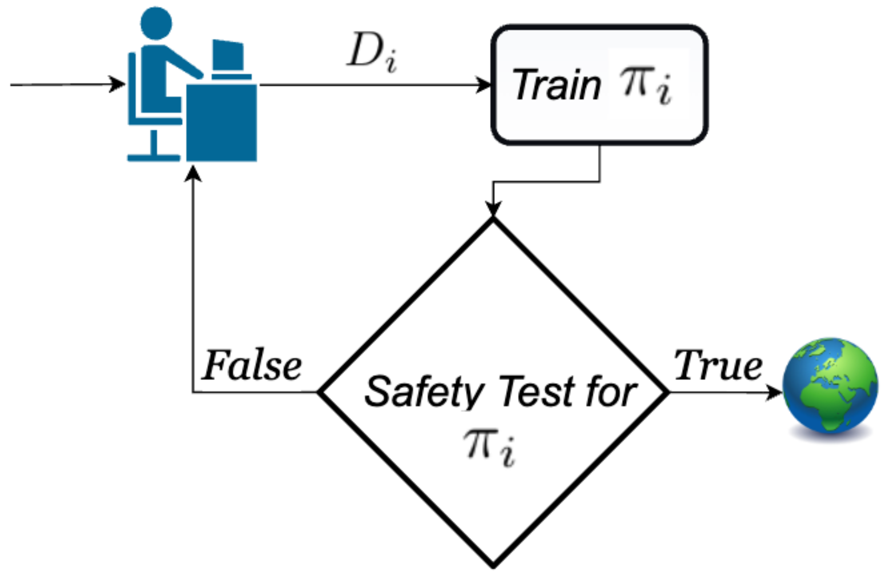
- Offline / Batch RL
- Transfer Learning
- Curriculum Learning / Meta RL
- Advice

Prior agent is ~~optimal~~ OK

How much data do I need?

How good is my policy? (OPE)

How sure am I about the policy's performance? (HCOPE)



"Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems." Levine, S.; Kumar, A.; Tucker, G.; and Fu, J. 2020.

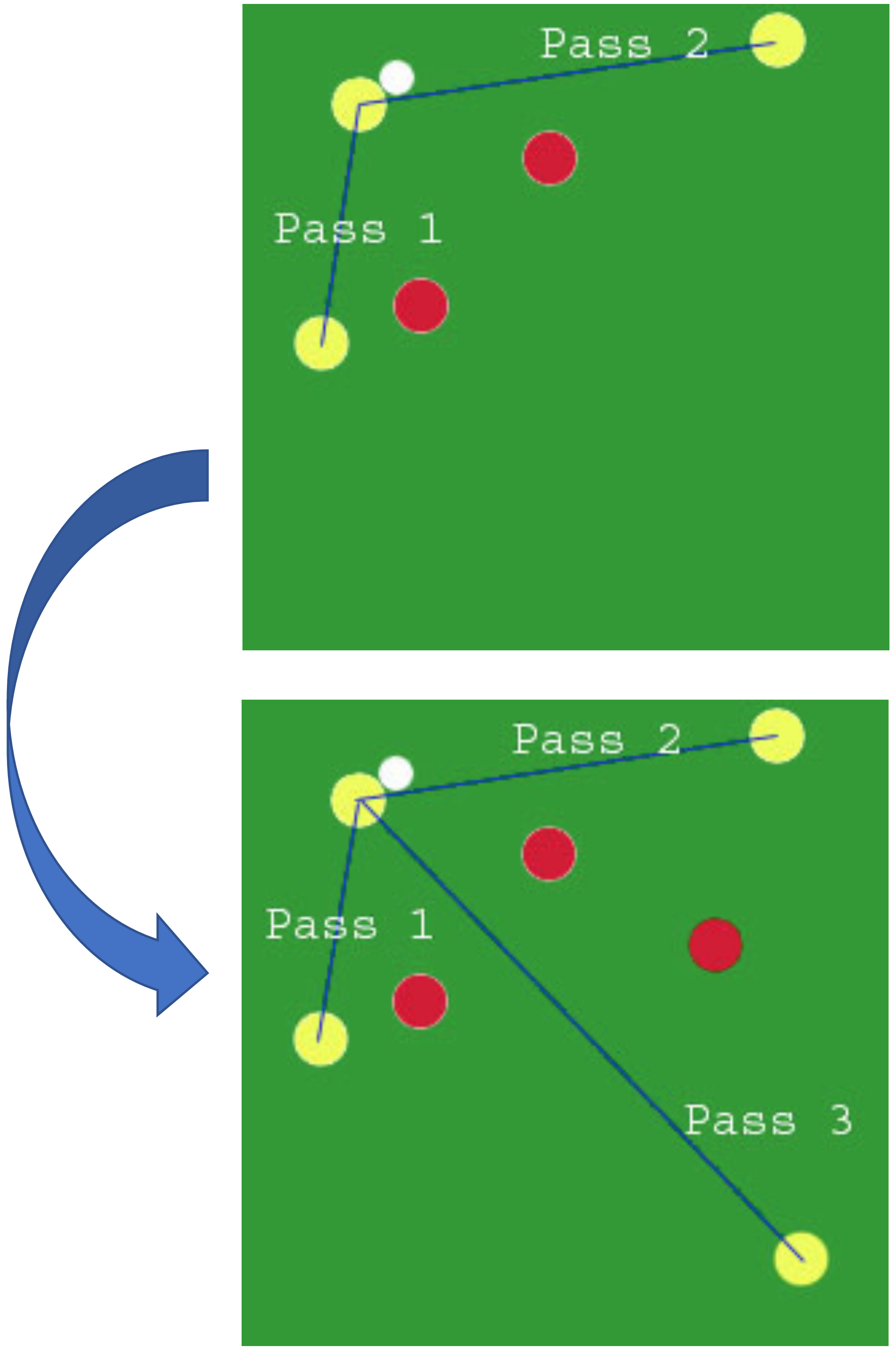
Agent → Agent Transfer

- Offline / Batch RL
- Transfer Learning
- Curriculum Learning / Meta RL
- Advice

What to transfer? (From whom to transfer?)

How to transfer?

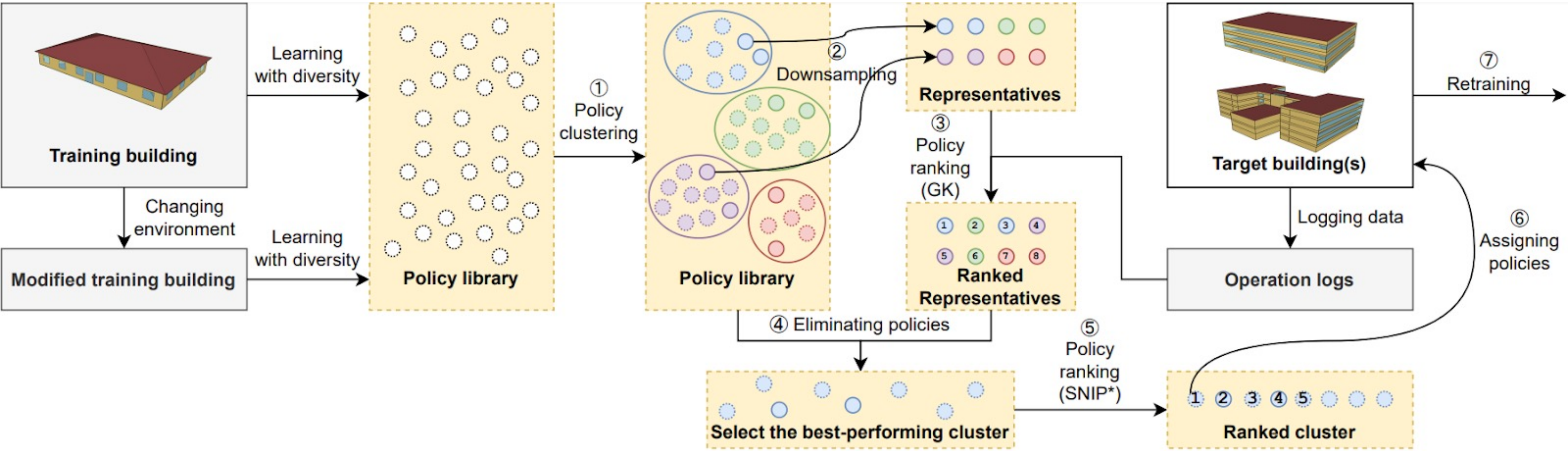
When to transfer?



"Transfer Learning for Reinforcement Learning Domains: A Survey." Taylor, and Peter Stone. 2009.

Agent → Agent Transfer

- Offline / Batch RL
- Transfer Learning
- Curriculum Learning / Meta RL
- Advice



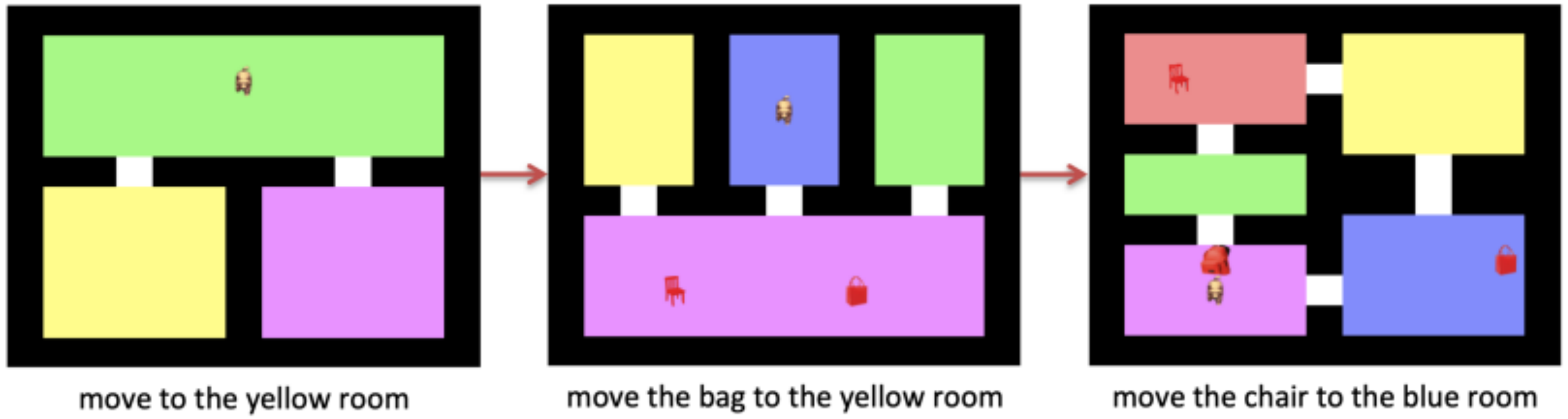
"Mitigating an Adoption Barrier of Reinforcement Learning-based Control Strategies in Buildings." Aakash Krishna G.S., Tianyu Zhang, Omid Ardakanian, Taylor. Energy & Buildings, 2023

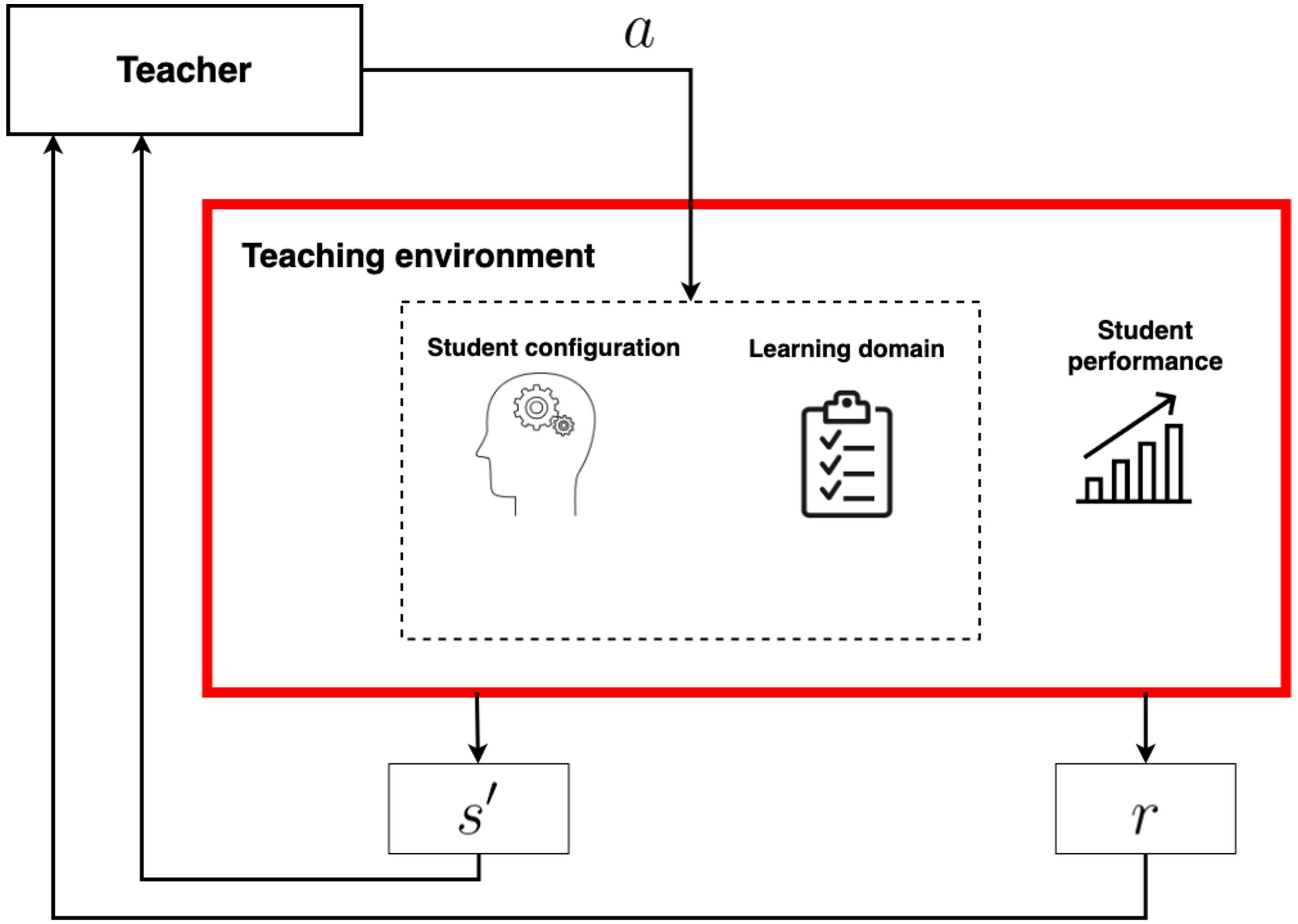
Agent → Agent Online Learning

- Offline / Batch RL
- Transfer Learning
- Curriculum Learning / Meta RL
- Advice

Curriculum Learning

Meta RL





"Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey." Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Taylor, and Peter Stone. 2020.

Muslimani, Lewandowski, Luo, Schuurmans

Agent → Agent Advice

- Offline / Batch RL
- Transfer Learning
- Curriculum Learning / Meta RL
- Advice

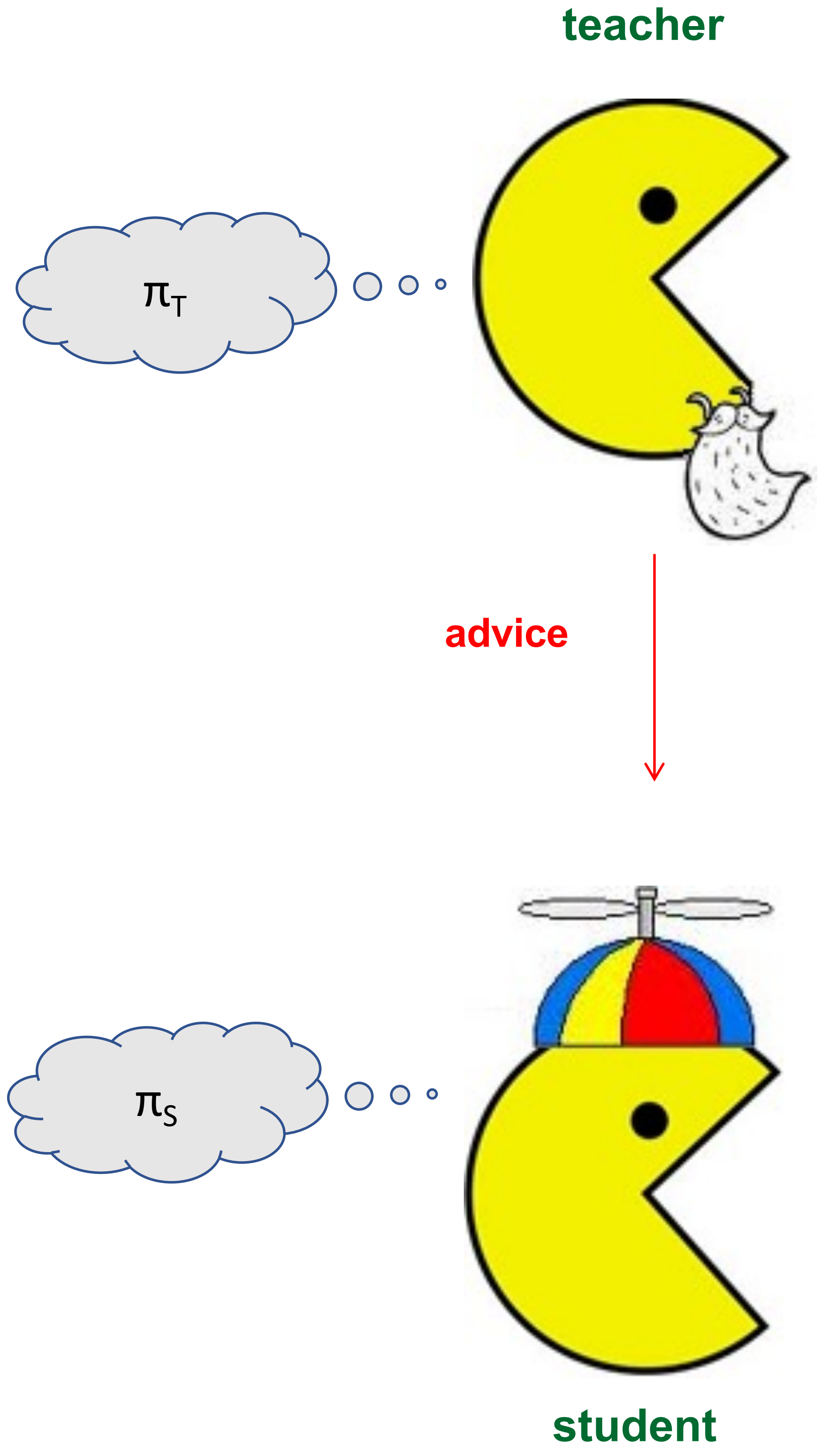
On-demand advice

Who initiates?

When do they provide?

Is there a cost?

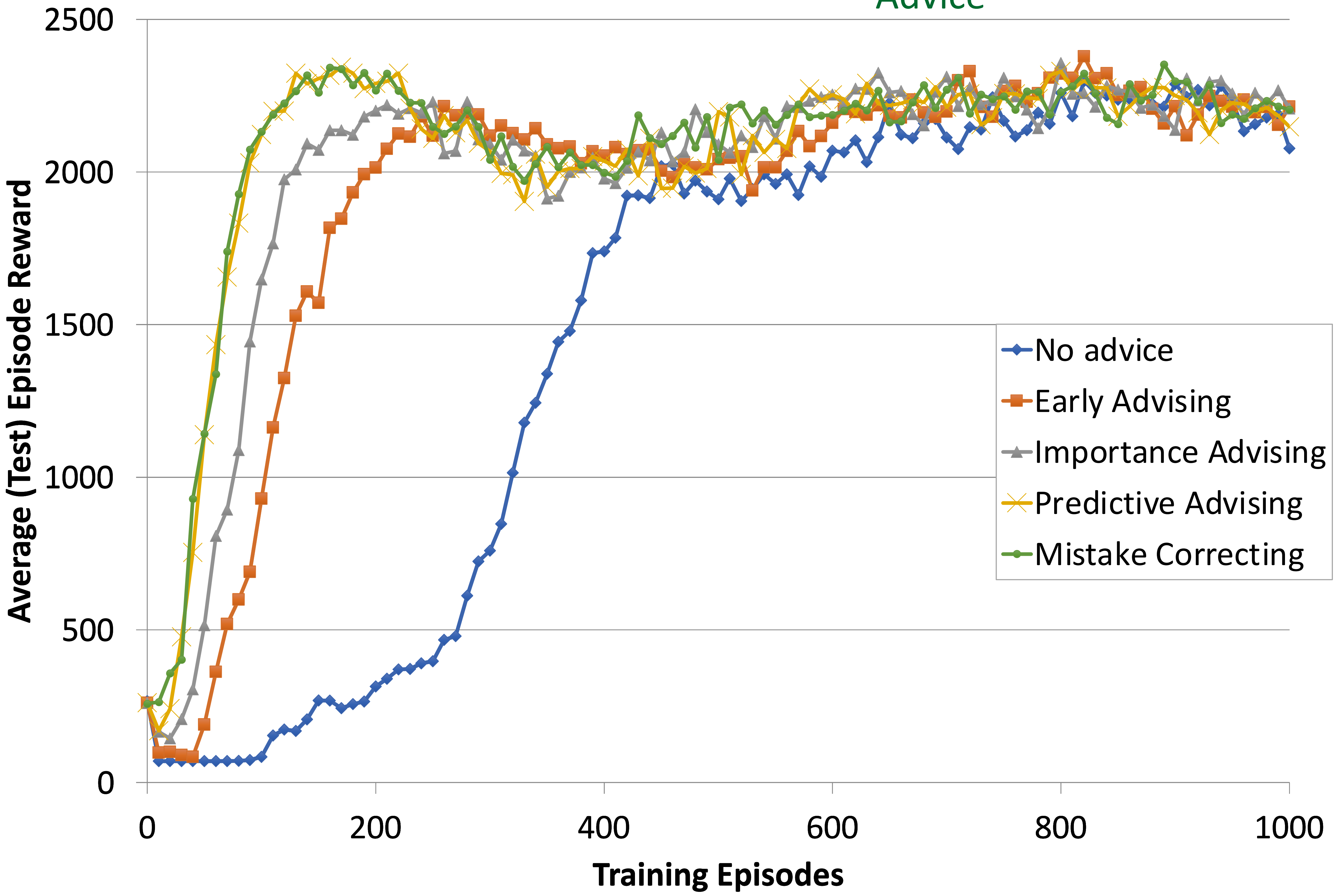
Are there multiple teachers?



“A Conceptual Framework for Externally-Influenced Agents: An Assisted Reinforcement Learning Review.” Adam Bignold, Francisco Cruz, Taylor, Tim Brys, Richard Dazeley, Peter Vamplew, and Cameron Foale. 2021

Agent → Agent Advice

- Offline / Batch RL
- Transfer Learning
- Curriculum Learning / Meta RL
- Advice



"Integrating Reinforcement Learning with Human Demonstrations of Varying Ability." Taylor, Halit Bener Suay, and Sonia Chernova. AAMAS-11.

Multi-Agent Advisor Q-Learning.

Subramanian, S., Taylor, K. Larson, & M. Crowley. JMLR-22

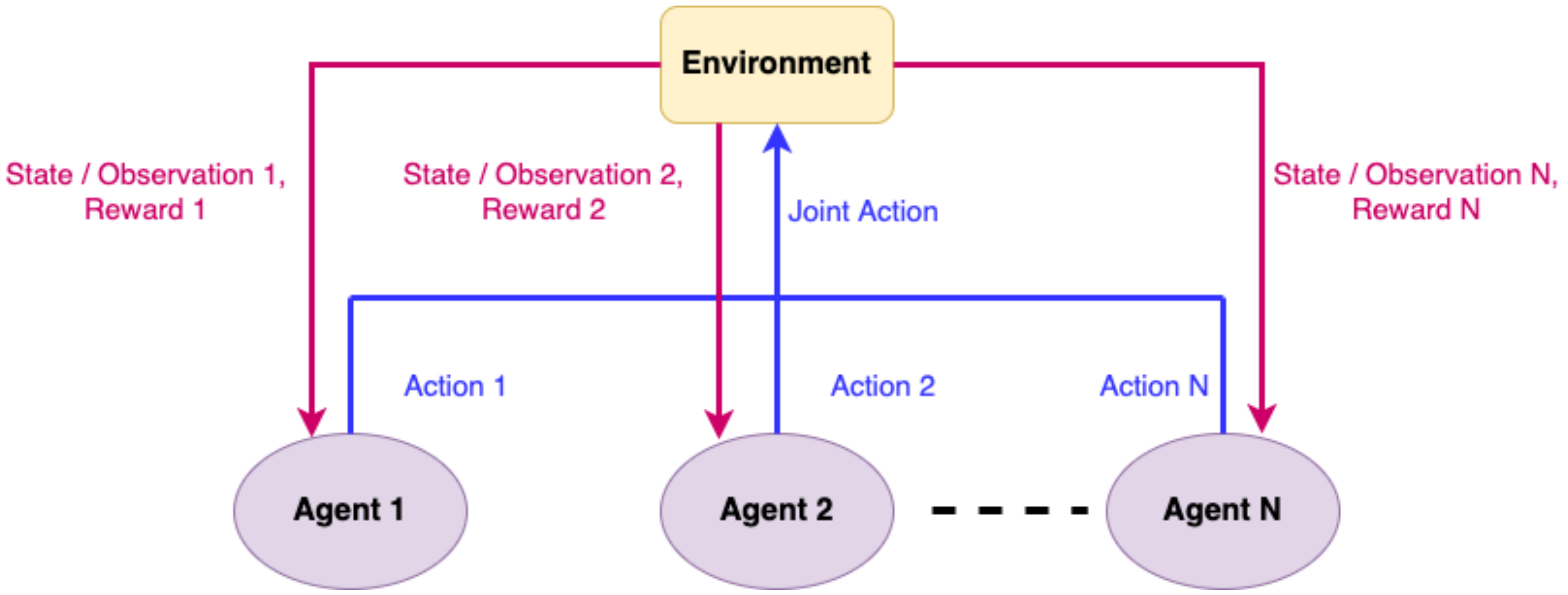
Sriram Ganapathi Subramanian

Kate Larson

Mark Crowley

“Multi-Agent Advisor Q-Learning.” S. Ganapathi Subramanian, S., Taylor, K. Larson, & M. Crowley. 2022.

ADvising Multiple Intelligent Agents (ADMIRAL)



Improving MARL sample efficiency

- We introduce: Multi-agent action advising for MARL
  - General methods (no assumption on advisor or type of environment)
  - Two practical algorithms to learn from advisor
  - Principled method to evaluate the advisor
  - Theoretical guarantees of convergence

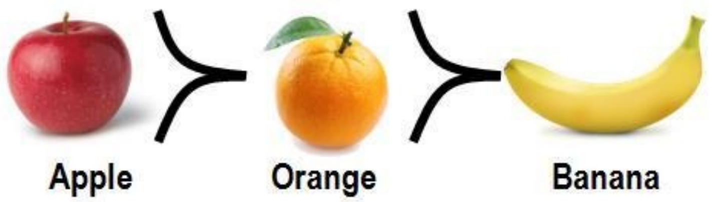
Human → Agent

- Offline / Batch RL
- Demonstrations
- Curriculum Learning / Meta RL
- Advice

Reward signal?

No

- Demonstrations
- Feedback
- Preferences



Yes

- Demonstrations
- Feedback
- Preferences
- Action Advice
- Shaping Rewards

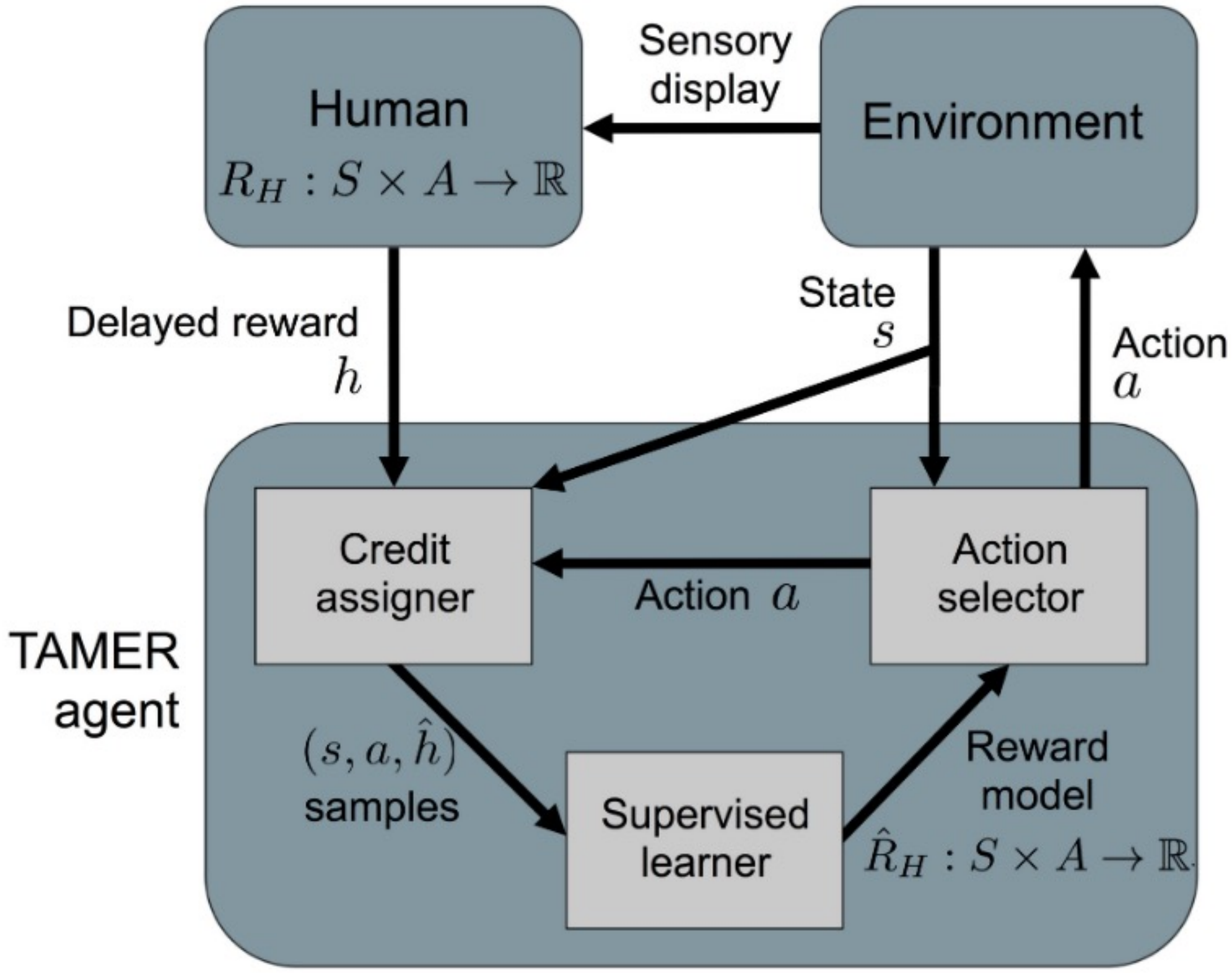
"A Conceptual Framework for Externally-Influenced Agents: An Assisted Reinforcement Learning Review." Adam Bignold, Francisco Cruz, Taylor, Tim Brys, Richard Dazeley, Peter Vamplew, and Cameron Foale. 2021.

Human $\rightarrow$ Agent Feedback (!R)

- Offline / Batch RL
- Demonstrations
- Curriculum Learning / Meta RL
- Advice

Thomaz & Breazeal 2006: Anticipator

TAMER, Knox & Stone 2009



Human → Agent Feedback (!R)

- Offline / Batch RL
- Demonstrations
- Curriculum Learning / Meta RL
- Advice

Thomaz & Breazeal 2006: Anticipator

TAMER, Knox & Stone 2009: Numeric, Return

SABL, Loftin+ 2015

Feedback history *h*

Observation: “sit”, Action:  , Feedback: “Bad Dog”

Observation: “sit”, Action:  , Feedback: “Good Boy”

…

Really make sense to assign numeric rewards to these?

Human → Agent Feedback (!R)

- Offline / Batch RL
- Demonstrations
- Curriculum Learning / Meta RL
- Advice

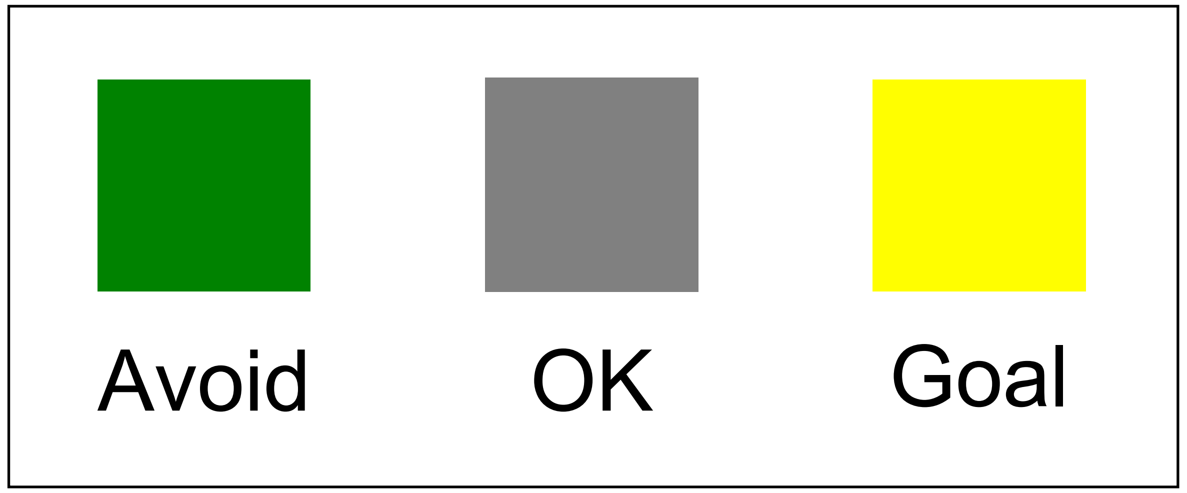
Thomaz & Breazeal 2006: Anticipator

TAMER, Knox & Stone 2009: Numeric, Return

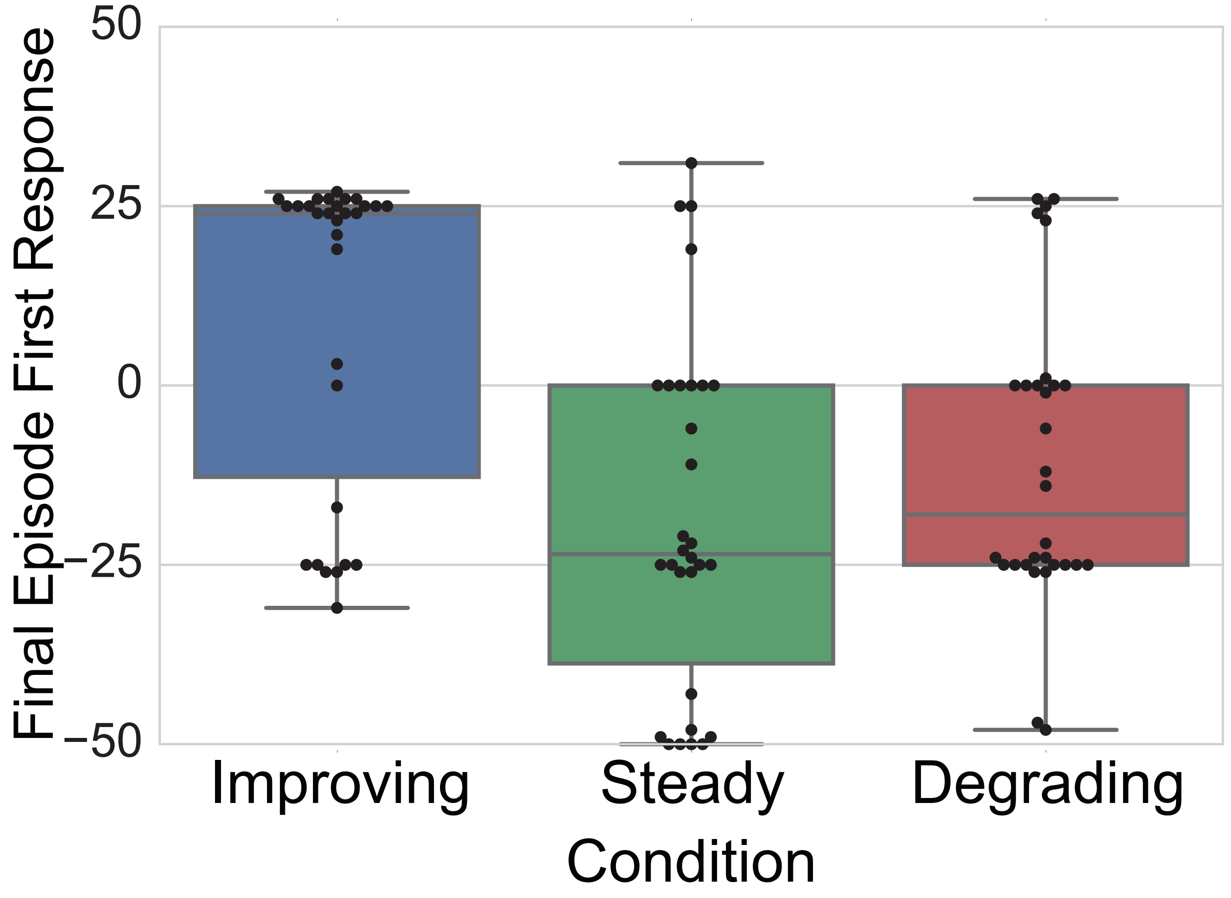
SABL, Loftin+ 2015: Categorical

COACH, McGlashlin+ 2017

Feedback can be Relative



Feedback can be Relative



Improving Condition: bad → bad → alright

Steady Condition: alright → alright → alright

Degrading Condition: good → good → alright

Advantage Function!

Human → Agent Feedback (!R)

- Offline / Batch RL
- Demonstrations
- Curriculum Learning / Meta RL
- Advice

Thomaz & Breazeal 2006: Anticipator

TAMER, Knox & Stone 2009: Numeric, Return

SABL, Loftin+ 2015: Categorical

COACH, McGlashlin+ 2017: Advantage Function

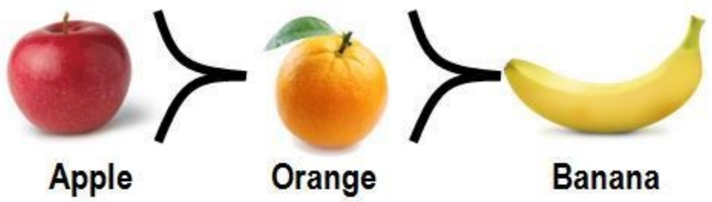
Human → Agent

- Offline / Batch RL
- Demonstrations
- Curriculum Learning / Meta RL
- Advice

Reward signal?

No

- Demonstrations
- Feedback
- Preferences



Yes

- Demonstrations
- Feedback
- Preferences
- Action Advice
- Shaping Rewards

“A Conceptual Framework for Externally-Influenced Agents: An Assisted Reinforcement Learning Review.” Adam Bignold, Francisco Cruz, Taylor, Tim Brys, Richard Dazeley, Peter Vamplew, and Cameron Foale. 2021.

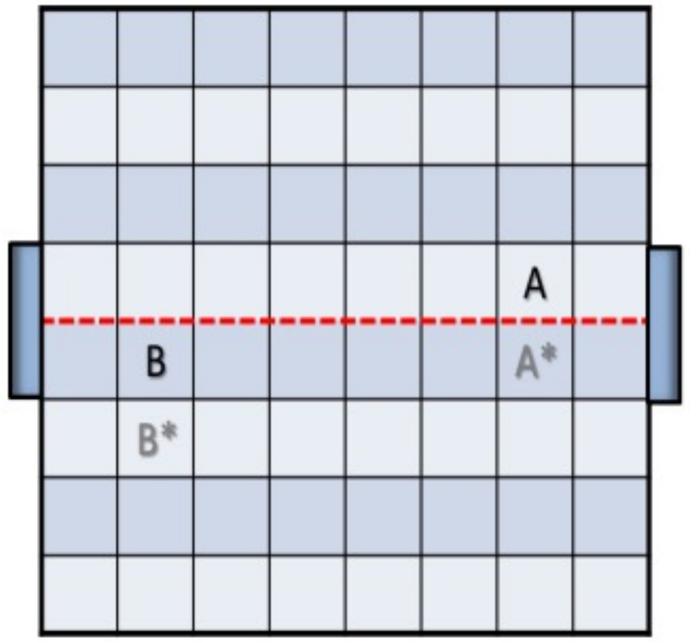
Human → Agent Bootstrapping

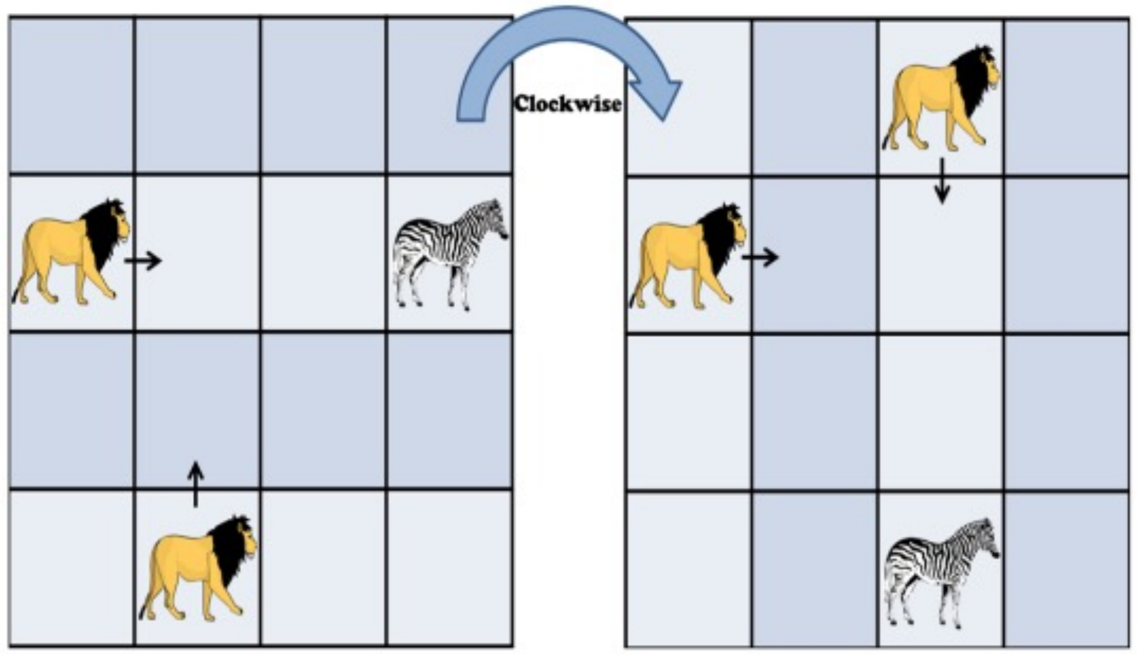
- Offline / Batch RL
- Demonstrations
- Curriculum Learning / Meta RL
- Advice

Lay person

Subject Matter Expert

Programmer





Leveraging Human Knowledge in Tabular Reinforcement Learning: A Study of Human Subjects. Ariel Rosenfeld, Matthew E. Taylor, and Sarit Kraus. IJCAI-17

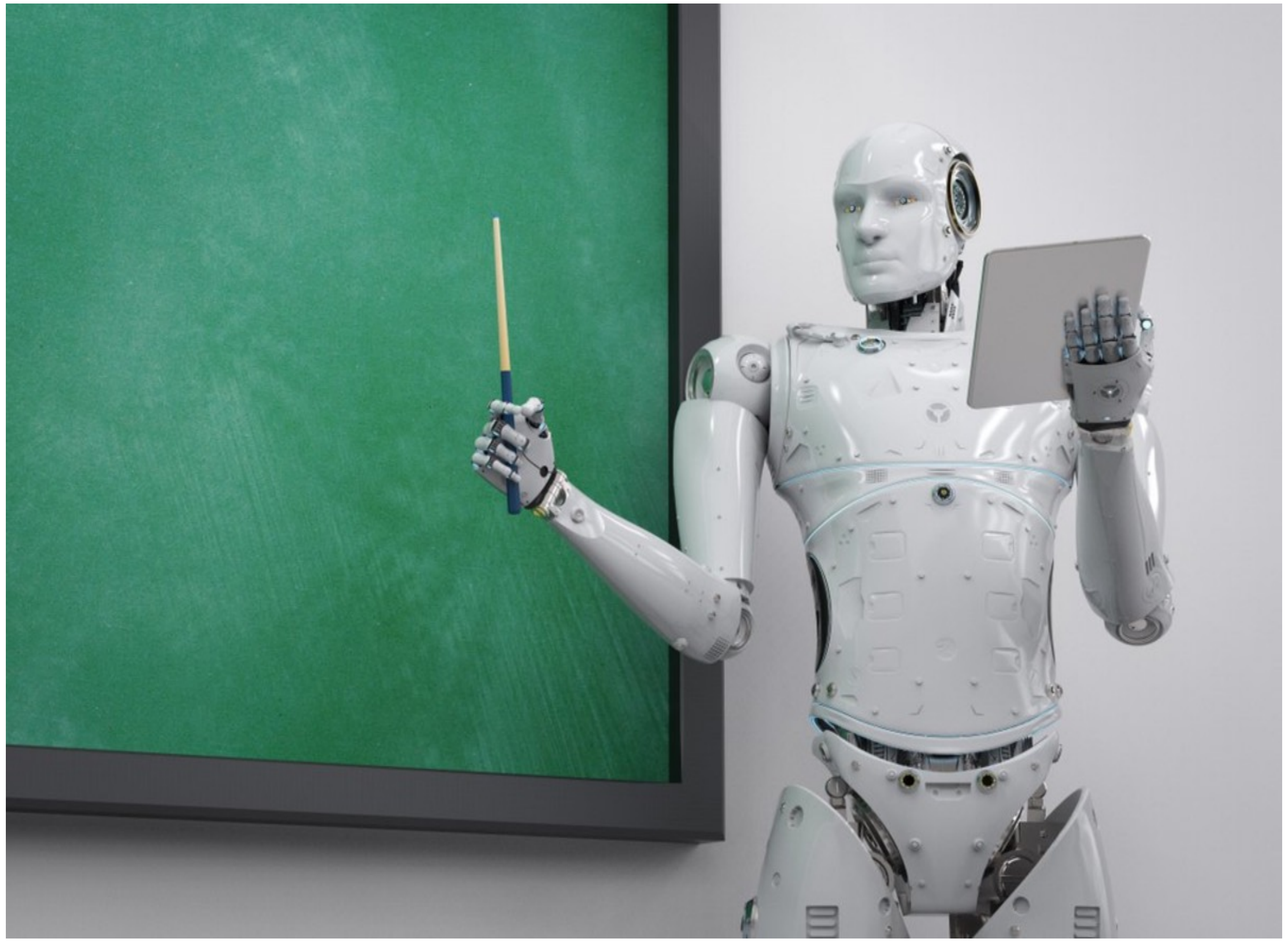
Agent → Human: ITS

Convey information

Model user’s understanding

Model user’s learning

→ Sequential decision tasks



https://hassanmachmouchiblog.files.wordpress.com/2021/01/robot-teachers.png

Agent → Human: ITS

How to practice

How to support

When to support

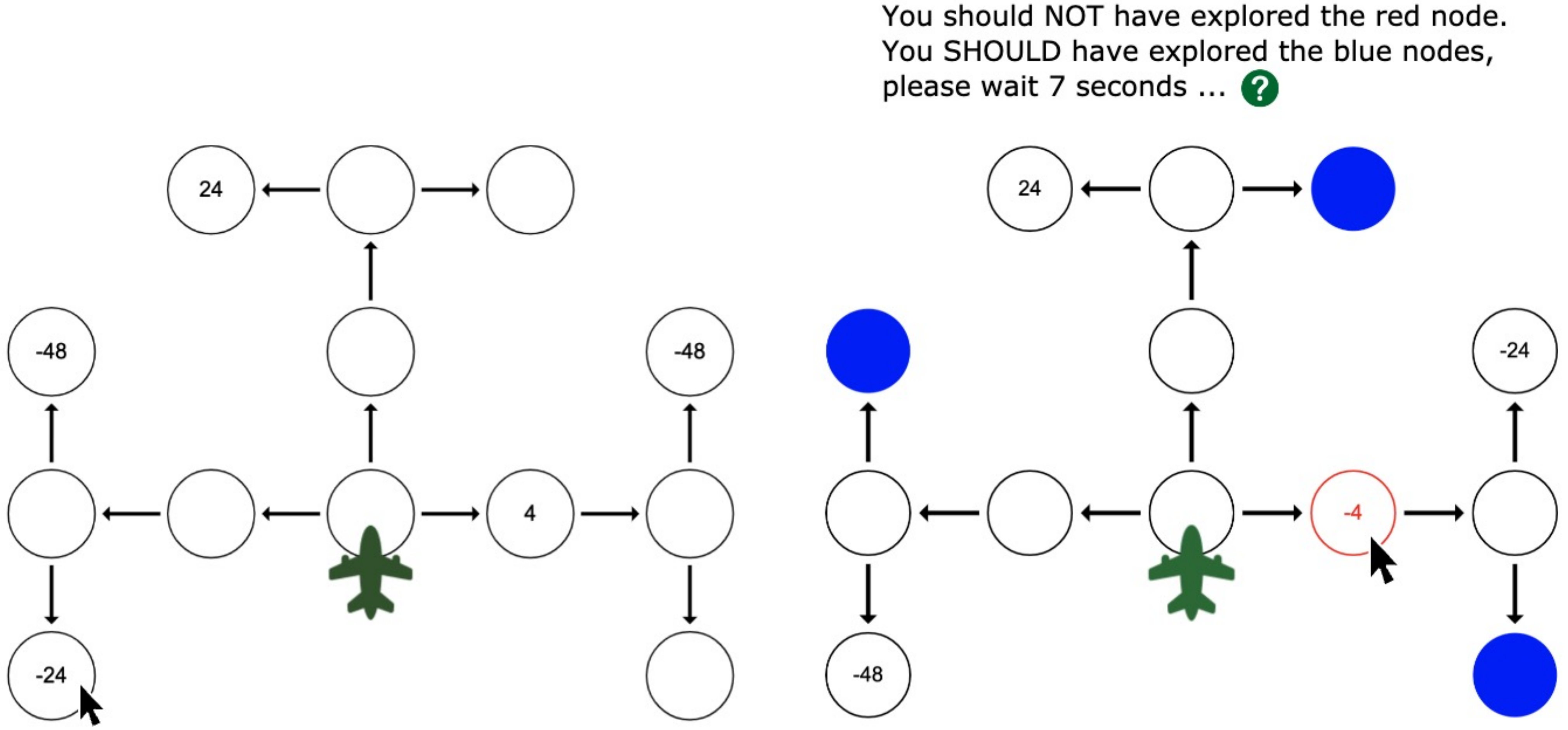


Fig. 1: The Mouselab-MDP paradigm. (Left) Participants click to reveal the value at future states. (Right) ITS provides feedback on each planning operation. The question mark represents optional elaborated feedback.

C^2 Tutor: Helping People Learn to Avoid Present Bias During Decision Making. Calarina Muslimani, Saba Gul, Taylor, Carrie Demmans Epp, Christabel Wayllace. AIED-23.

Agent → Human: Pilot Training

Shortage of pilots

Lots of knowledge

Hands on training

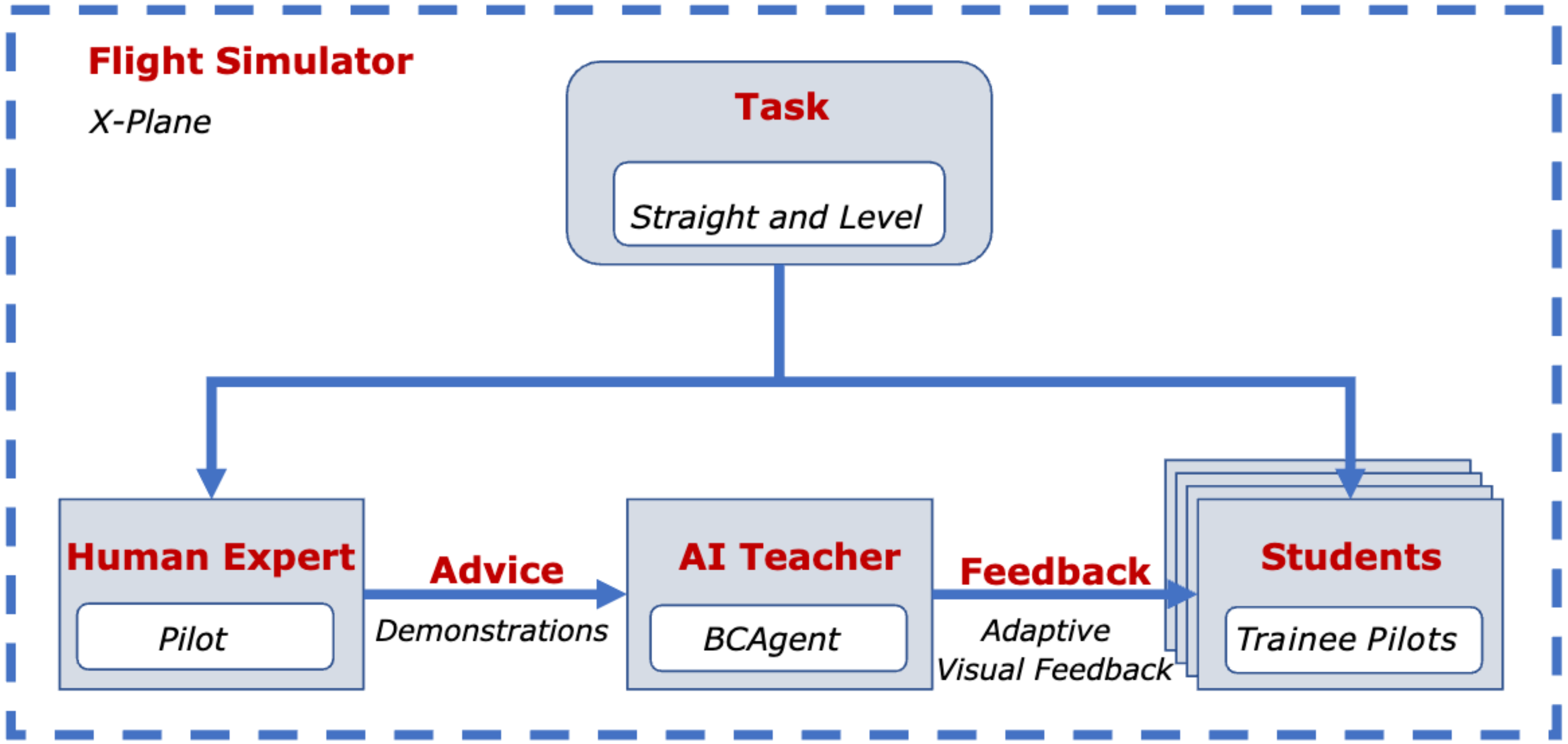


Figure 1: System Architecture

Augmenting Flight Training with AI to Efficiently Train Pilots. Michael Guevarra, Srijita Das, Christabel Wayllace, Carrie Demmans Epp,Taylor, Alan Tay. AAAI-23 Demo.

24 E
20:04
SELECT
CONTROL
FUEL
LEFT
RIGHT
QTY
AIRSPEED
KNOTS
VERTICAL SPEED
2 MIN.
X-PLANE 530
COM/VLOC
VLOC

Program Synthesis

Write better code faster

Program Optimization with Locally Improving search (POLIS)

- A system for improving programs w.r.t. reward
- Local search algorithm exploits program structure
- Generate effective & short programs

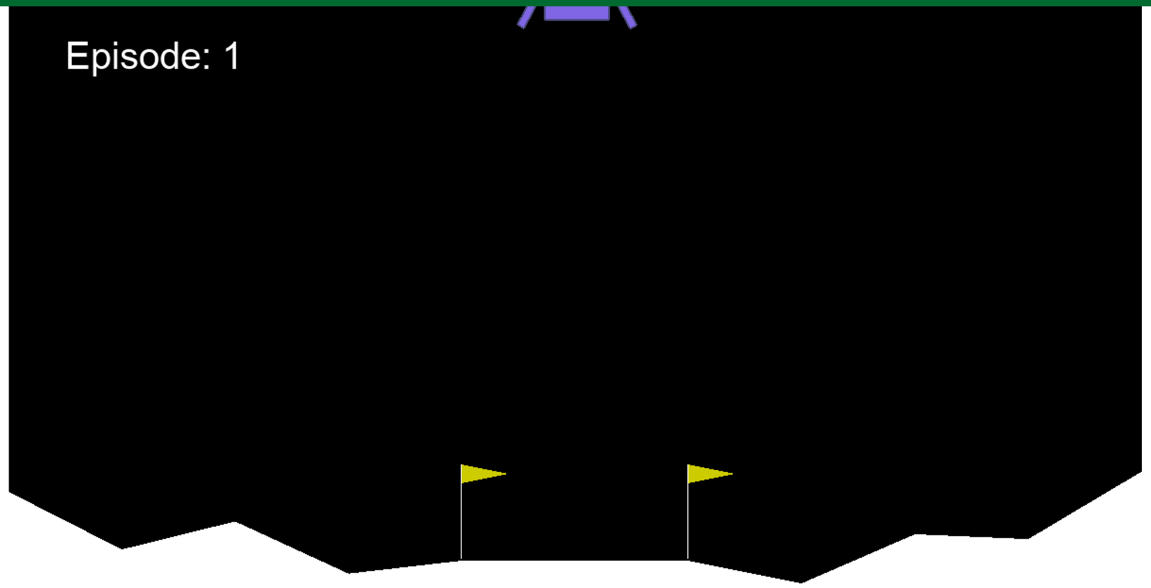
```python
def max_sum_slice(xs):
    max_ending = max_so_far = 0
    for x in xs:
        max_ending = max(0, max_ending + x)
        max_so_far = max(max_so_far, max_ending)
    return max_so_far
```

Copilot

Can You Improve My Code? Optimizing Programs with Local Search.
Fatemeh Abdollahi, Saqib Ameen, Levi Lelis, Taylor. IJCAI-23

POLIS

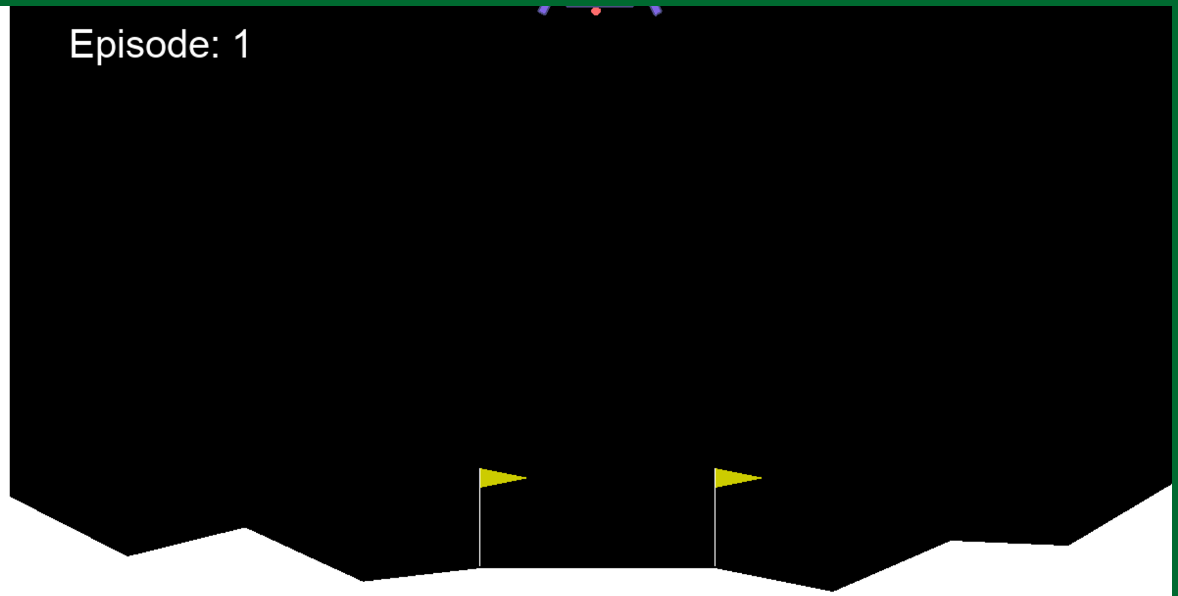
polis, plural **poleis**, ancient Greek city-state.... There were several hundred poleis, the history and constitutions of most of which are known only sketchily .... most ancient Greek history is recounted in terms of the histories of Athens, Sparta, and a few others.



Original program

Average score ~ -75

POLIS



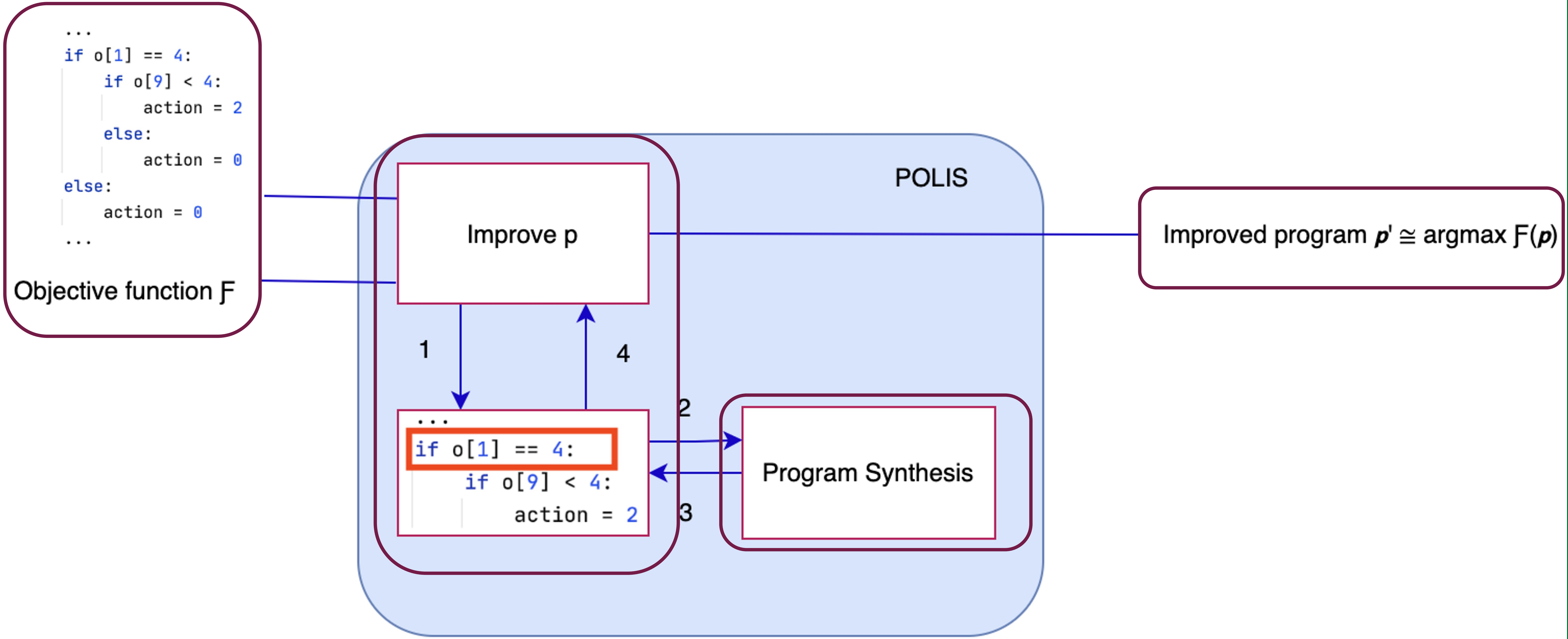

```python
def initial(o):
    if o[1] > 1 and o[1] < 1.1 and (o[4] < 0.12):
        action = 2
    elif o[1] > 1 and o[3] < -0.7 and o[0] < -0.05:
        action = 3
    elif o[1] > 1 and o[3] < -0.8 and o[0] > 0.1:
        action = 1
    elif o[0] < -0.15 and o[4] > 0.1:
        action = 3
    elif o[0] < 0.13 and o[4] < -0.1:
        action = 1
    elif o[1] < 0.8 and o[1] > 0.2:
        action = 2
    elif o[1] <= 0.2 and o[4] > 0.1:
        action = 3
    elif o[1] <= 0.2 and o[4] < -0.1:
        action = 1
    else:
        action = 0
    return action
```

```python
def improved(o):
    if o[3] > -0.038:
        action = 0
    elif o[7] > 0.036:
        action = 2
    elif o[5] < -0.1:
        action = 1
    elif o[0] and o[6]:
        action = 2
    elif o[5] > 0.959:
        action = 0
    elif o[3] < -0.388:
        action = 2
    elif o[4] > 0.1:
        action = 3
    elif o[2] > 0.28:
        action = 1
    else:
        action = 3
    return action
```

POLIS improved program

Average score ~ +190

POLIS



...
if o[1] == 4:
if o[9] < 4:
action = 2
else:
action = 0
else:
action = 0
...
Objective function Ƒ
POLIS
Improve p
1
4
2
3
...
if o[1] == 4:
if o[9] < 4:
action = 2
Program Synthesis
Improved program p' ≅ argmax Ƒ(p)

Bottom-Up Search (BUS)

cost 1 1 1 1 1 1

$$S \rightarrow \sqrt{S} \mid S + S \mid S - S \mid x_1 \mid x_2 \mid x_3$$

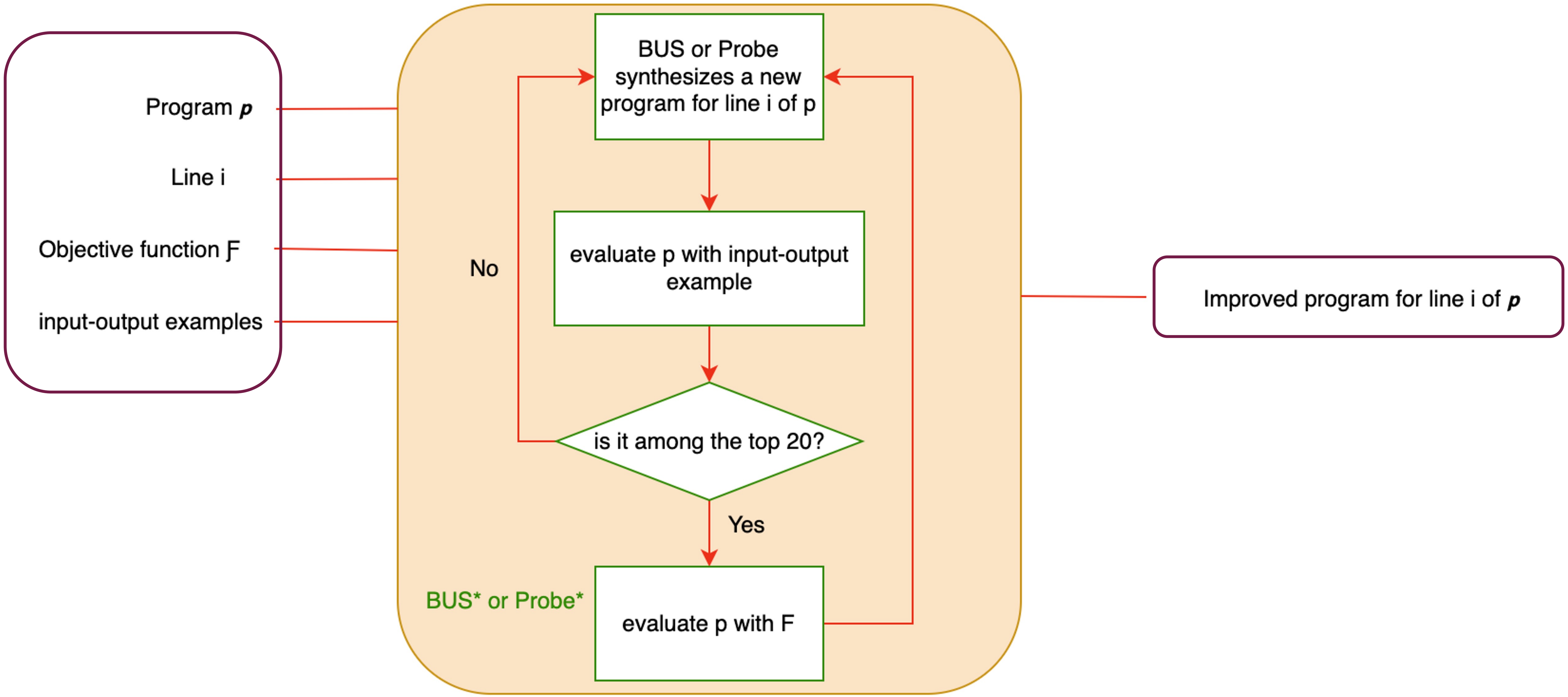
Cost	# Programs	Bank
1	3	$\{x_1, x_2, x_3\}$
$\boxed{2}$	3	$\{\sqrt{x_1}, \boxed{\sqrt{x_2},} \sqrt{x_3}\}$
3	75	$\{\sqrt{\sqrt{x_1}}, \sqrt{\sqrt{x_2}}, \sqrt{\sqrt{x_3}}, x_1 + x_1, \cdots, x_1 - x_1, x_1 - x_2, \cdots\}$
4	147	$\{\sqrt{\sqrt{\sqrt{x_1}}}, \cdots, \sqrt{x_1 + x_1}, \cdots, \sqrt{x_1 - x_1}, \cdots, \sqrt{x_1} + x_1, \cdots, \sqrt{x_1} - x_1, \cdots\}$
5	12K	$\{\cdots\}$
6	70K	$\{\cdots\}$
7	...	$\{\cdots \boxed{, \sqrt{\sqrt{x_1 + x_2} + x_3},} \cdots\}$

Guided BUS: Probe (Barke et. al. 2020)

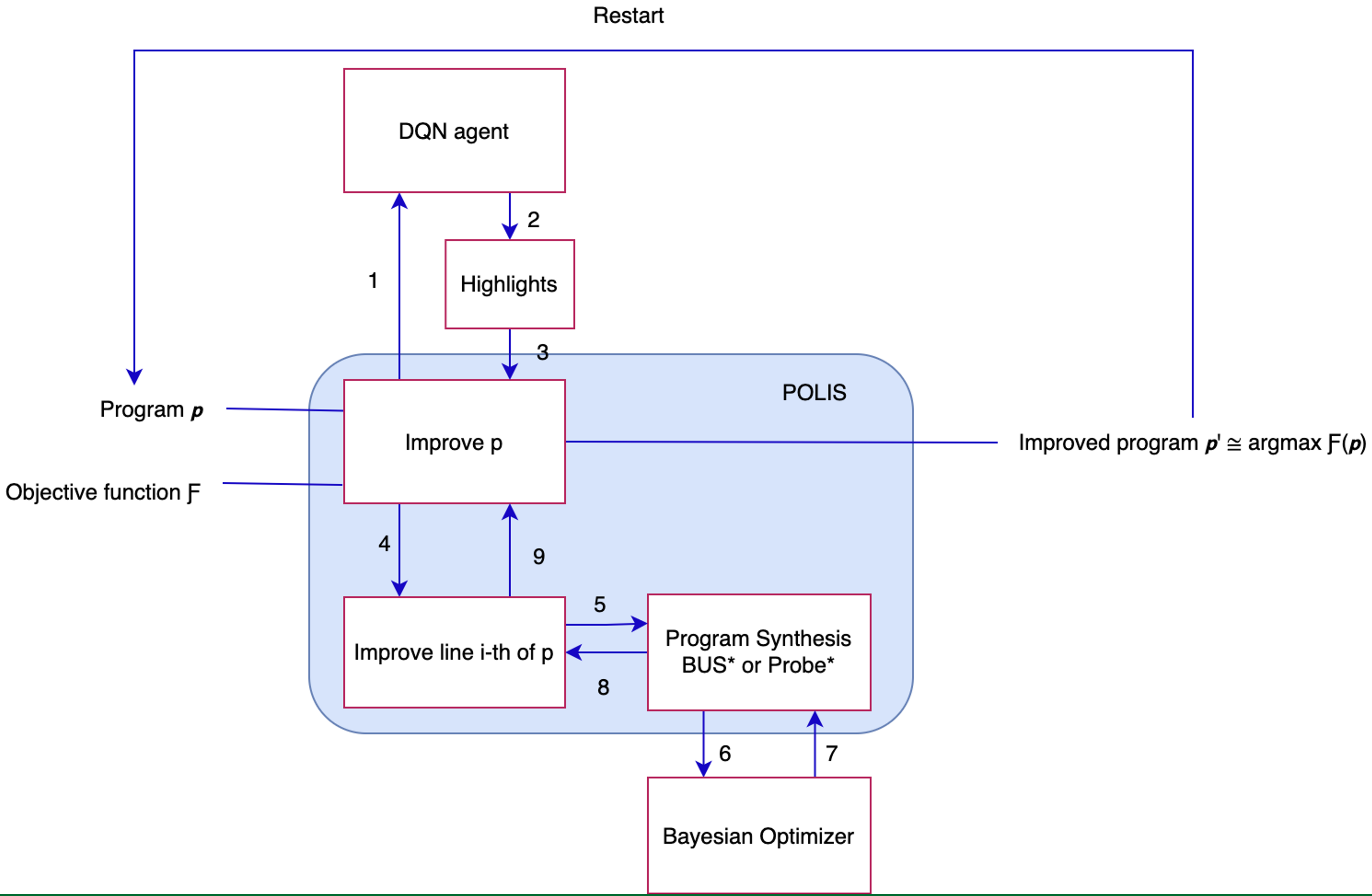
cost		1	1	5	2	2	2
	$S \rightarrow$	$\sqrt{S}$	$S + S$	$S - S$	x_1	x_2	x_3

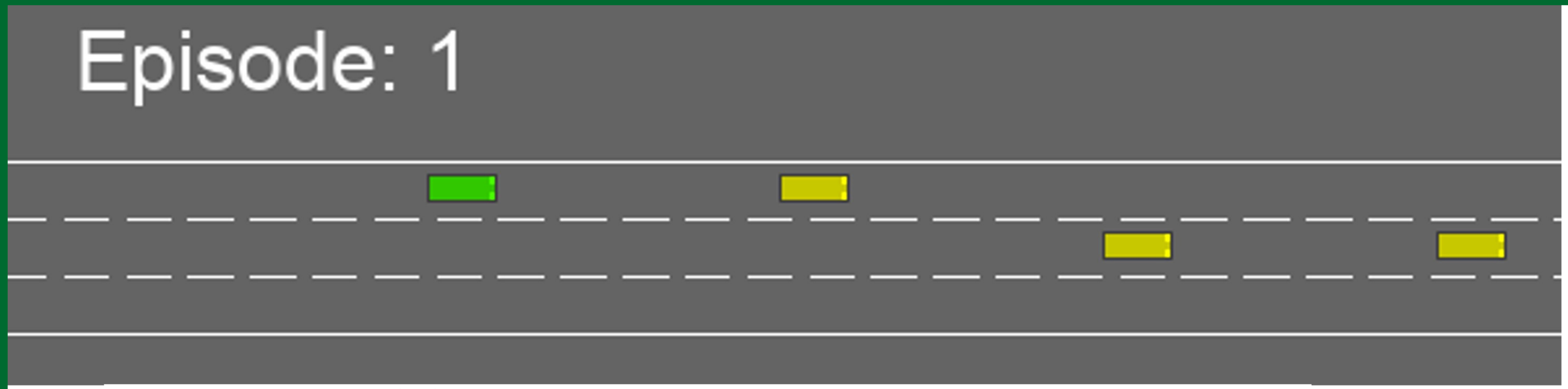
Cost	# Programs	Bank
2	3	$\{x_1, x_2, x_3\}$
3	3	$\{\sqrt{x_1}, \sqrt{x_2}, \sqrt{x_3}\}$
4	3	$\{\sqrt{\sqrt{x_1}}, \sqrt{\sqrt{x_2}}, \sqrt{\sqrt{x_3}}\}$
5	12	$\{\sqrt{\sqrt{\sqrt{x_1}}}, \cdots, x_1 + x_1, x_1 + x_2, x_1 + x_3, \cdots\}$
6	48	$\{\sqrt{\sqrt{\sqrt{x_1}}}, \cdots, \sqrt{x_1 + x_2}, \cdots, x_1 + \sqrt{x_1}, \cdots, \sqrt{x_1} + x_1\}$
7	93	$\{\cdots\}$
8	354	$\{\cdots\}$
9	3200	$\{\cdots\}$
10	...	$\{\cdots, \sqrt{\sqrt{x_1 + x_2 + x_3}}, \cdots\}$

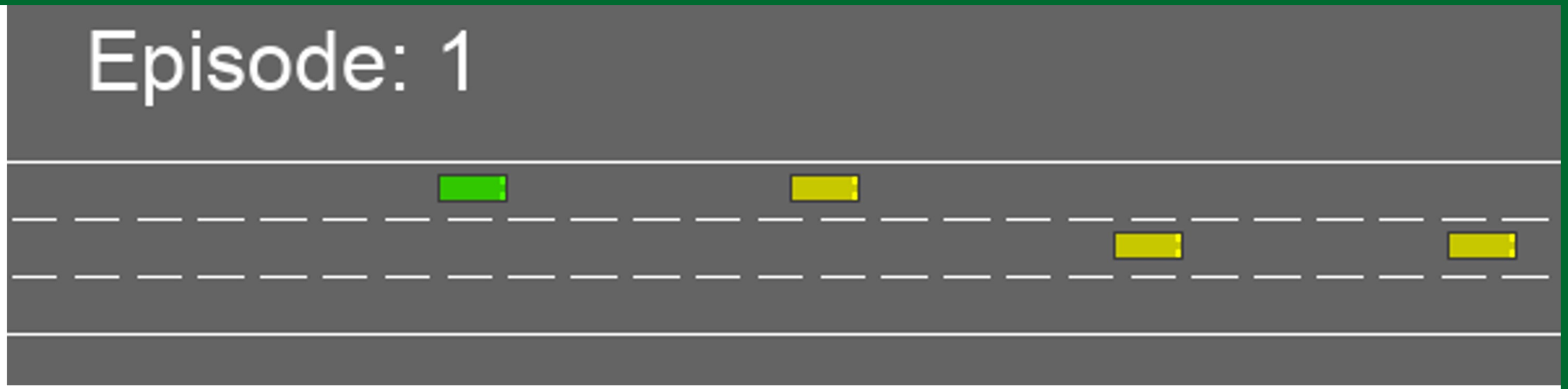
How does POLIS use BUS and Probe?



Experimental details








```python
def initial(o):
    if (o[5] == o[1] and o[5]-o[1] > 200) or\
            (o[9] == o[1] and o[9]-o[1] > 200):
        action = 4
    elif (o[5] == o[1] and o[5]-o[1] <= 200) or\
            (o[9] == o[1] and o[9]-o[1] <= 200):
        if o[1] == 4:
            if o[9] < 4:
                action = 2
            else:
                action = 0
        else:
            action = 0
    else:
        action = 3
    return action
```

```python
def improved(o):
    if o[1] and o[3]:

        action = 4
    elif (o[5] == o[1] and o[5]-o[1] <= 200) or\
            (o[9] == o[1] and o[9]-o[1] <= 200):
        if o[1] == o[5]:
            if o[1] < 7.93:
                action = 2
            else:
                action = 0
        else:
            action = 2
    else:
        action = 1
    return action
```

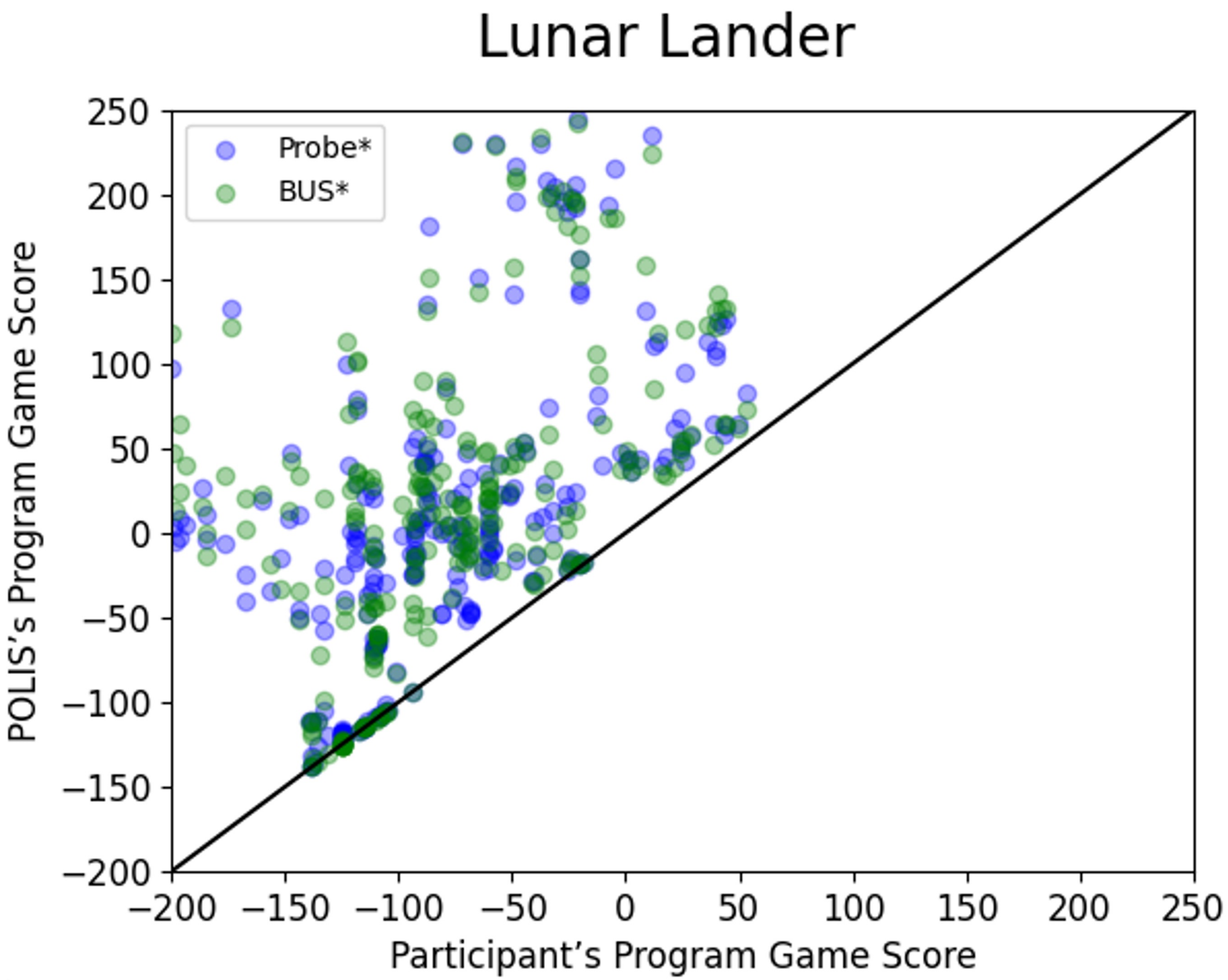
Instead of left, go to the right

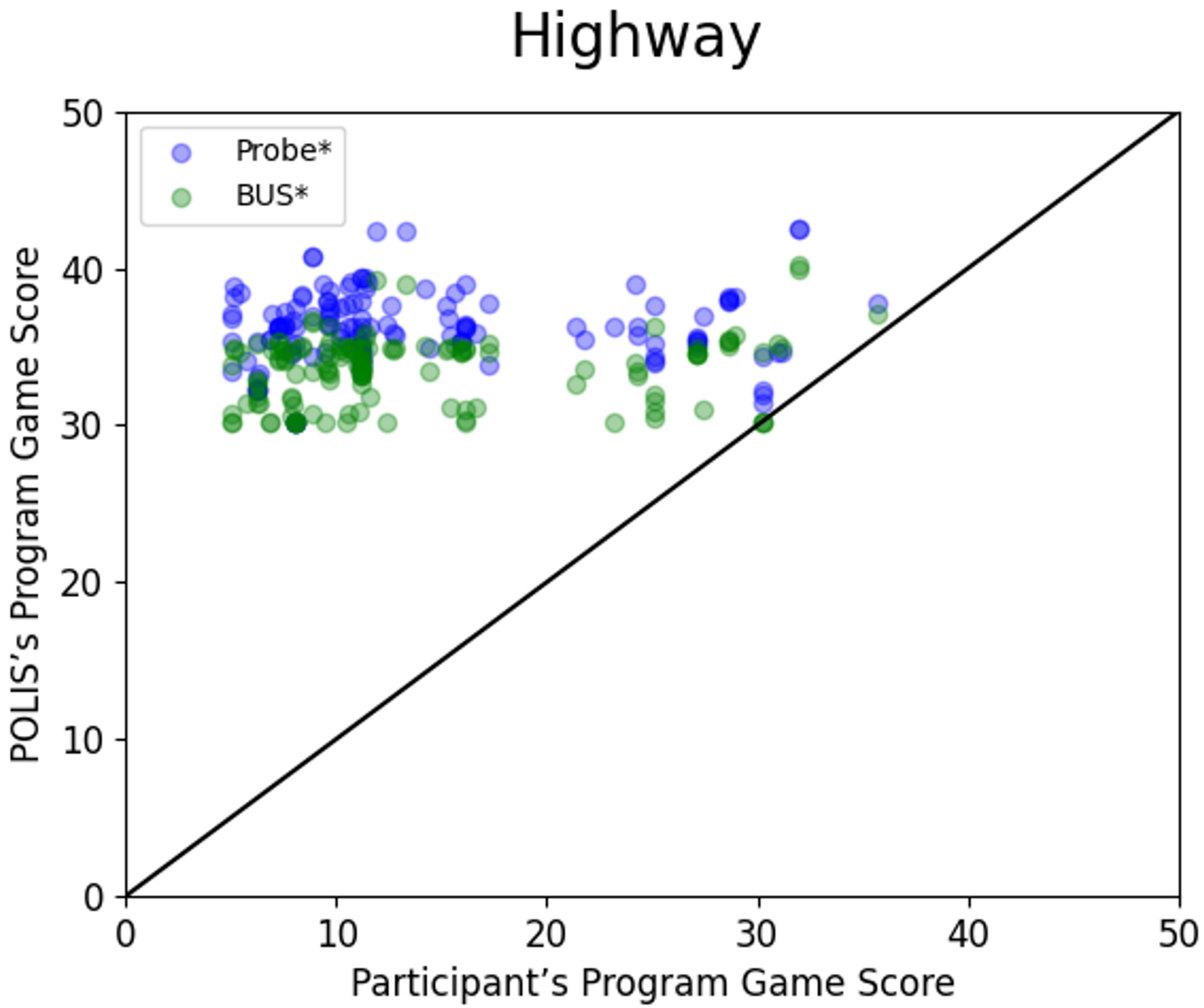
Do nothing instead of increasing speed

Score ~6.8

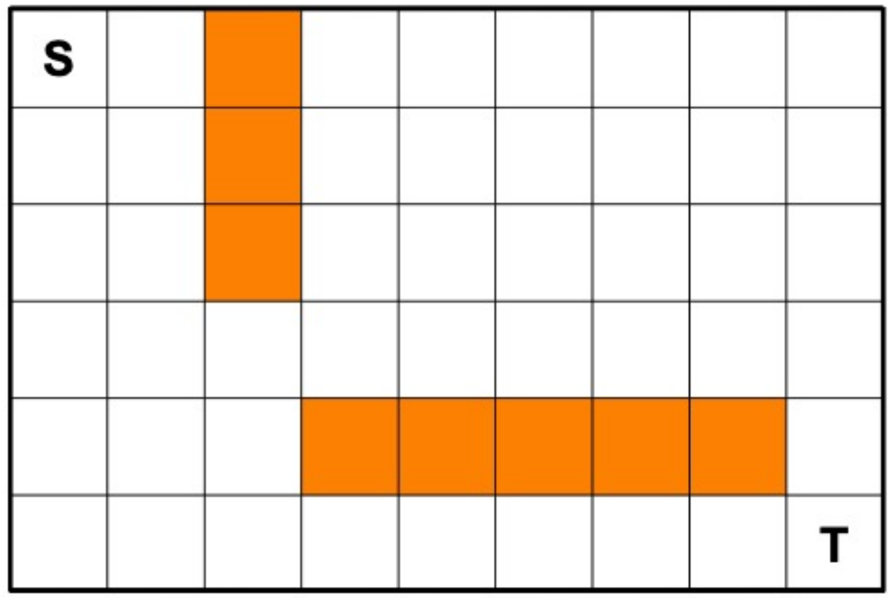
Score ~39

Computational results: POLIS results

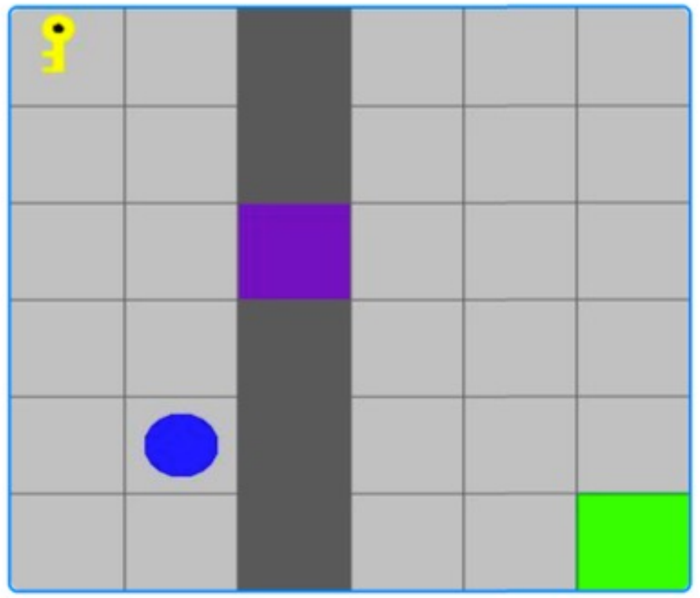




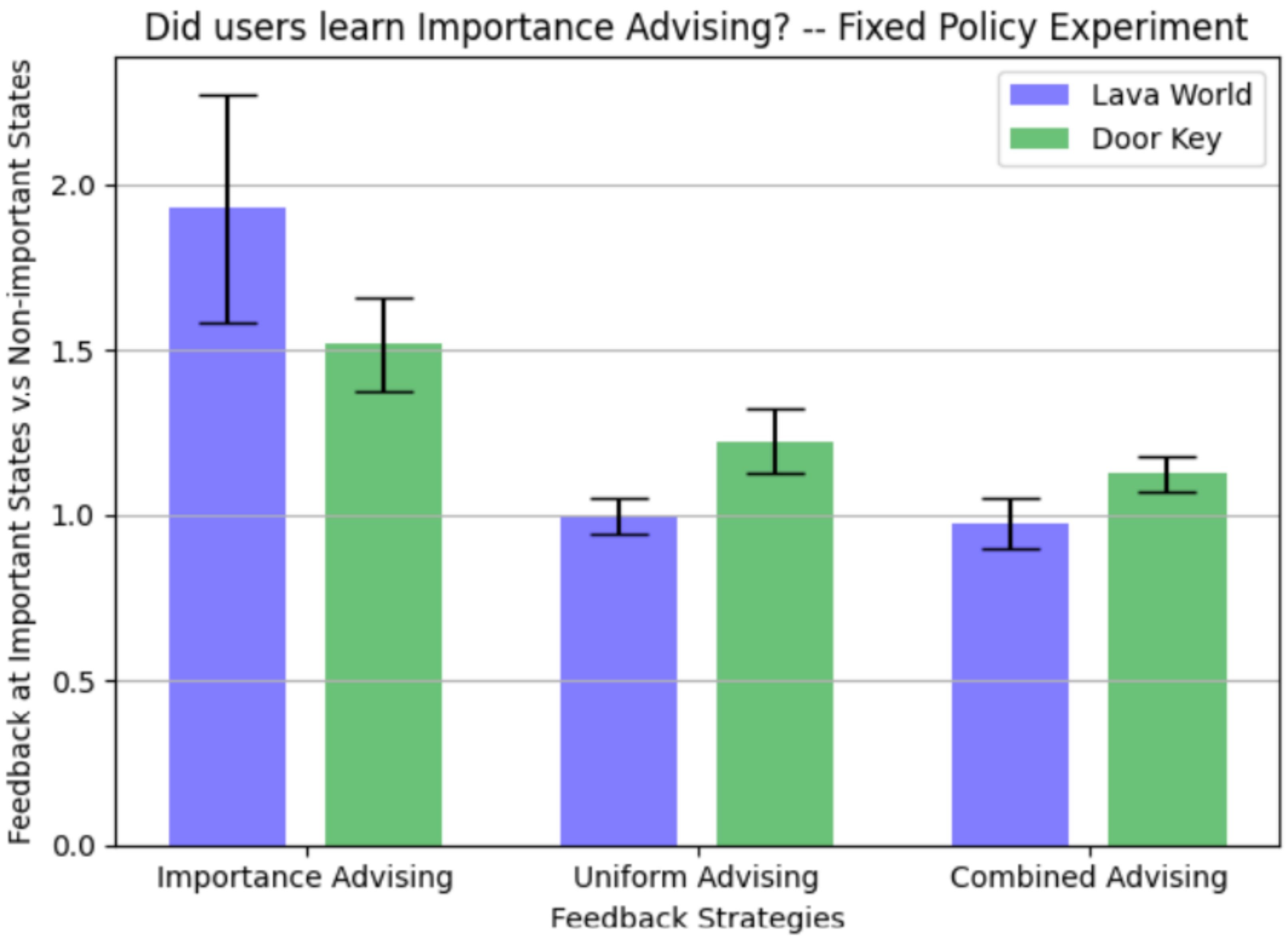
Research Question 1:
Can we teach people how to be better teachers?



(a) Lava World



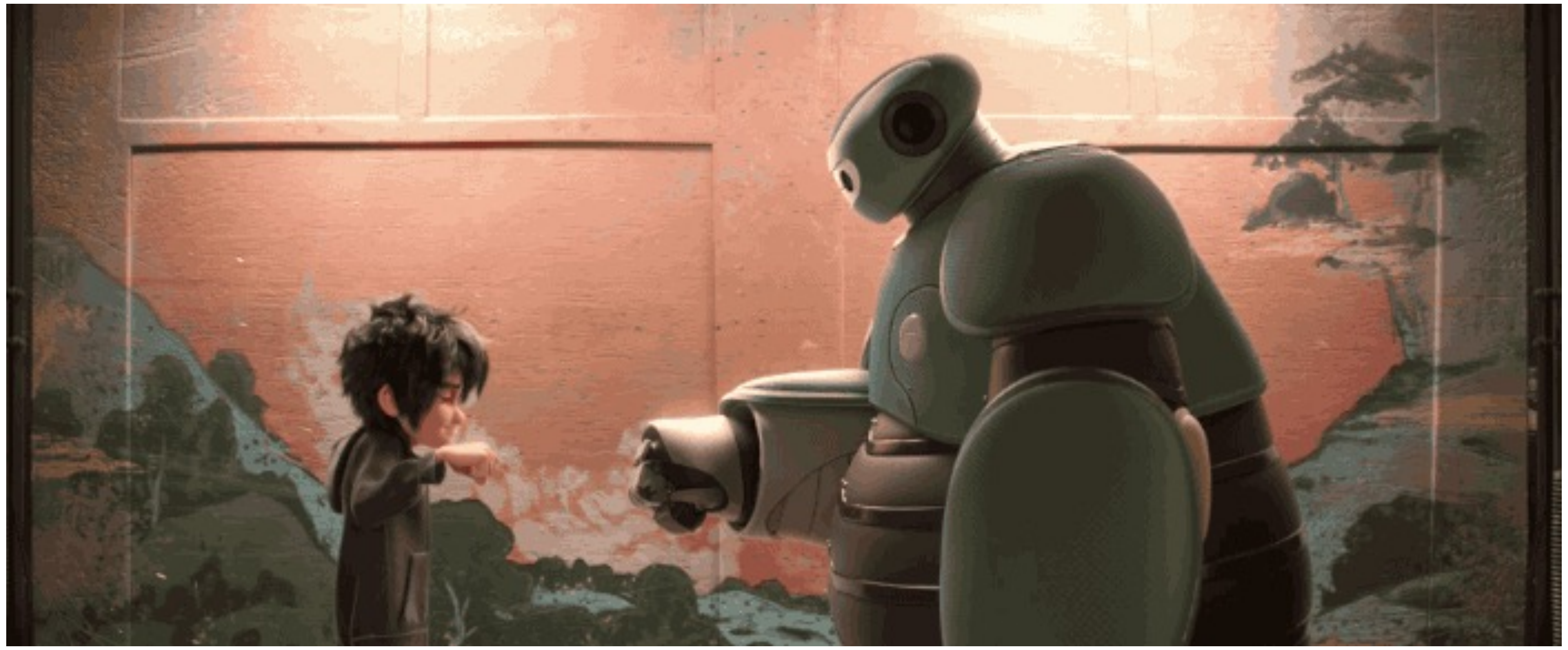
(b) Door Key



Muslimani et al., 2021

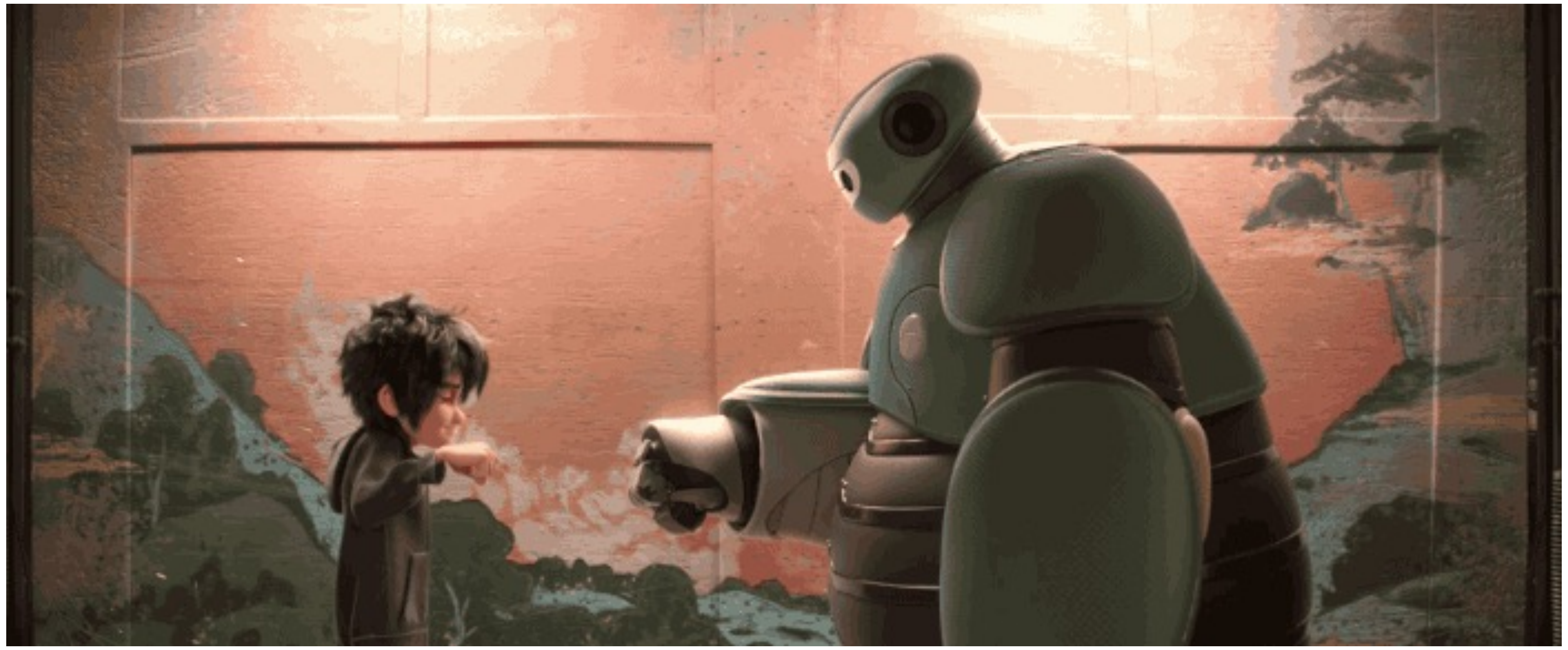
Research Question 2:
Can we adapt our algorithms to better learn from human teachers?

- Figure out what human feedback means?



Research Question 2:
Can we adapt our algorithms to better learn from human teachers?

- Figure out what human feedback means?



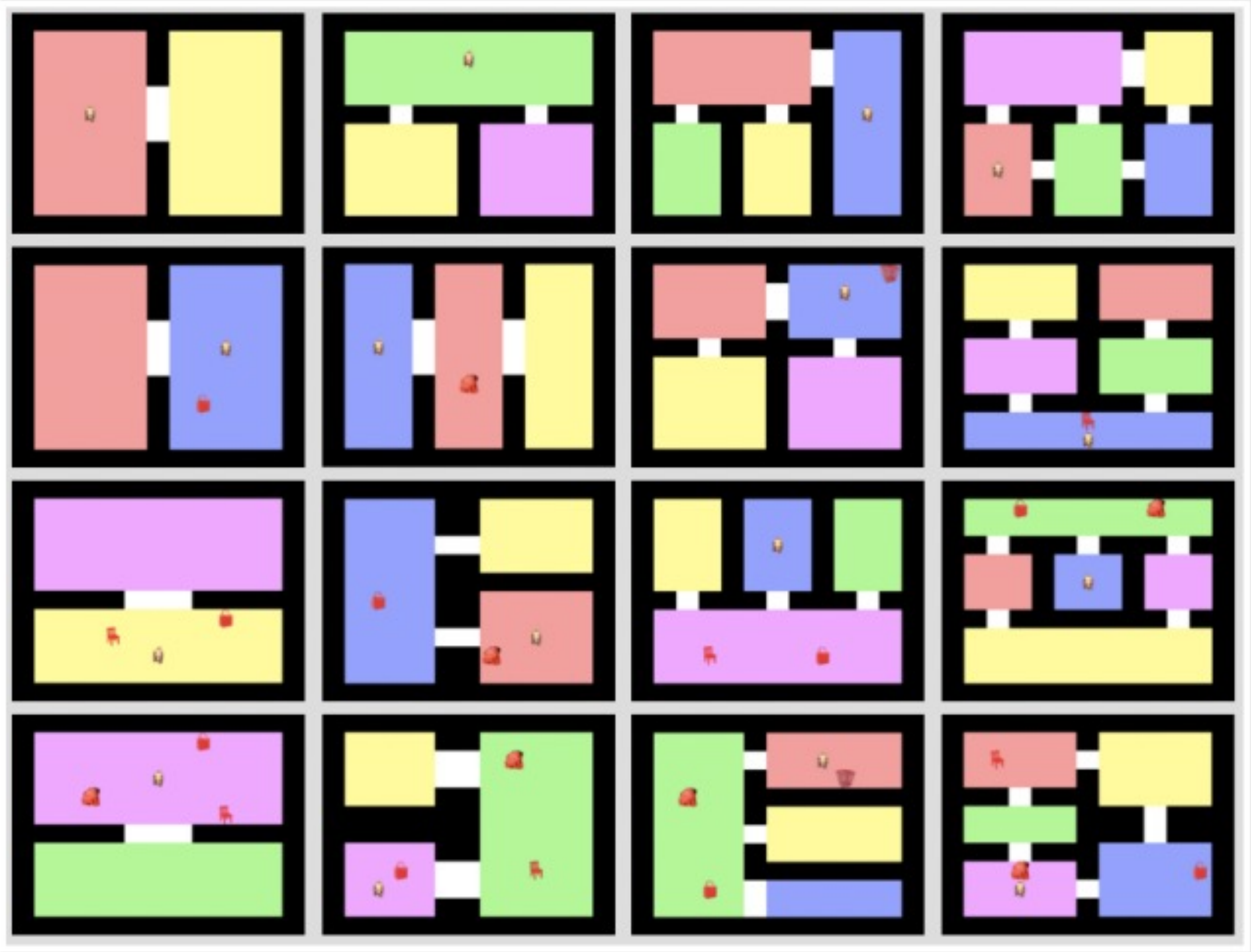
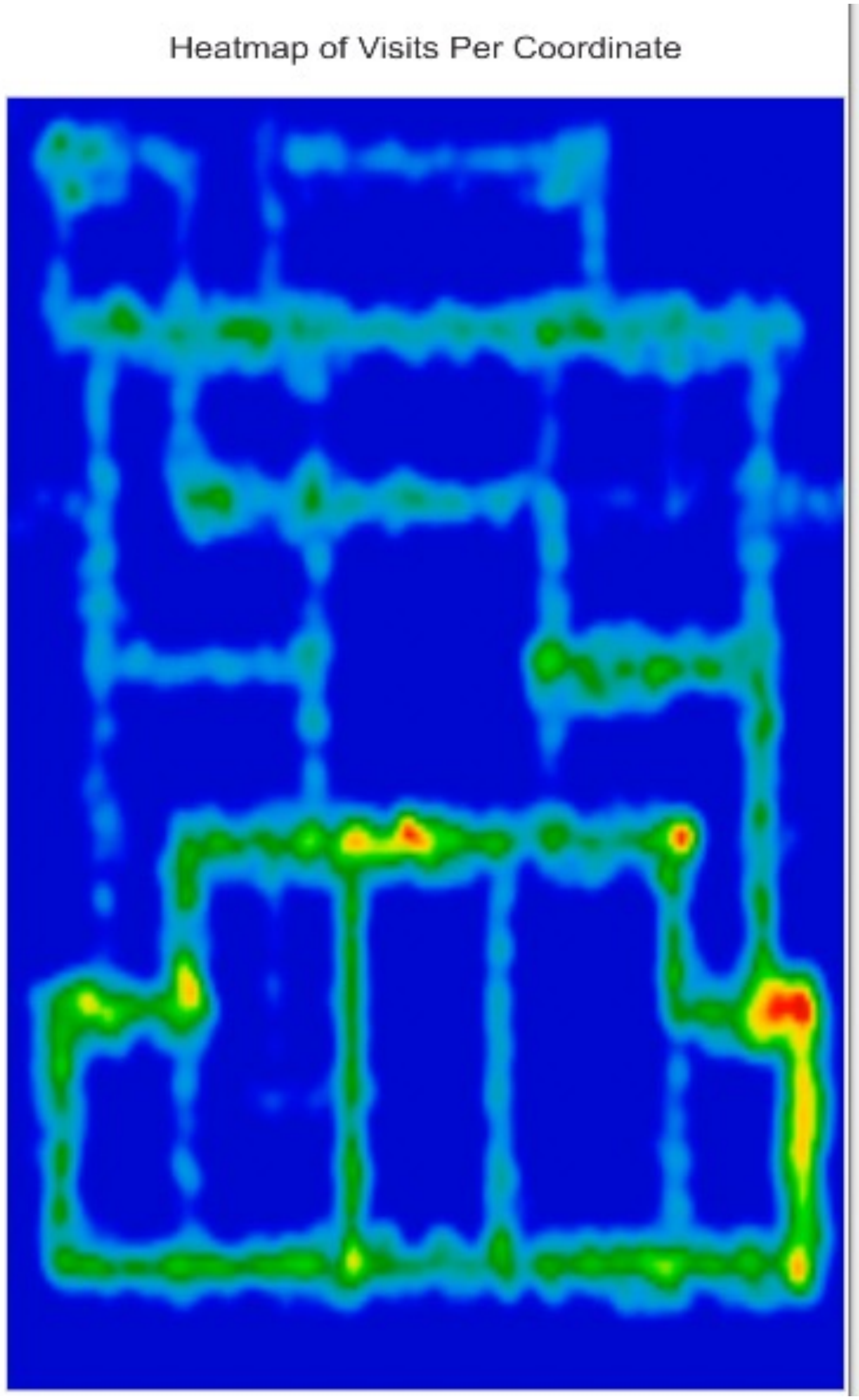


Figure 2: The library of 16 environments is organized by the number of rooms and objects. There is a command list for each environment.

Peng et al., 2018

Research Question 3: Will Explainability Help?

- Explanations can help people select better agent and/or better anticipate agent's actions

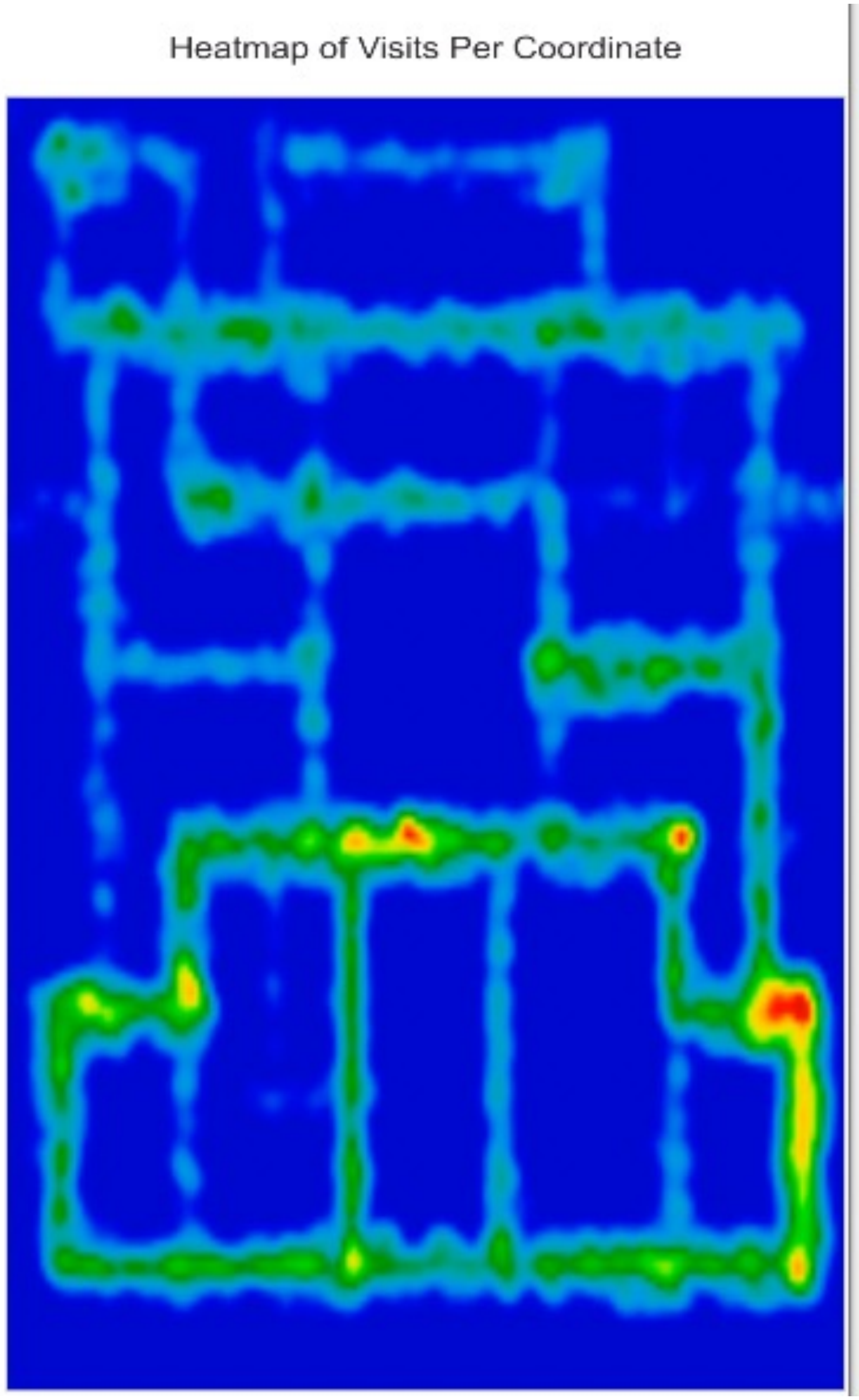


Davis-Pearson et al., under submission

Research Question 3: Will Explainability Help?

- Explanations can help people select better agent and/or better anticipate agent’s actions

- Knowing what the agents knows should let teacher better target how they help
  → seems obvious…

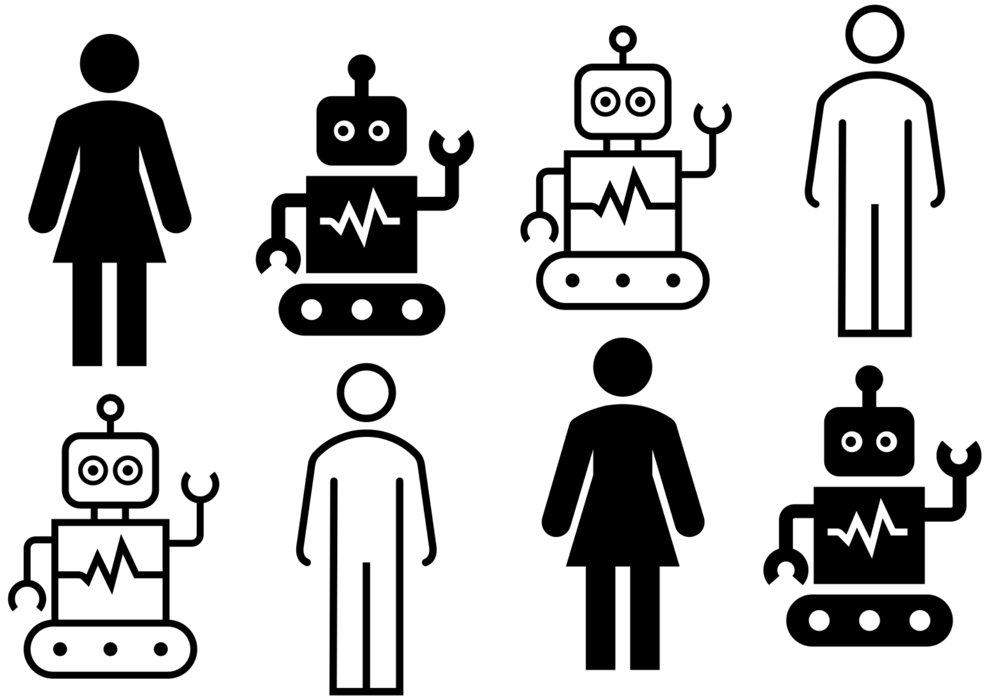


Davis-Pearson et al., under submission

Research Question 4:
When is one type of help preferred?

- Teacher competence?
- Student capabilities?
- Speed of simulation?
- ...

Multi-agent, Multi-human Teaming



cogment™

The **first platform** to allow the design, training, and deployment of complex **intelligence ecosystems**, mixing **humans and artificial agents** of various kinds

It orchestrates heterogeneous ML & non-ML agents with real-time human interaction.

air

AI Redefined

Human Input Parsing Platform for Openai Gym

In a web browser, human subjects can interact with Atari games, MuJuCo robots, etc.

- Give demonstrations
- Provide feedback
- Identify errors

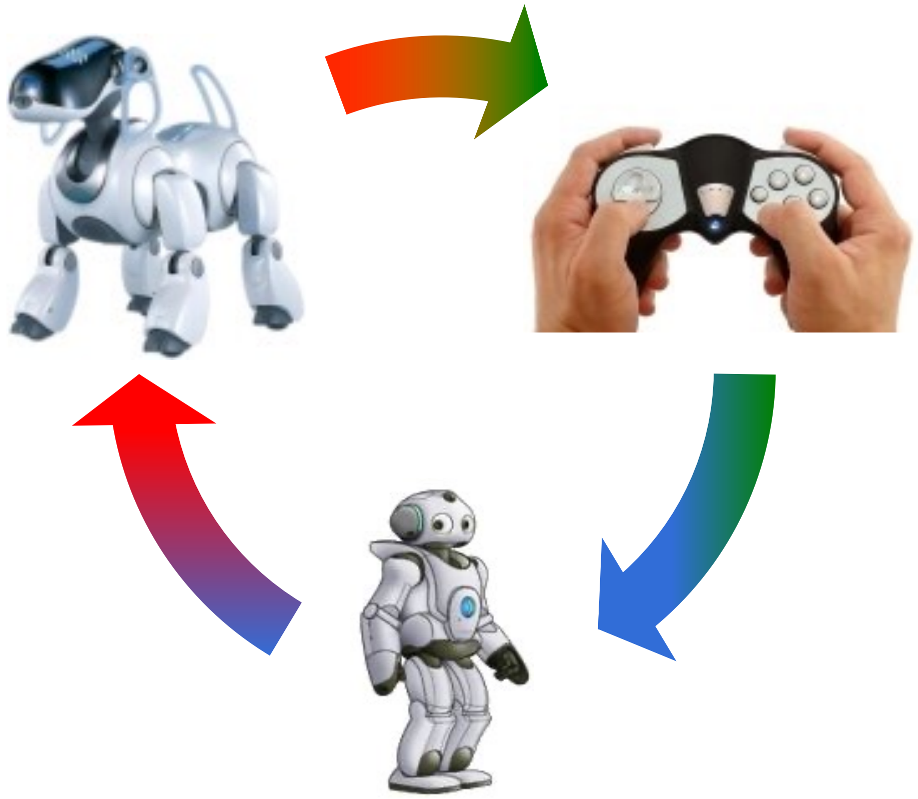
Enable scaling up & out of HitL RL

- Built-in AWS support
- Integrate with MTurk

HIPPO GYM

hippogym.irll.ca

Conclusion: Many more questions!



We should cheat whenever possible

Lots of room for improvement

- Learning from agents/data
- Learning from humans
- Teaching humans

The Intelligent Robot Learning Laboratory

http://irll.ca
http://cogment.ai/