

University of New South Wales
School of Computer Science and Engineering

Explainable Robotics Systems in Reinforcement Learning Scenarios

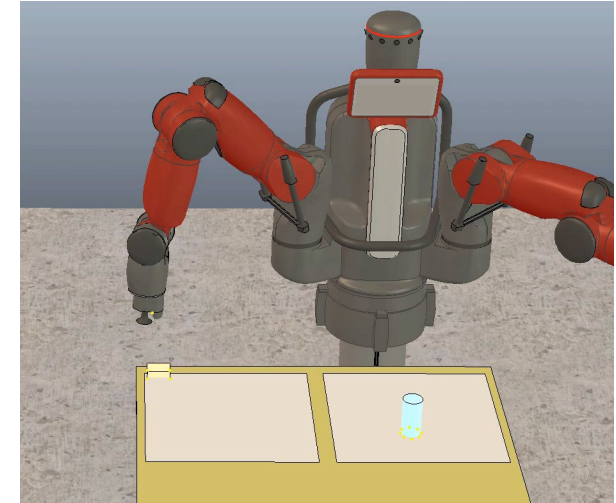
Dr. Francisco Cruz

f.cruz@unsw.edu.au

<https://www.franciscocruz.org/>

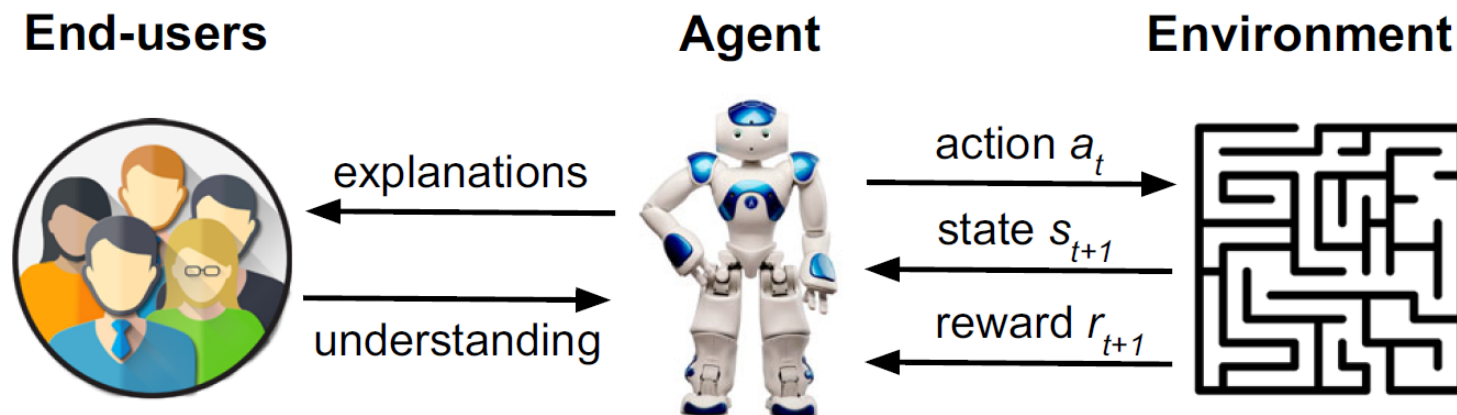
Outline

- Introduction and reinforcement learning background.
- Explainable artificial intelligence.
- Memory-based method.
- Learning-based and introspection-based methods.
- Non-episodic and continuous domains.
- Evaluation of resources.
- Evaluation by non-expert end-users.
- Conclusions and future work.



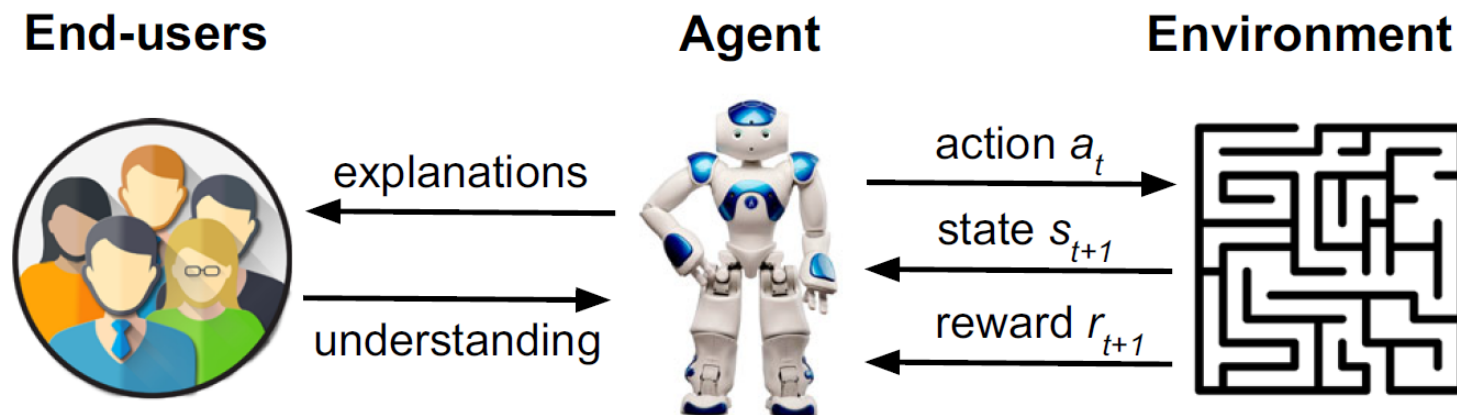
Reinforcement Learning

- Cognitive agents are able to autonomously learn new tasks by interacting with the environment.
- Reinforcement learning (RL) has been shown a successful method for agents to acquire new skills by exploring their environment.
- In human–robot environments, it is crucial that end-users may correctly understand their robotic team-partners.



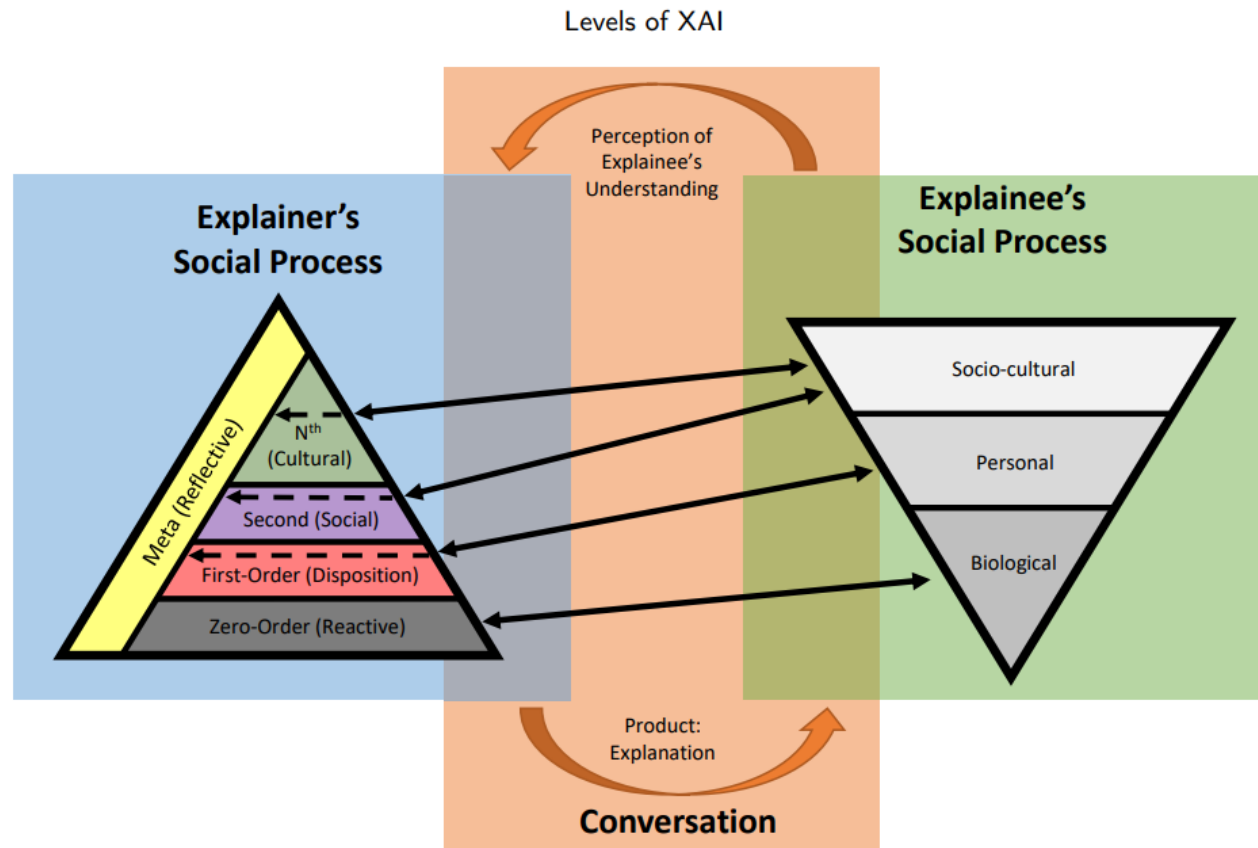
Explainable Robotic Systems

- A robot can provide featured-based or goal-driven explanations.
- **Not acceptable.** I choose action *left* because it maximizes future collected reward OR I choose action *right* because it is the next one following the optimal policy.
- Using the probability of success is possible to create human-like explanations.



Explainable Artificial Intelligence¹

- AI explanations aligned to human communication.



¹ Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., & Cruz, F. "Levels of Explainable Artificial Intelligence for Human-aligned Conversational Explanations". *Artificial Intelligence*, 299, 103525. 2021.

Memory-based Method²

- From a non-expert end-user perspective, most relevant questions: 'why?' and 'why not?'. For instance
 - Why did you step forward in the last movement?
 - Why did you not turn to the right in this situation?
- We propose MXRL to compute P_s and N_t using an episodic memory.
- We implement a list of state-action pairs (TList).

² Cruz, F., Dazeley, R., Vamplew, P. "Memory-based explainable reinforcement learning". In *Proceedings of the 32nd Australasian Joint Conference on Artificial Intelligence (AI2019)*, pp. 66-67, Adelaide, Australia, 2019.

Memory-based Method

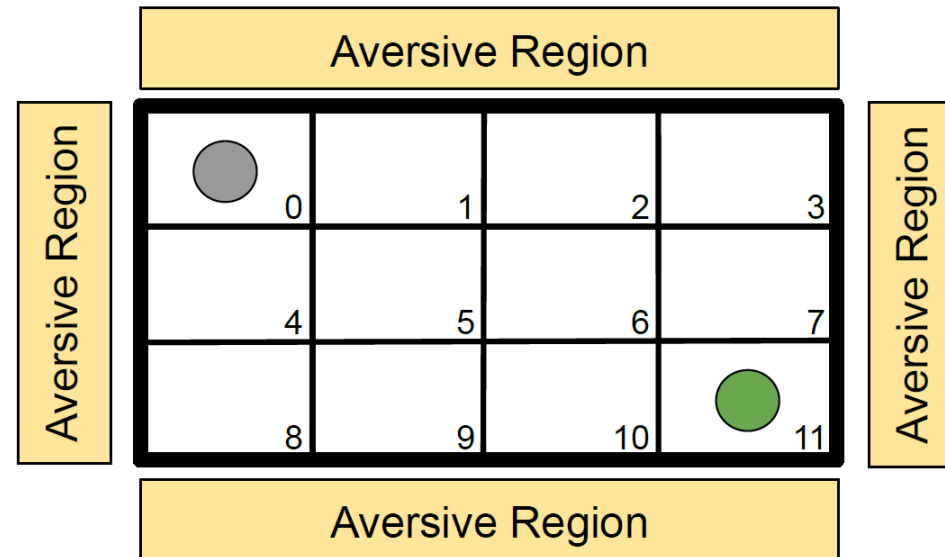
- MXRL algorithm.

Algorithm 1 Memory-based explainable reinforcement learning approach with the on-policy method SARSA to compute the probability of success and the number of transitions to the goal state.

```
1: Initialize  $Q(s, a), T_t, T_s, P_s, N_t$ 
2: for each episode do
3:   Initialize  $T_{List}[]$ 
4:   Choose an action using  $a_t \leftarrow \text{SELECTACTION}(s_t)$ 
5:   repeat
6:     Take action  $a_t$ 
7:     Save state-action transition  $T_{List}.add(s, a)$ 
8:      $T_t[s][a] \leftarrow T_t[s][a] + 1$ 
9:     Observe reward  $r_{t+1}$  and next state  $s_{t+1}$ 
10:    Choose next action  $a_{t+1}$  using softmax action selection method
11:     $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$ 
12:     $s_t \leftarrow s_{t+1}; a_t \leftarrow a_{t+1}$ 
13:  until  $s$  is terminal (goal or aversive state)
14:  if  $s$  is goal state then
15:    for each  $s, a \in T_{List}$  do
16:       $T_s[s][a] \leftarrow T_s[s][a] + 1$ 
17:    end for
18:  end if
19:  Compute  $P_s \leftarrow T_s/T_t$ 
20:  Compute  $N_t$  for each  $s \in T_{List}$  as  $\text{pos}(s, T_{List}) + 1$ 
21: end for
```

Memory-based Method²

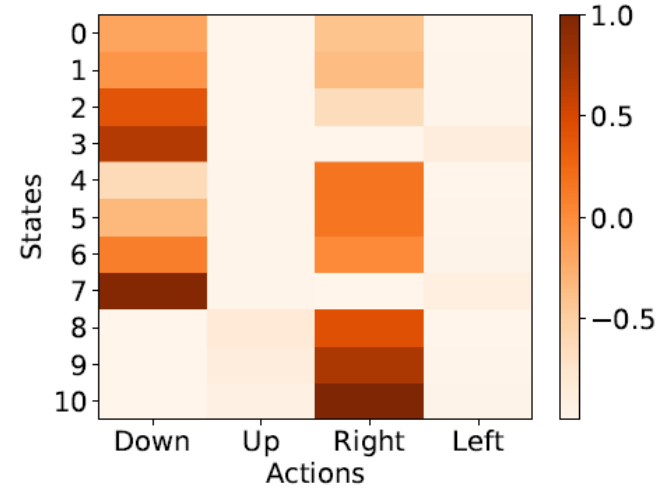
- Experimental setup: A 3x4 grid world scenario.
- Four allowed actions in this scenario: down, up, right, and left.



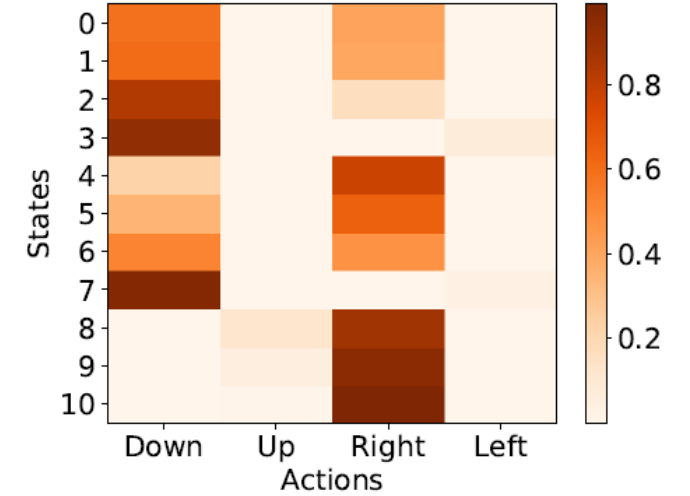
² Cruz, F., Dazeley, R., Vamplew, P. "Memory-based explainable reinforcement learning". In *Proceedings of the 32nd Australasian Joint Conference on Artificial Intelligence (AI2019)*, pp. 66-67, Adelaide, Australia, 2019.

Memory-based Method

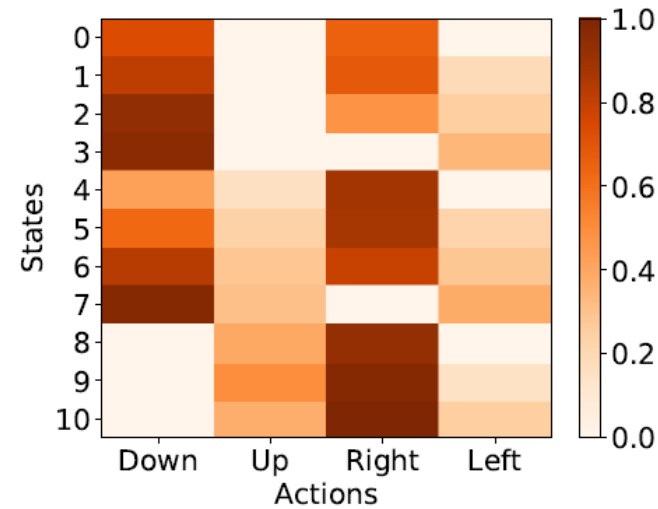
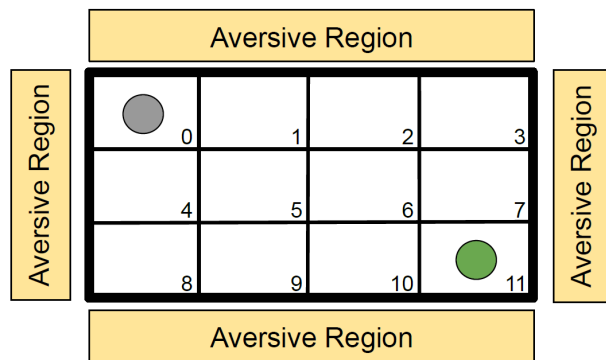
- Experimental results.



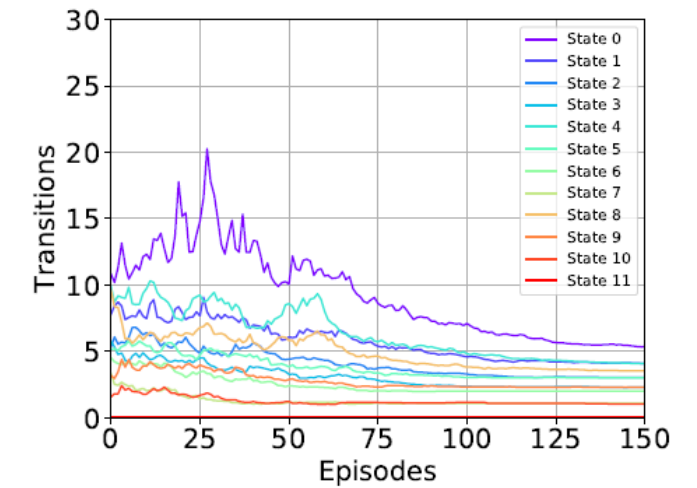
(a) Q-values.



(b) Softmax values.

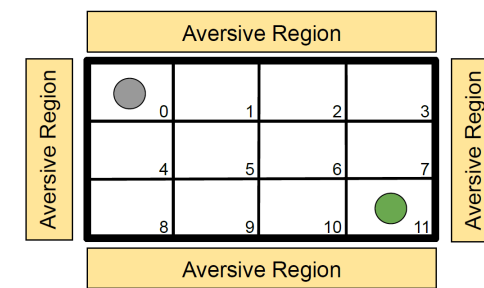


(c) Probability of success.



(d) Number of transitions.

Memory-based Method²



- In this context, one possible question to the artificial agent is:
 - *Why did you choose action down when in state 0?*

- Using Q-values to explain this is pointless for a non-expert user.

$$Q(s=0; a=\text{down}) = -0.181$$

$$Q(s=0; a=\text{up}) = -0.998$$

$$Q(s=0; a=\text{right}) = -0.411$$

$$Q(s=0; a=\text{left}) = -0.998$$

- If we use P_s , the agent may answer the end-user: *I chose to go down because that has a 73.6% probability of successfully reaching the goal.*

$$P_s(s=0; a=\text{down}) = 0.736$$

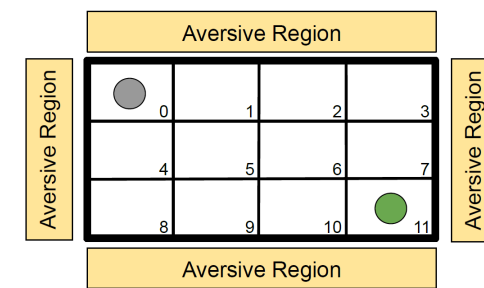
$$P_s(s=0; a=\text{up}) = 0$$

$$P_s(s=0; a=\text{right}) = 0.656$$

$$P_s(s=0; a=\text{left}) = 0$$

² Cruz, F., Dazeley, R., Vamplew, P. "Memory-based explainable reinforcement learning". In *Proceedings of the 32nd Australasian Joint Conference on Artificial Intelligence (AI2019)*, pp. 66-67, Adelaide, Australia, 2019.

Memory-based Method²



- Another possible question to the agent is:
 - *Why did you not choose to go left when in state 0?*

- Using Q-values to explain this is pointless for a non-expert user.

$$Q(s=0; a=\text{down}) = -0.181$$

$$Q(s=0; a=\text{up}) = -0.998$$

$$Q(s=0; a=\text{right}) = -0.411$$

$$Q(s=0; a=\text{left}) = -0.998$$

- If we use P_s , one possible answer is: *I did not choose left because that has a zero probability of success, whereas by choosing down has a 73.6% probability of success, which was higher than other actions.*

$$P_s(s=0; a=\text{down}) = 0.736$$

$$P_s(s=0; a=\text{up}) = 0$$








$$P_s(s=0; a=\text{right}) = 0.656$$










$$P_s(s=0; a=\text{left}) = 0$$

² Cruz, F., Dazeley, R., Vamplew, P. "Memory-based explainable reinforcement learning". In *Proceedings of the 32nd Australasian Joint Conference on Artificial Intelligence (AI2019)*, pp. 66-67, Adelaide, Australia, 2019.

Memory-based in a Hierarchical Scenario³

- Spaceship problem:

	Goal first high-level task		Agent
	Goal second high-level task		Black Holes
	Goal third high-level task		Shield
			WormHole

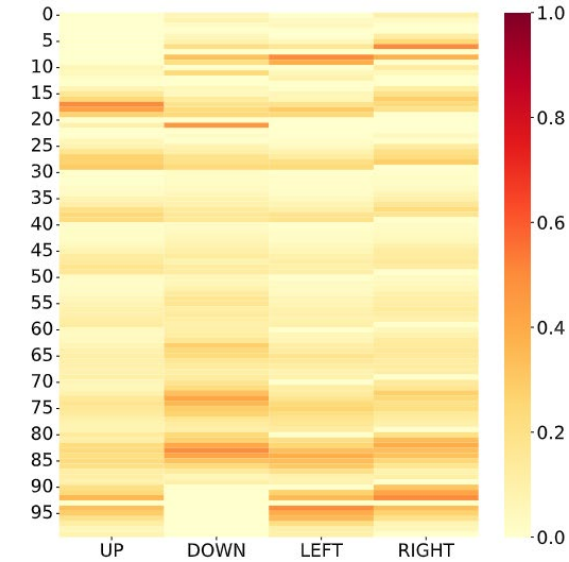
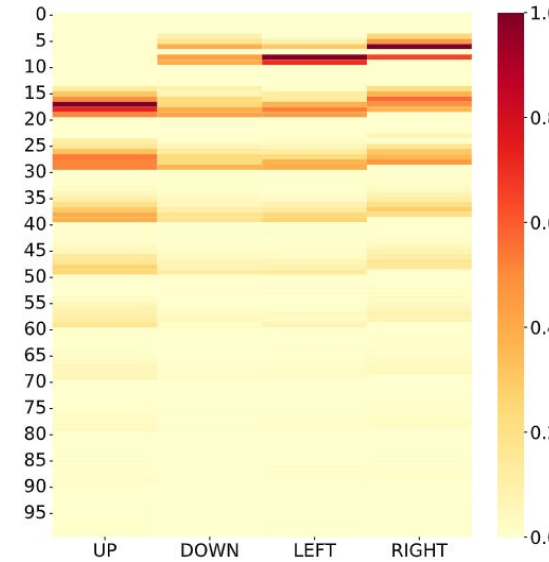
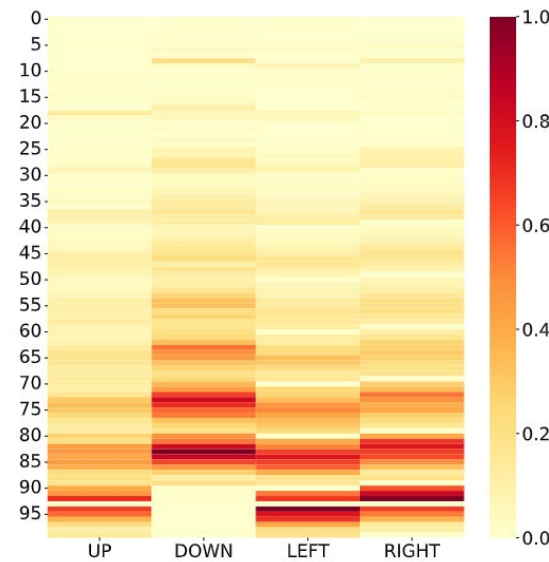
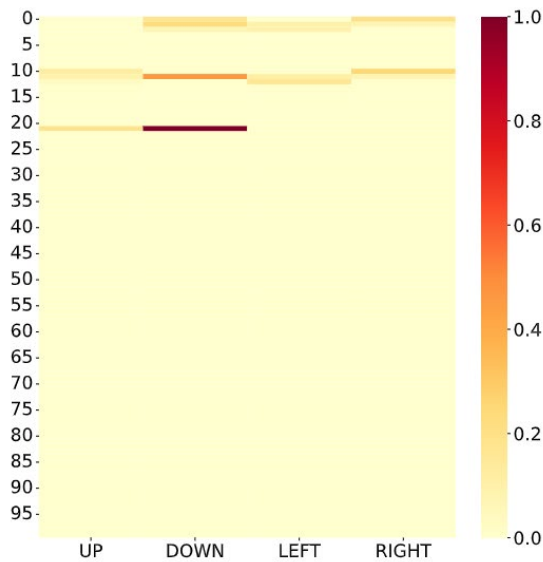
0	1	2	3	4	5	6	7	8	9
									
10	11	12	13	14	15	16	17	18	19
									
20	21	22	23	24	25	26	27	28	29
									
30	31	32	33	34	35	36	37	38	39
									
40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69
70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	87	88	89
90	91	92	93	94	95	96	97	98	99
									

³ Muñoz, H., Portugal, E., Ayala A., Fernandes, B., Cruz, F. "Explaining Agent's Decision-making in a Hierarchical Reinforcement Learning Scenario". Accepted at the IEEE 41st International Conference of the Chilean Computer Society (SCCC 2022). In press.

Memory-based Hierarchical Method³

- Spaceship problem:

0	1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18	19
20	21	22	23	24	25	26	27	28	29
30	31	32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49
50	51	52	53	54	55	56	57	58	59
60	61	62	63	64	65	66	67	68	69
70	71	72	73	74	75	76	77	78	79
80	81	82	83	84	85	86	87	88	89
90	91	92	93	94	95	96	97	98	99



High-level tasks

General task

³ Muñoz, H., Portugal, E., Ayala A., Fernandes, B., Cruz, F. "Explaining Agent's Decision-making in a Hierarchical Reinforcement Learning Scenario". Accepted at the IEEE 41st International Conference of the Chilean Computer Society (SCCC 2022). In press.

Learning- and Introspection-based Methods⁴

- Goal-driven explanations.

Algorithm 2 Explainable reinforcement learning approach to compute the probability of success using the learning-based approach.

```
1: Initialize  $Q(s, a), \mathbb{P}(s_t, a_t)$ 
2: for each episode do
3:   Initialize  $s_t$ 
4:   Choose an action  $a_t$  from  $s_t$ 
5:   repeat
6:     Take action  $a_t$ 
7:     Observe reward  $r_{t+1}$  and next state  $s_{t+1}$ 
8:     Choose next action  $a_{t+1}$  using softmax action
       selection method
9:      $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1})$ 
        $- Q(s_t, a_t)]$ 
10:     $\mathbb{P}(s_t, a_t) \leftarrow \mathbb{P}(s_t, a_t) + \alpha[\varphi_{t+1} + \mathbb{P}(s_{t+1}, a_{t+1})$ 
        $- \mathbb{P}(s_t, a_t)]$ 
11:     $s_t \leftarrow s_{t+1}; a_t \leftarrow a_{t+1}$ 
12:   until  $s_t$  is terminal (goal or aversive state)
13: end for
```

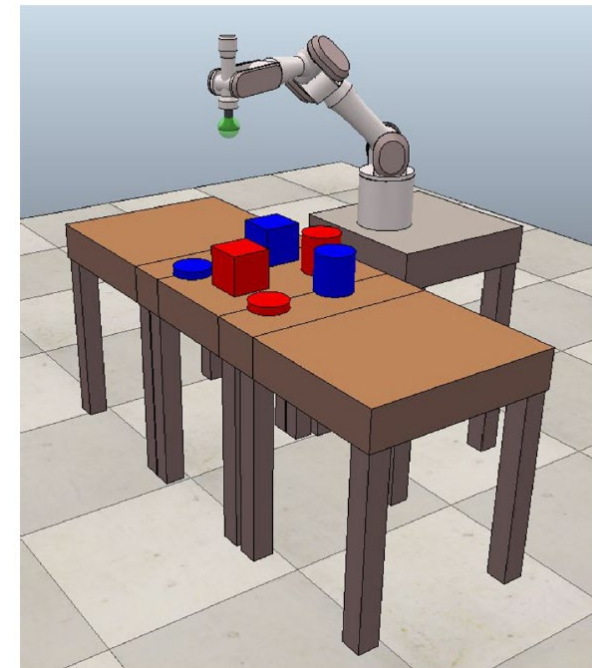
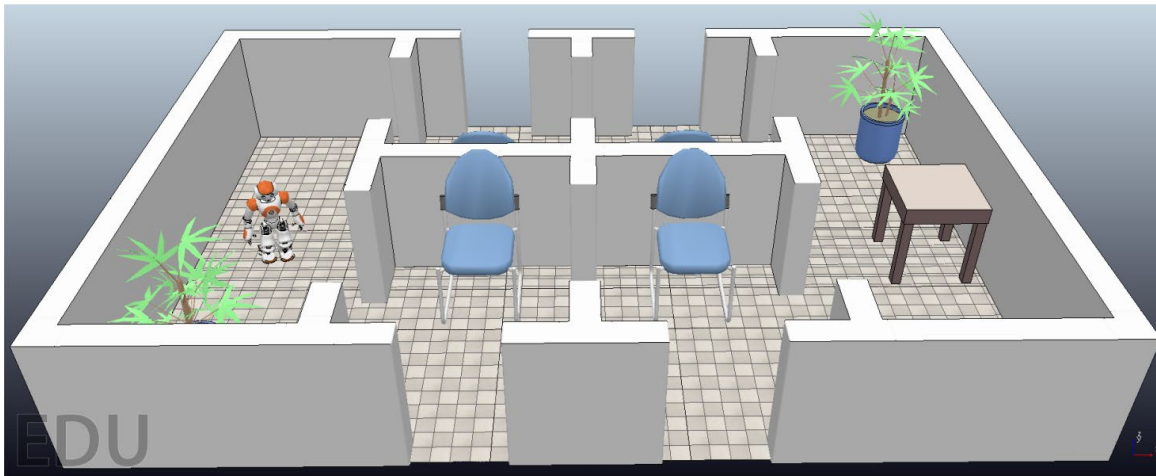
Algorithm 3 Explainable reinforcement learning approach to compute the probability of success using the introspection-based approach.

```
1: Initialize  $Q(s, a), \hat{P}_s$ 
2: for each episode do
3:   Initialize  $s_t$ 
4:   Choose an action  $a_t$  from  $s_t$ 
5:   repeat
6:     Take action  $a_t$ 
7:     Observe reward  $r_{t+1}$  and next state  $s_{t+1}$ 
8:     Choose next action  $a_{t+1}$  using softmax action
       selection method
9:      $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1})$ 
        $- Q(s_t, a_t)]$ 
10:     $s_t \leftarrow s_{t+1}; a_t \leftarrow a_{t+1}$ 
11:   until  $s_t$  is terminal (goal or aversive state)
12:    $\hat{P}_s \approx \left[ (1 - \sigma) \cdot \left( \frac{1}{2} \cdot \log_{10} \frac{Q(s_t, a_t)}{R^T} + 1 \right) \right]_{\hat{P}_s \leq 1}$ 
        $_{\hat{P}_s \geq 0}$ 
13: end for
```

⁴ Cruz, F., Dazeley, R., Vamplew, P., Moreira, I. "Explainable Robotic Systems: Understanding Goal-driven Actions in a Reinforcement Learning Scenario". *Neural Computing and Applications*. Springer. 2021.

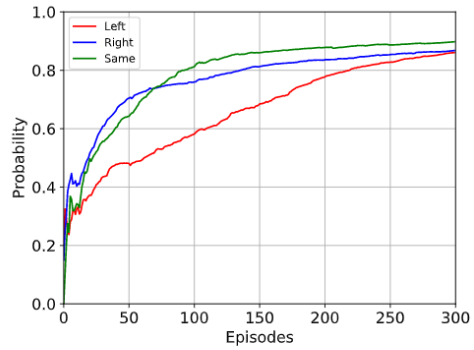
Learning- and Introspection-based Methods⁴

- Deterministic and stochastic navigation task.
- Continuous sorting object task.
- Real-world scenario.

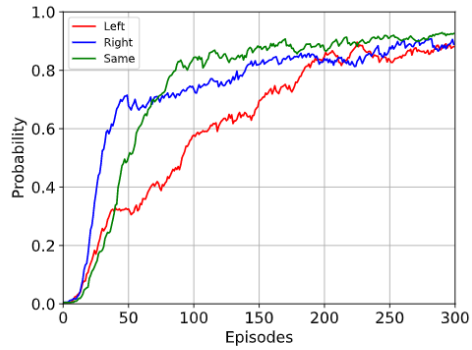


⁴ Cruz, F., Dazeley, R., Vamplew, P., Moreira, I. "Explainable Robotic Systems: Understanding Goal-driven Actions in a Reinforcement Learning Scenario". *Neural Computing and Applications*. Springer. 2021.

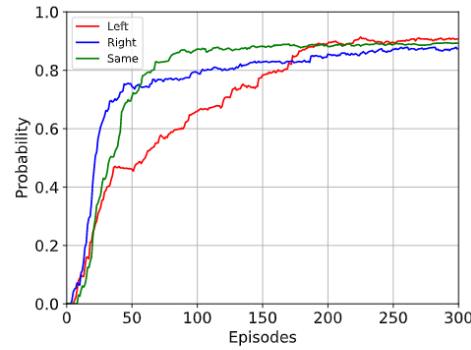
Learning- and Introspection-based Methods⁴



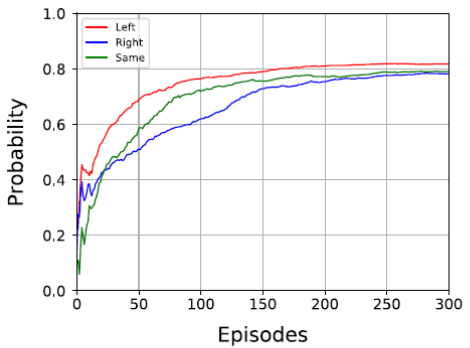
(d) Memory-based approach.



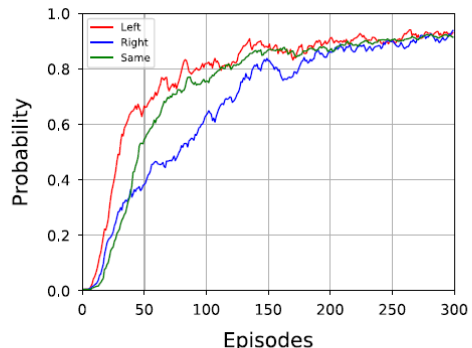
(e) Learning-based approach.



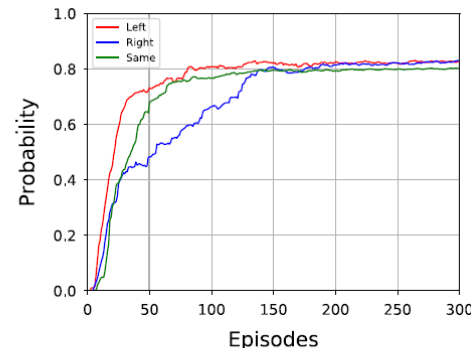
(f) Introspection-based approach.



(d) Memory-based approach.



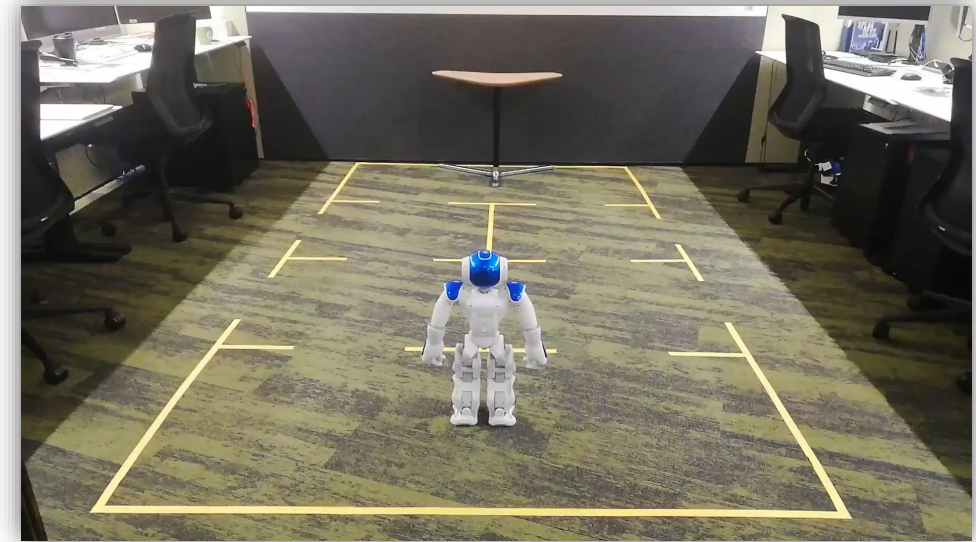
(e) Learning-based approach.



(f) Introspection-based approach.



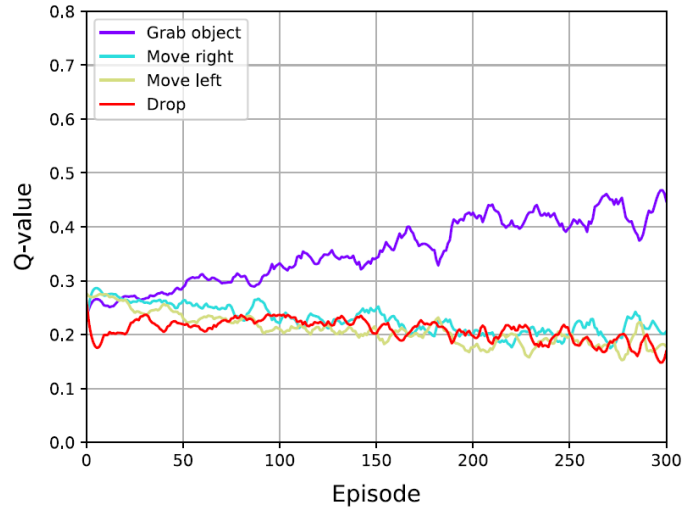
Deterministic and stochastic tasks.



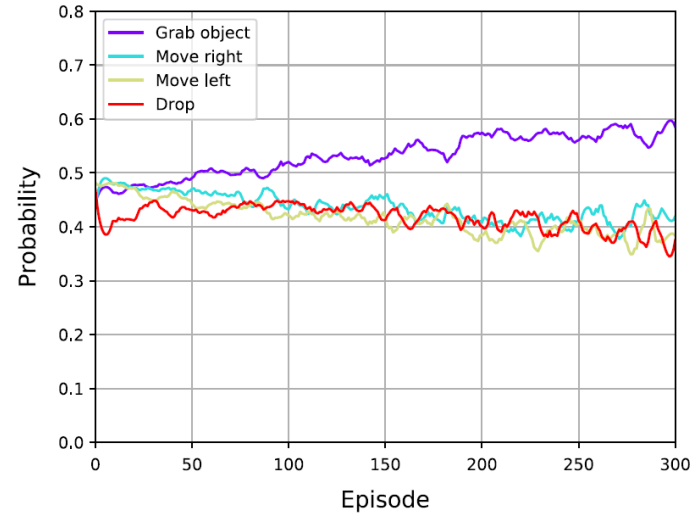
Explanation. I chose to go left because that has a 87.6% probability of reaching the goal successfully

⁴ Cruz, F., Dazeley, R., Vamplew, P., Moreira, I. "Explainable Robotic Systems: Understanding Goal-driven Actions in a Reinforcement Learning Scenario". *Neural Computing and Applications*. Springer. 2021.

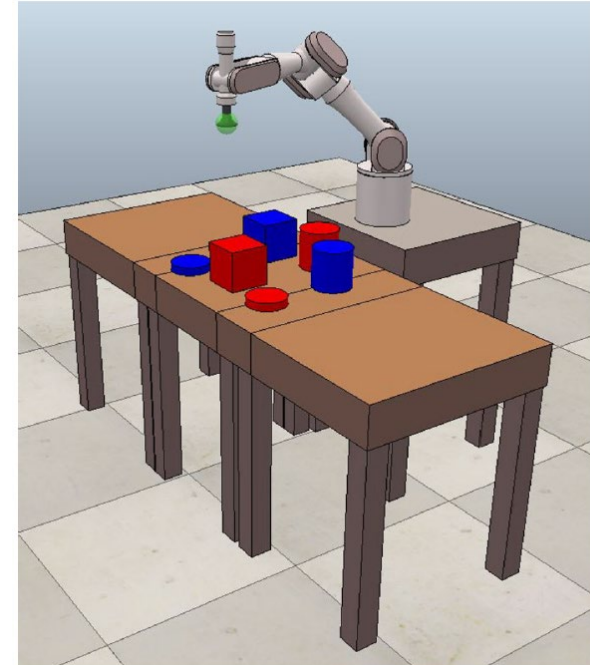
Learning- and Introspection-based Methods⁴



(a) Q-values.



(b) Introspection-based approach.



Question. Why the action move right or move left have not been chosen by the agent.

Explanation. I have selected the action grab object because doing so, I have 59% chances of sorting all the objects successfully, while moving left I have only 38% probability of being successful.

⁴ Cruz, F., Dazeley, R., Vamplew, P., Moreira I. "Explainable Robotic Systems: Understanding Goal-driven Actions in a Reinforcement Learning Scenario". *Neural Computing and Applications*. Springer. 2021.

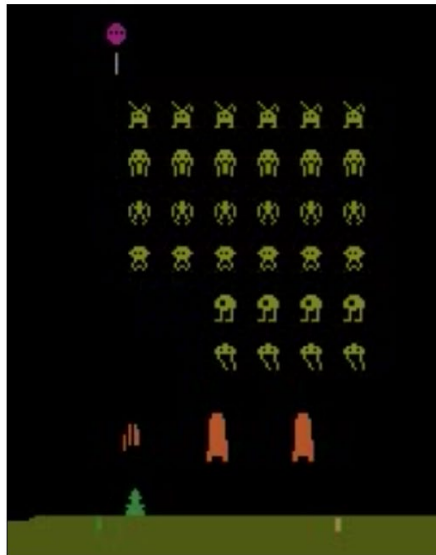
Non-episodic and Continuous Domains⁵

- Introspection method along with Rainbow deep RL algorithm

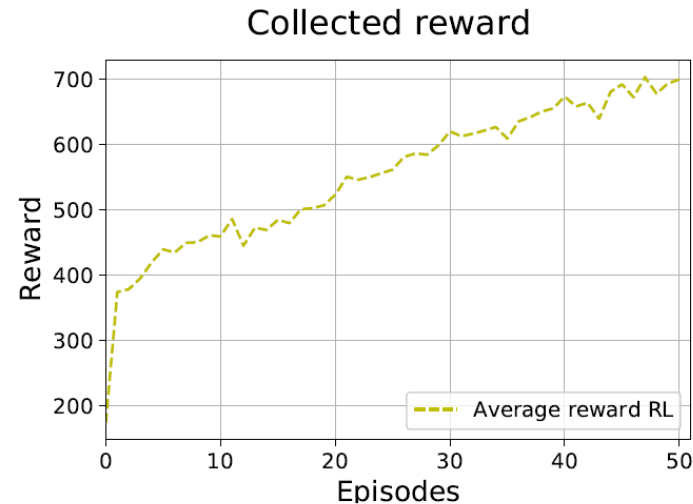
- Maximal reward per step. $\hat{P}_s \approx \frac{1}{2} \cdot \log_{10} \frac{Q(s, a)}{RS} + 1$



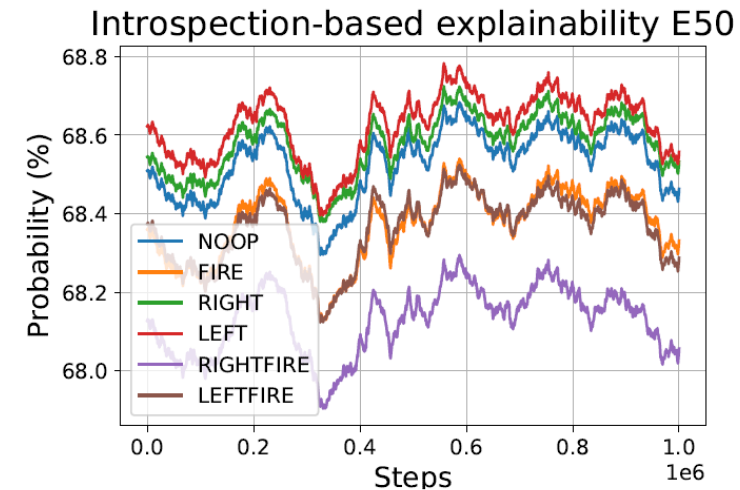
(a) Initial state.



(b) Extra alien ship.



(a) Evaluation reward per episode.



(b) Probability of success.

⁵ Ayala, A., Cruz, F., Fernandes, B., Dazeley, R. "Explainable Deep Reinforcement Learning Using Introspection in a Non-episodic Task". International Conference on Development and Learning (ICDL), Workshop on Human-aligned Reinforcement Learning for Autonomous Agents and Robots, Beijing, China, 2021.

Non-episodic and Continuous Domains⁶

- Drone scenario in Webots.

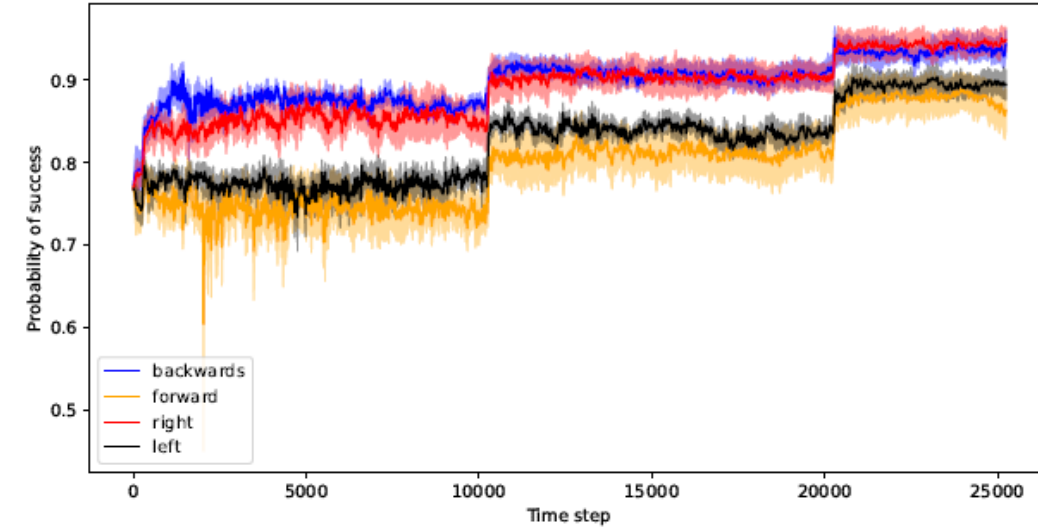


Figure 3: Probabilities of success for every available action in a spot close to the top-left corner.

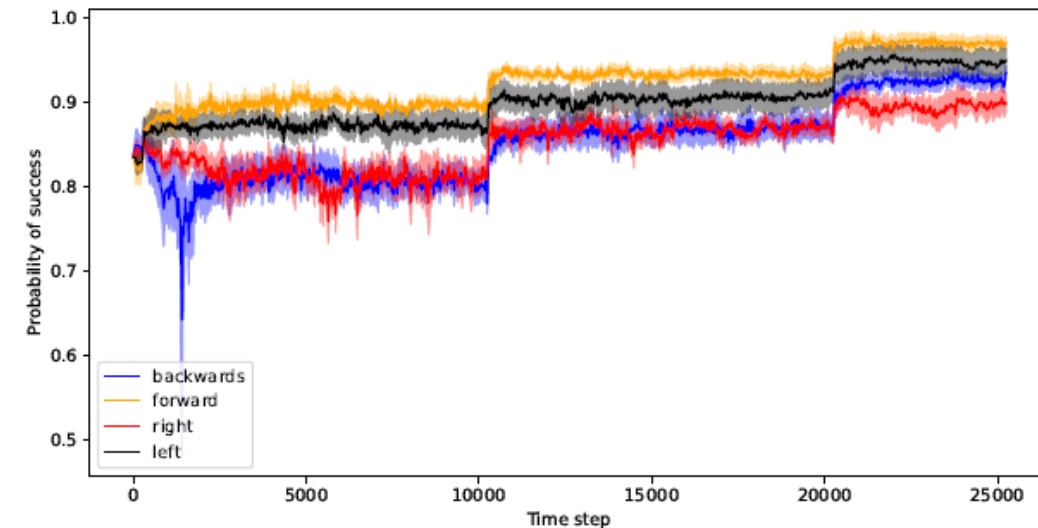


Figure 4: Probabilities of success for every available action in a spot close to the bottom-right corner.

⁶ Schroeter, N., Cruz, F., Wermter, S. "Introspection-based Explainable Reinforcement Learning in Episodic and Non-episodic Scenarios". Accepted at the Australian Conference on Robotics and Automation (ACRA 2022). In press.

Evaluation of Resources⁷

- Memory and CPU usage in the car racing game.

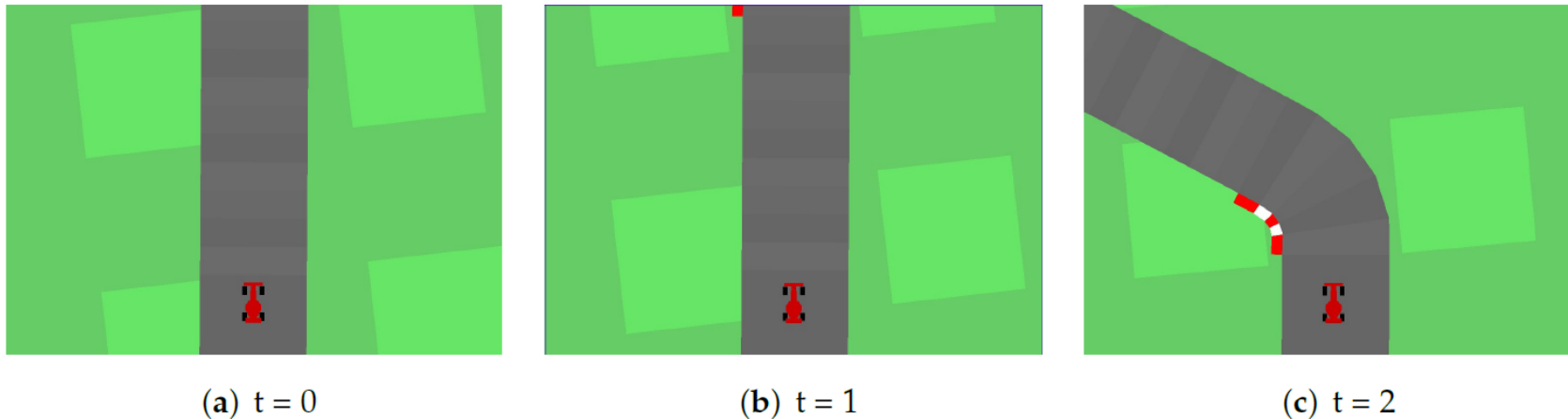
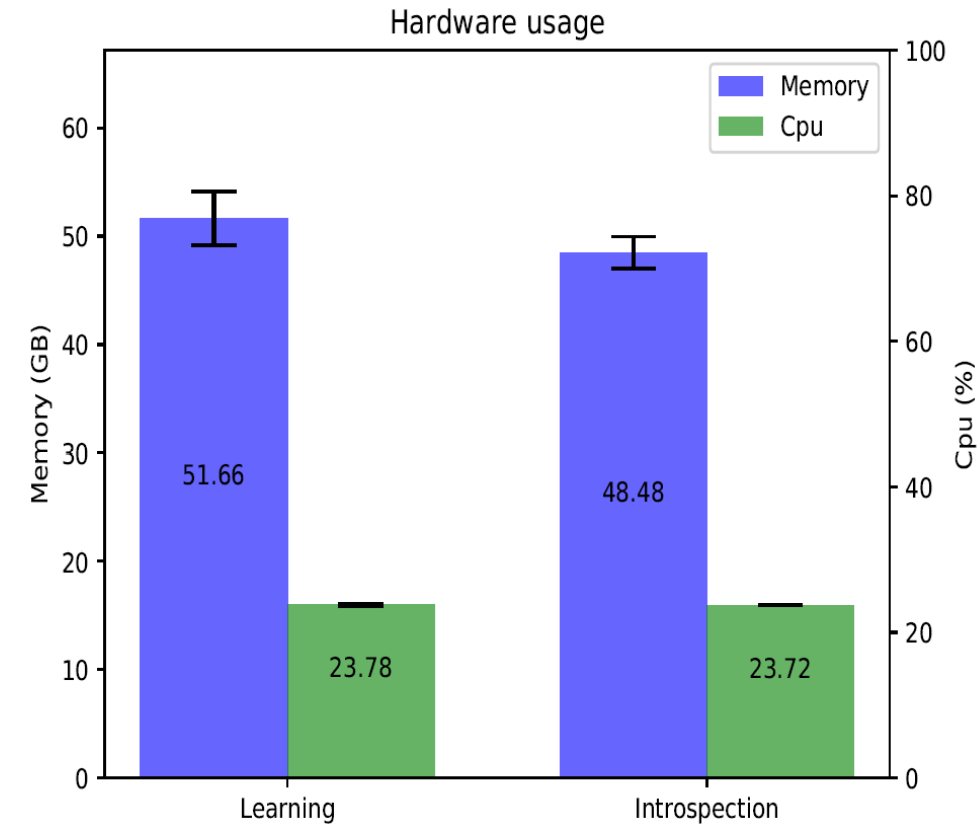
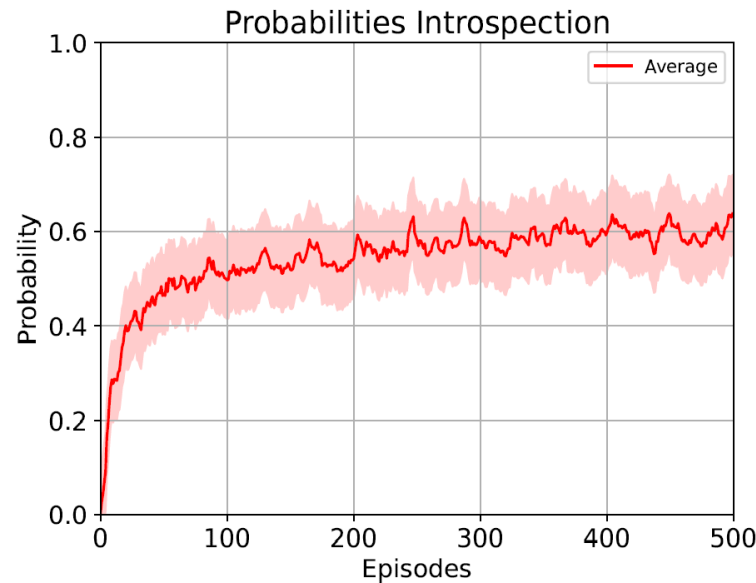
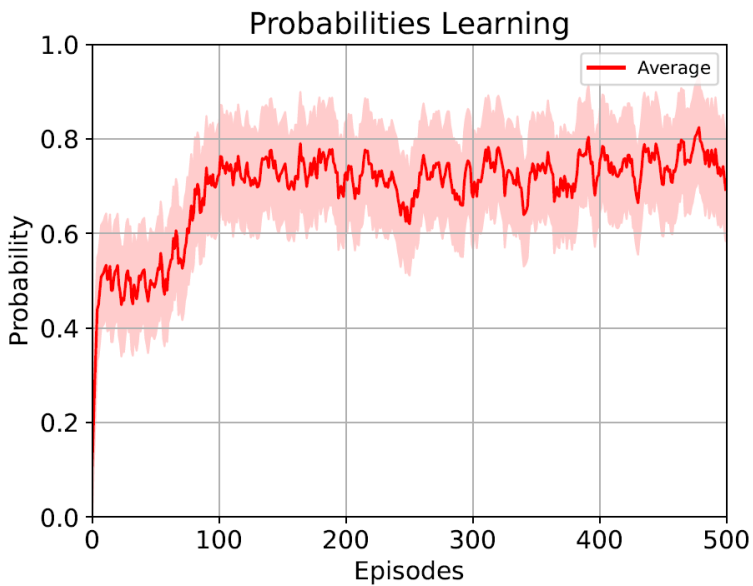


Figure 3. The input is represented by three consecutive images of 96×96 (matrix of $96 \times 96 \times 3$) from the car racing game. The images in the figure are examples since they were previously processed in a gray scale.

⁷ Portugal, E., Cruz, F., Ayala, A., Fernandes, B. "Analysis of Explainable Goal-Driven Reinforcement Learning in a Continuous Simulated Environment". *Algorithms*, 15(3), 91. 2022.

Evaluation of Resources⁷

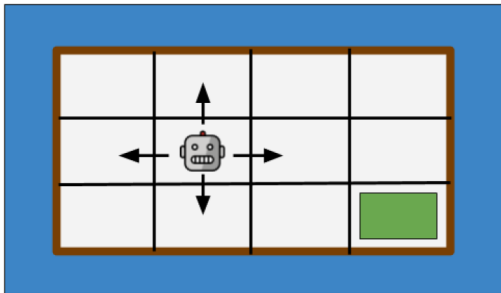
- Memory and CPU usage in the car racing game.



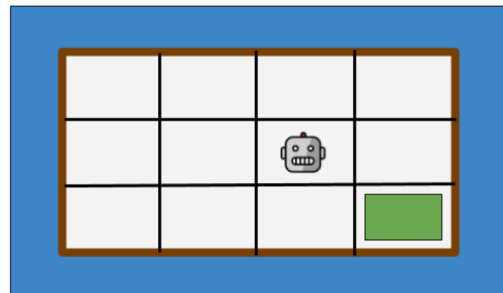
⁷ Portugal, E., Cruz, F., Ayala, A., Fernandes, B. "Analysis of Explainable Goal-Driven Reinforcement Learning in a Continuous Simulated Environment". *Algorithms*, 15(3), 91. 2022.

Evaluating Goal-driven Explanations by Non-experts End-users⁸

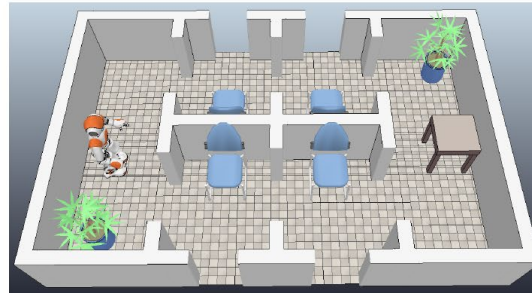
- User study using Amazon Mechanical Turk with 228 participants.



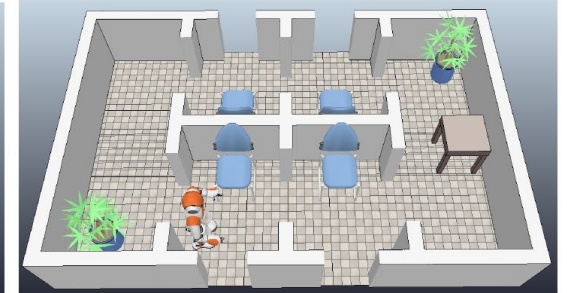
(a) Initial state.



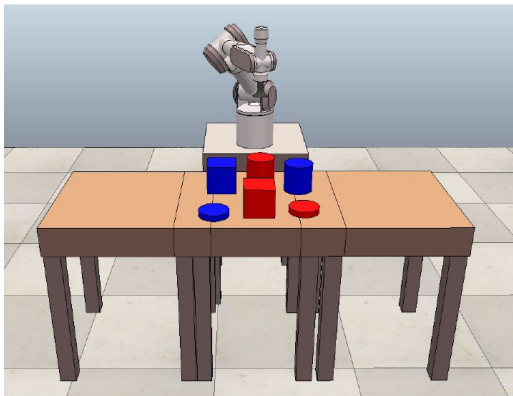
(b) State after 'go east' action.



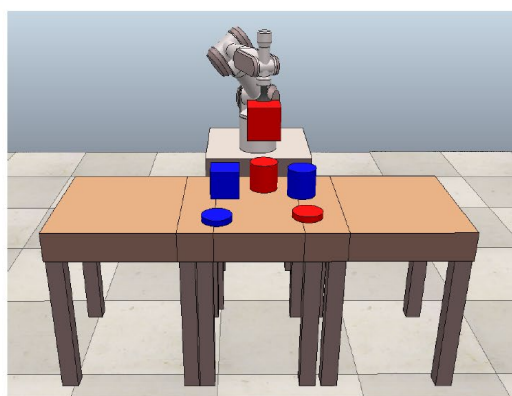
(a) Initial state.



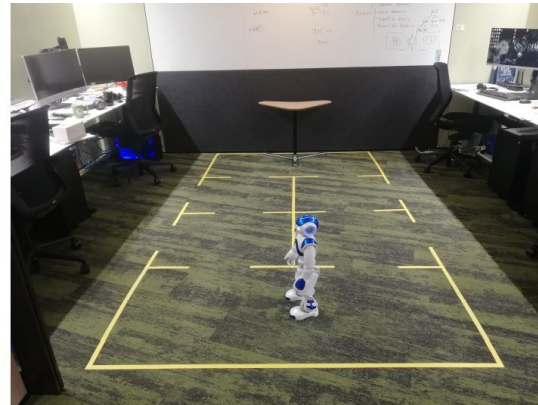
(b) State after 'move right' action.



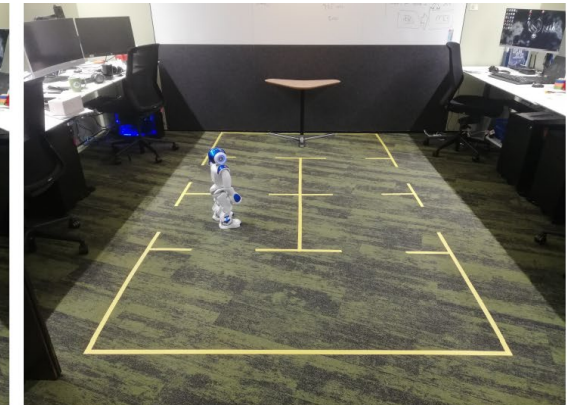
(a) Initial state.



(b) State after 'grab an object' action.



(a) Initial state.



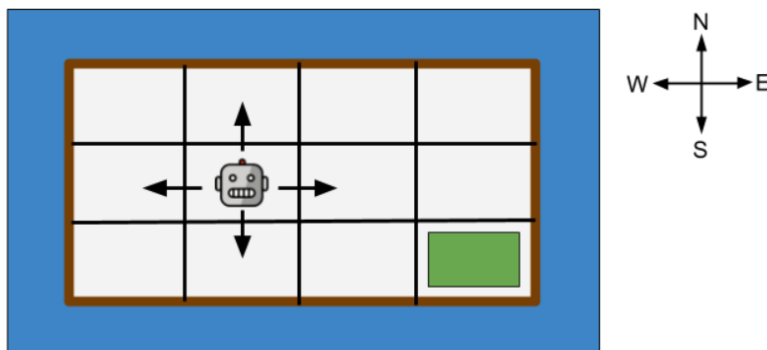
(b) State after 'move to the left' action.

⁸ Cruz, F., Young, C., Dazeley, R., Vamplew, P. "Evaluating Human-like Explanations for Robot Actions in Reinforcement Learning Scenarios". IEEE/RSS International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 2022.

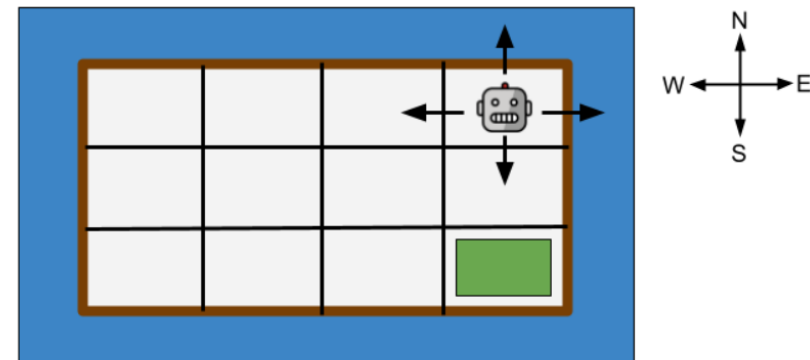
Evaluating Goal-driven Explanations by Non-experts End-users⁸

- Technical, human-like and standalone, counterfactual explanations.
 - **[S]** After performing 'go east' from (1,1). Why did you move to the east?
 - **[T]** I moved to the east because it has a Q-value of -0.411
 - **[H]** I moved to the east because it has a 65.6% probability of reaching the green position
 - **[C]** After performing 'go south' from (3,0). Why you did not move to the east?
 - **[T]** I did not move to the east because it has a Q-value of -0.998, while moving south has a Q-value of 0.181
 - **[H]** I did not move to the east because it has 0% probability of reaching the green position, instead moving south has 73.6% probability

Initial situation



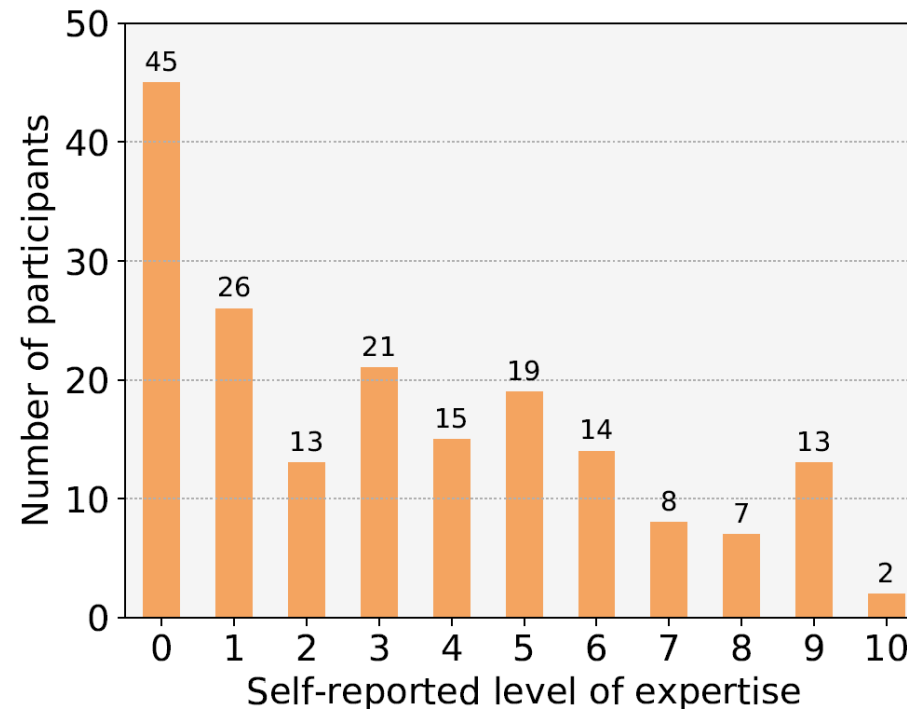
Initial situation



⁸ Cruz, F., Young, C., Dazeley, R., Vamplew, P. "Evaluating Human-like Explanations for Robot Actions in Reinforcement Learning Scenarios". IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 2022.

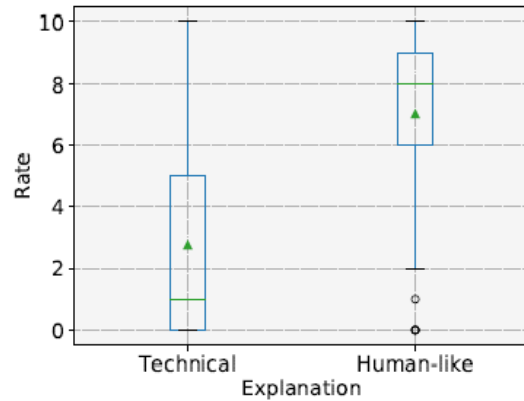
Evaluating Goal-driven Explanations by Non-experts End-users⁸

- Most of participants reported no previous expertise in machine learning.

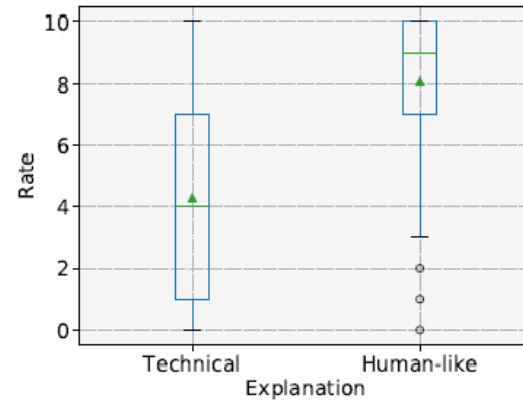


⁸ Cruz, F., Young, C., Dazeley, R., Vamplew, P. "Evaluating Human-like Explanations for Robot Actions in Reinforcement Learning Scenarios". IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 2022.

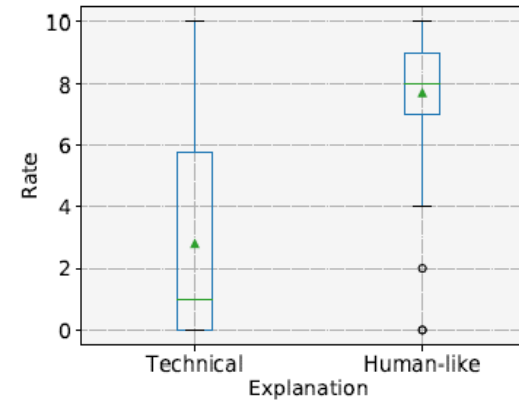
Evaluating Goal-driven Explanations by Non-experts End-users⁸



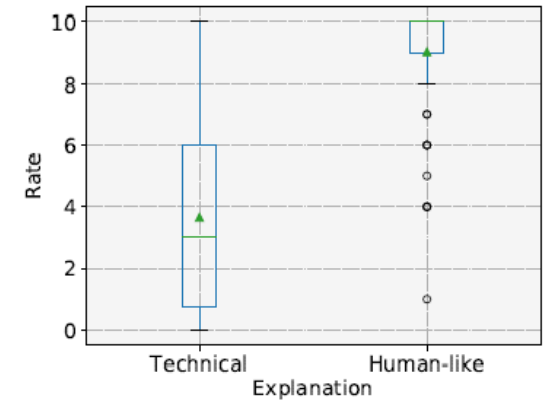
(a) Island scenario – Go east.



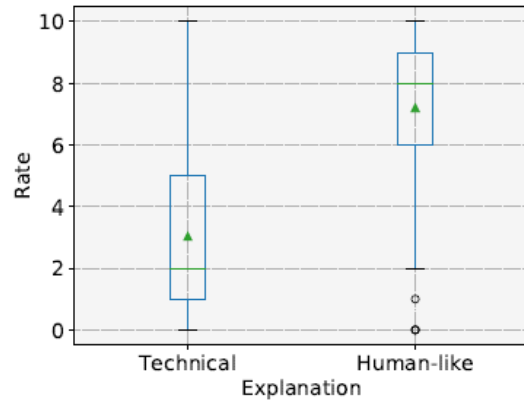
(b) Island scenario – Go south.



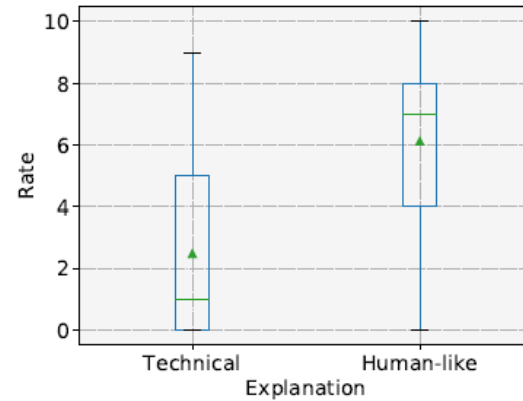
(c) Navigation scenario – Move right.



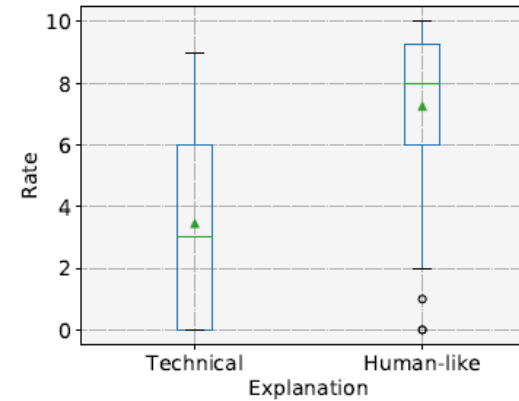
(d) Navigation scenario – Move straight.



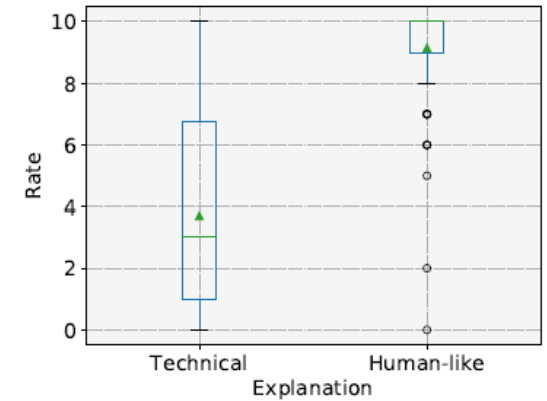
(e) Robot arm – Move to the right.



(f) Robot arm – Grab an object.



(g) Real-world Nao – Move straight.

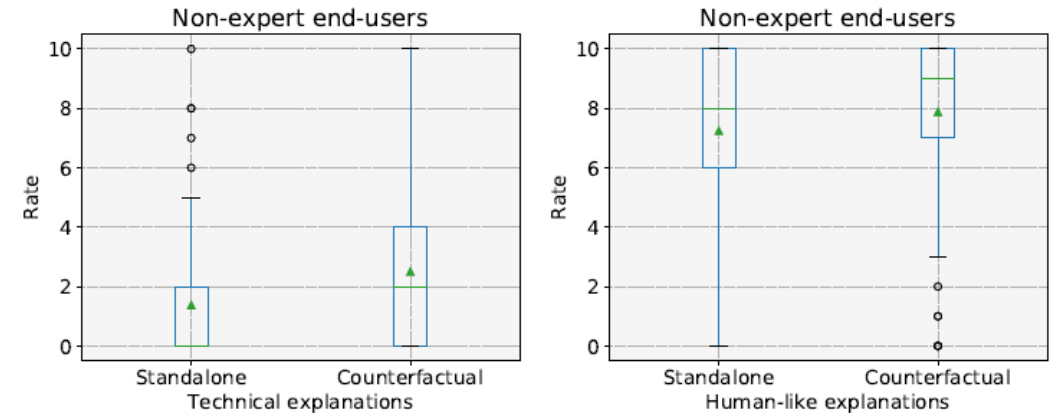
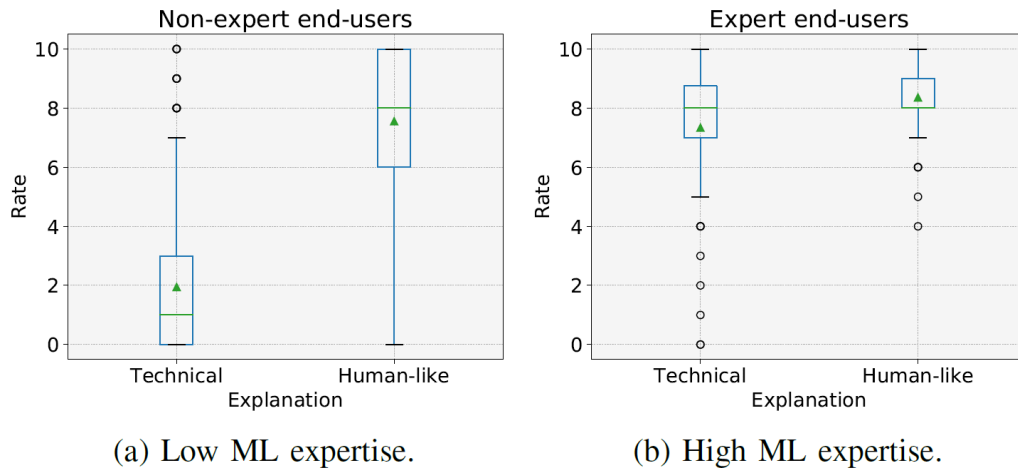


(h) Real-world Nao – Move to the left.

⁸ Cruz, F., Young, C., Dazeley, R., Vamplew, P. "Evaluating Human-like Explanations for Robot Actions in Reinforcement Learning Scenarios". IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 2022.

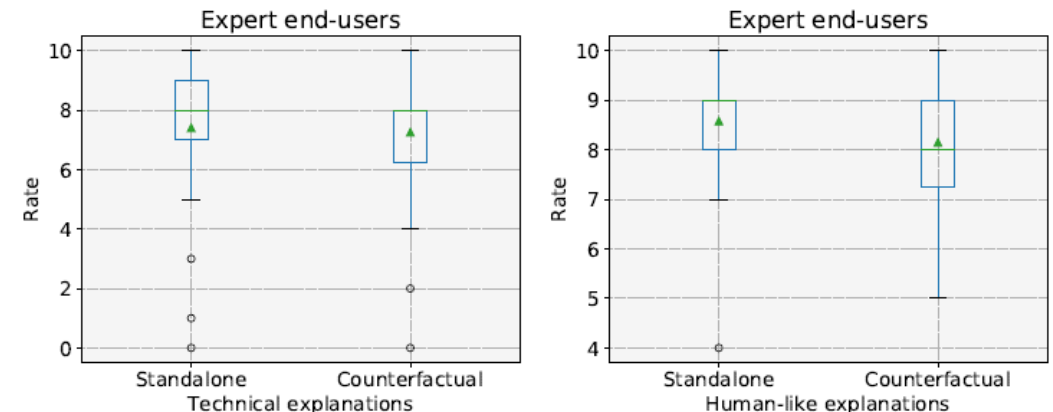
Evaluating Goal-driven Explanations by Non-experts End-users⁸

- Expert and non-expert end-users.



(a) Non-experts – Technical.

(b) Non-experts – Human-like.



(c) Experts – Technical.

(d) Experts – Human-like.

⁸ Cruz, F., Young, C., Dazeley, R., Vamplew, P. "Evaluating Human-like Explanations for Robot Actions in Reinforcement Learning Scenarios". IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 2022.

Conclusions and Future Work

- Human-like explanations are in general well accepted by non-expert end-users.
- Combination of goal-driven and feature-based explanations is needed.
- Interaction between mechanisms.
- Real-world high-dimensional robot learning.

Recruiting PhD Students

- Australian Government Research Training Program (RTP) Scholarship
 - ~37,000 AUD per year + Tuition fee scholarship
 - Health insurance
 - 3.5 years



University of New South Wales
School of Computer Science and Engineering

Explainable Robotics Systems in Reinforcement Learning Scenarios

Dr. Francisco Cruz

f.cruz@unsw.edu.au

<https://www.franciscocruz.org/>