On Generating Explanations for Reinforcement Learning Policies: An Empirical Study

Mikihisa (Miki) Yuasa[1], Huy T. Tran[1], and Ramavarapu S. Sreenivas[2]

[1]Department of Aerospace Engineering / [2]Department of Industrial & Enterprise Systems Engineering, The Grainger College of Engineering, University of Illinois Urbana-Champaign

LIRA@illinois
Lab for Intellignet Robots and Agents

INTRODUCTION

Black-box nature of neural networks: While learning-based methods have advanced robot decision-making and control, their lack of interpretability raises concerns for **safety-critical applications** like autonomous vehicles.

Need for explainability: Formal methods, such as **Linear Temporal Logic (LTL)**, offer a structured way to interpret robot policies by prioritizing constraints based on importance.

Limitations of existing approaches

Existing approaches typically learn from a fixed set of example trajectories. This makes them highly dependent on the provided data and unable to access the policy's internal logic, such as its preference or certainty for actions in unobserved states.

Contribution

- Introduce a novel method that overcomes these limitations by **assuming direct access** to the target policy itself.
- Automatically find an explanation in the form of a **Linear Temporal Logic (LTL)** formula.

Key Idea

Compare the **action distributions** of the target policy with policies explicitly optimized for candidate LTL formulas, allowing for a **more nuanced evaluation** that captures the agent's underlying intent.

Experimental Setup

We tested our method in **three distinct simulated environments**.

Target Policy Goals

CtF: Capture the red flag while avoidng the red agent.

Parking: Park in a designated spot while avoiding walls and another car.

Robot Nav.: Navigate to a goal while avoiding hazards.

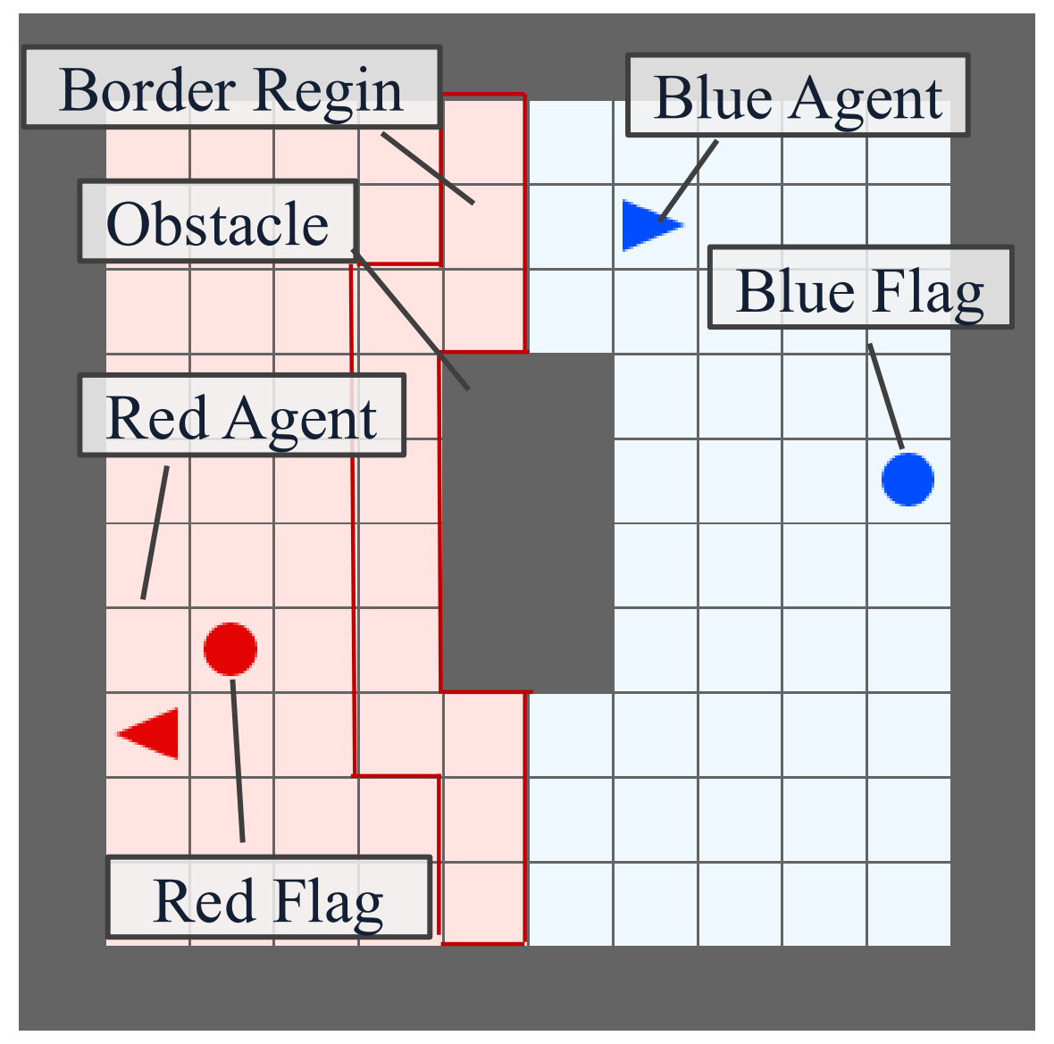

Fig 2. Capture-the-Flag (CtF)

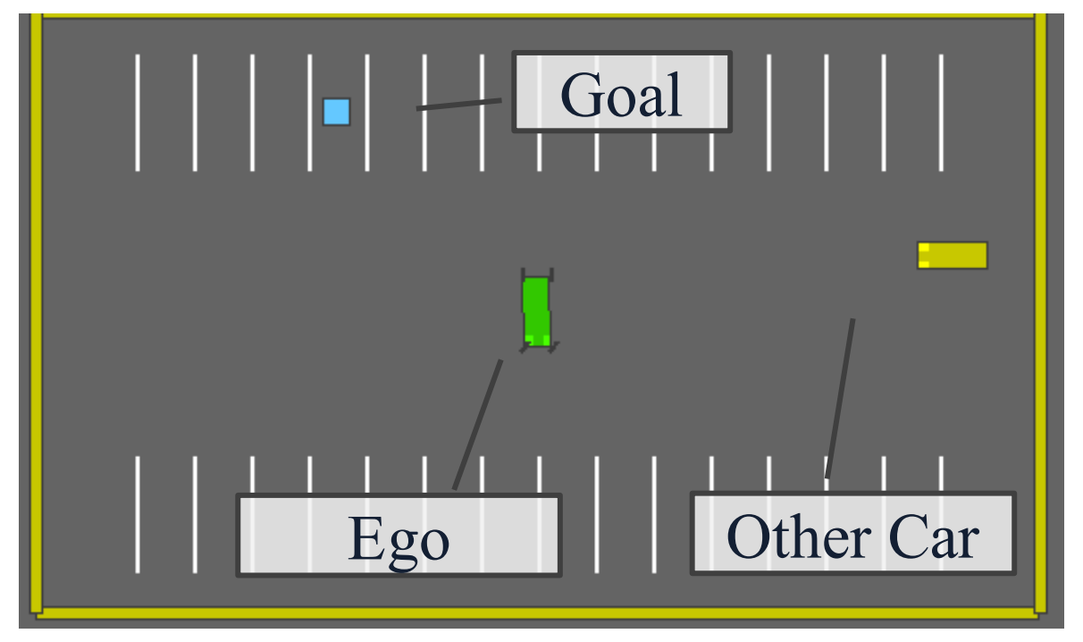

Fig 3. Adversarial Parking

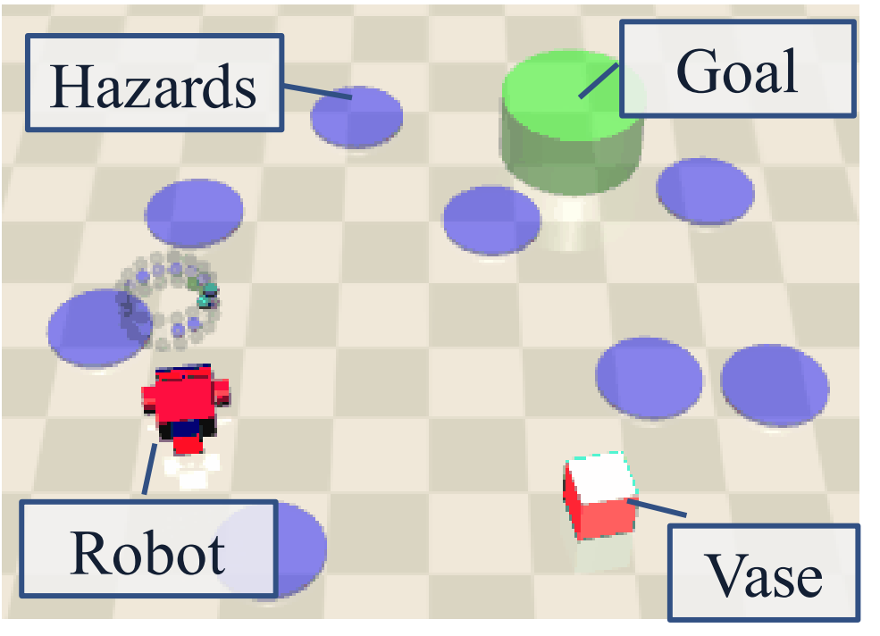

Fig 4. Robot Navigation

METHOD

We frame the problem as a search for the LTL formula that produces a policy most similar to the target policy (π_{tar}).

Explanation Structure

We define explanations as LTL formulas with a specific structure, capturing both a goal and a safety constraint:

$$\phi = \mathcal{F}(\phi_F) \wedge \mathcal{G}(\phi_G).$$

- $\mathcal{F}(\phi_F)$: *"Eventually, achieve some task ϕ_F."*
- $\mathcal{G}(\phi_G)$: *"Globally, always satisfy some safety constraint ϕ_G."*
- ϕ_F and ϕ_G are logical combinations of user-defined atomic predicates (e.g., distance_to_goal < 1).

Evaluation Metric: Weighted KL Divergence

How do we measure similarity? We compare the policies' action distributions ($\pi(a|s)$) in key states.

- **Utility Score** (U^ϕ)**:** We calculate the **weighted Kullback-Leibler (wKL) divergence** between distributions of the target policy and a candidate policy:

$$U^\phi = -\sum_s w_s D_{\text{KL}}\left(\pi_\phi(\cdot\,|s) \,\|\, \pi_{\text{tar}}(\cdot\,|s)\right).$$

- **Weighting** (w_s): The weights emphasize states where the target policy is most certain about its action, as these are the most informative states for understanding its intent.
- **Benefit:** This avoids trivial "catch-all" explanations (e.g., "*eventually do anything*") that might be satisfied by trajectories but do not capture the policy's specific logic.

The Search Algorithm

We employ a **greedy local-search algorithm** to navigate the space of possible LTL explanations.

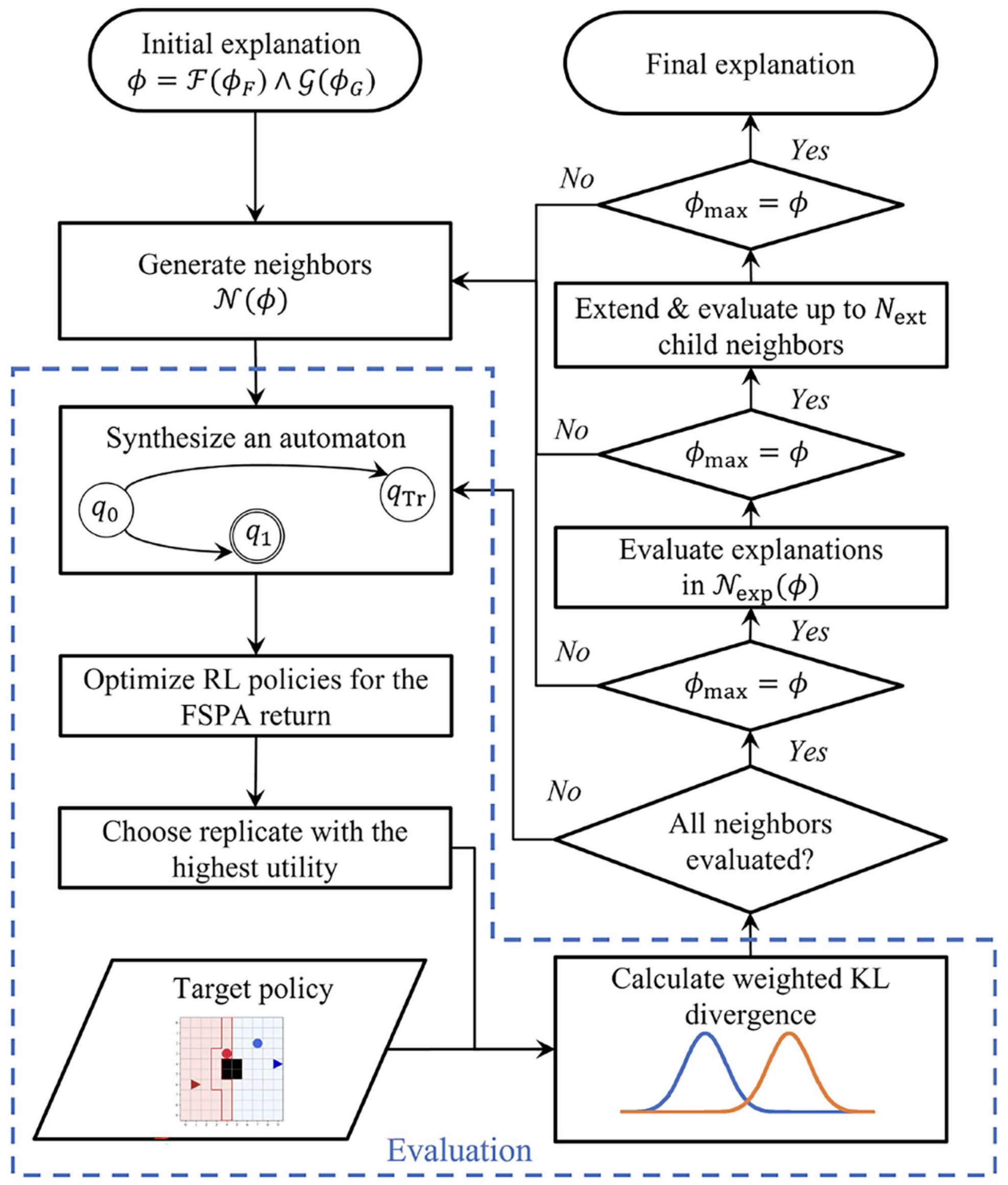
Initial explanation
$\phi = \mathcal{F}(\phi_F) \wedge \mathcal{G}(\phi_G)$
Generate neighbors
$\mathcal{N}(\phi)$
Synthesize an automaton
q_0
q_1
q_{Tr}
Optimize RL policies for the FSPA return
Choose replicate with the highest utility
Target policy
Calculate weighted KL divergence
Evaluation
Final explanation
Yes
No
$\phi_{\max} = \phi$
Extend & evaluate up to N_{ext} child neighbors
Yes
No
$\phi_{\max} = \phi$
Evaluate explanations in $\mathcal{N}_{\text{exp}}(\phi)$
Yes
No
$\phi_{\max} = \phi$
Yes
No
All neighbors evaluated?

Fig 1. Overview of our proposed search algorithm.

RESULTS

Table I. CtF & Parking Results.
Target policies were successfully found in Searches 1 & 2 (CtF) and 1, 2, & 3 (Parking).

Search	CtF explanations	wKL div. [-]	Searched specs [%]	Parking explanations	wKL div. [-]	Searched specs [%]
1	$\boldsymbol{\mathcal{F}(\psi_{ba,rf} \wedge \neg\psi_{ra,bf}) \wedge \mathcal{G}(\neg\psi_{ba,ra} \vee \psi_{ba,bt})}$	$\mathbf{8.00 \times 10^{-8}}$	**8.13**	$\boldsymbol{\mathcal{F}(\psi_{g}) \wedge \mathcal{G}(\neg\psi_{o} \wedge \neg\psi_{w})}$	**0.00**	**37.5**
2	$\boldsymbol{\mathcal{F}(\psi_{ba,rf} \wedge \neg\psi_{ra,bf}) \wedge \mathcal{G}(\neg\psi_{ba,ra} \vee \psi_{ba,bt})}$	$\mathbf{8.00 \times 10^{-8}}$	**10.2**	$\boldsymbol{\mathcal{F}(\psi_{g}) \wedge \mathcal{G}(\neg\psi_{o} \wedge \neg\psi_{w})}$	**0.00**	**29.2**
3	$\mathcal{F}(\neg\psi_{\text{ba,bt}}) \wedge \mathcal{G}((\neg\psi_{\text{ba,ra}} \wedge \psi_{\text{ba,rf}}) \vee (\neg\psi_{\text{ra,bf}}))$	1.46×10^{-7}	6.56	$\boldsymbol{\mathcal{F}(\psi_{g}) \wedge \mathcal{G}(\neg\psi_{o} \wedge \neg\psi_{w})}$	**0.00**	**37.5**
4	$\mathcal{F}(\psi_{\text{ba,rf}}) \wedge \mathcal{G}((\neg\psi_{\text{ba,bt}} \wedge \neg\psi_{\text{ra,bf}}) \vee (\neg\psi_{\text{ba,ra}}))$	7.42×10^{-7}	8.44	$\mathcal{F}(\psi_{\text{o}}) \wedge \mathcal{G}(\neg\psi_{\text{g}} \vee \neg\psi_{\text{w}})$	4.61×10^{-4}	44.8
5	$\mathcal{F}(\neg\psi_{\text{ba,bt}} \wedge \neg\psi_{\text{ra,bf}}) \wedge \mathcal{G}((\neg\psi_{\text{ba,rf}}) \vee (\neg\psi_{\text{ba,ra}}))$	1.57×10^{-6}	6.56	$\mathcal{F}(\psi_{\text{o}}) \wedge \mathcal{G}(\neg\psi_{\text{g}} \vee \neg\psi_{\text{w}})$	4.61×10^{-4}	33.3
6	$\mathcal{F}((\neg\psi_{\text{ra,bf}}) \vee (\psi_{\text{ba,rf}})) \wedge \mathcal{G}(\neg\psi_{\text{ba,ra}} \vee \psi_{\text{ba,bt}})$	5.95×10^{-6}	9.06	$\mathcal{F}(\psi_{\text{w}}) \wedge \mathcal{G}(\neg\psi_{\text{g}} \vee \psi_{\text{o}})$	6.40×10^{-4}	30.2
7	$\mathcal{F}(\psi_{\text{ba,ra}}) \wedge \mathcal{G}((\neg\psi_{\text{ba,bt}}) \vee (\neg\psi_{\text{ba,rf}} \wedge \neg\psi_{\text{ra,bf}}))$	1.18×10^{-5}	8.59	$\mathcal{F}(\neg\psi_{\text{o}} \wedge \psi_{\text{w}}) \wedge \mathcal{G}(\neg\psi_{\text{g}})$	7.35×10^{-4}	28.1
8	$\mathcal{F}((\psi_{\text{ba,ra}}) \wedge (\neg\psi_{\text{ba,rf}})) \wedge \mathcal{G}((\psi_{\text{ba,bt}}) \vee (\neg\psi_{\text{ra,bf}}))$	1.18×10^{-5}	11.4	$\mathcal{F}(\neg\psi_{\text{o}} \wedge \psi_{\text{w}}) \wedge \mathcal{G}(\neg\psi_{\text{g}})$	7.35×10^{-4}	27.1
9	$\mathcal{F}((\psi_{\text{ra,bf}}) \vee (\neg\psi_{\text{ba,bt}})) \wedge \mathcal{G}((\neg\psi_{\text{ba,ra}}) \vee (\psi_{\text{ba,rf}}))$	1.62×10^{-5}	7.19	-	-	-
10	$\mathcal{F}((\neg\psi_{\text{ba,ra}}) \vee (\neg\psi_{\text{ra,bf}})) \wedge \mathcal{G}((\psi_{\text{ba,bt}}) \vee (\psi_{\text{ba,rf}}))$	8.52×10^{-5}	6.56	-	-	-

Table II. Robot Navigation Results for LTL (left) & Non-LTL (right) Target Policies.

Search	Robot explanations	wKL div. [-]	Searched specs [%]	Robot explanations	wKL div. [-]	Searched specs [%]
1	$\boldsymbol{\mathcal{F}(\psi_{gl}) \wedge \mathcal{G}(\neg\psi_{hz} \wedge \neg\psi_{vs})}$	**0.00**	**21.8**	$\boldsymbol{\mathcal{F}(\psi_{gl} \vee \neg\psi_{vs}) \wedge \mathcal{G}(\neg\psi_{hz})}$	$\mathbf{3.0501 \times 10^{-4}}$	**39.6**
2	$\boldsymbol{\mathcal{F}(\psi_{gl}) \wedge \mathcal{G}(\neg\psi_{hz} \wedge \neg\psi_{vs})}$	**0.00**	**30.2**	$\mathcal{F}(\neg\psi_{\text{gl}}) \wedge \mathcal{G}(\neg\psi_{\text{hz}} \vee \neg\psi_{\text{vs}})$	3.0504×10^{-4}	29.2
3	$\boldsymbol{\mathcal{F}(\psi_{gl}) \wedge \mathcal{G}(\neg\psi_{hz} \wedge \neg\psi_{vs})}$	**0.00**	**26.0**	$\mathcal{F}(\neg\psi_{\text{gl}}) \wedge \mathcal{G}(\neg\psi_{\text{hz}} \vee \neg\psi_{\text{vs}})$	3.0504×10^{-4}	27.1
4	$\boldsymbol{\mathcal{F}(\psi_{gl}) \wedge \mathcal{G}(\neg\psi_{hz} \wedge \neg\psi_{vs})}$	**0.00**	**29.2**	$\mathcal{F}(\neg\psi_{\text{gl}}) \wedge \mathcal{G}(\neg\psi_{\text{hz}} \vee \neg\psi_{\text{vs}})$	3.0504×10^{-4}	27.1
5	$\boldsymbol{\mathcal{F}(\psi_{gl}) \wedge \mathcal{G}(\neg\psi_{hz} \wedge \neg\psi_{vs})}$	**0.00**	**21.9**	$\mathcal{F}(\neg\psi_{\text{gl}}) \wedge \mathcal{G}(\neg\psi_{\text{hz}} \vee \neg\psi_{\text{vs}})$	3.0504×10^{-4}	29.2
6	$\boldsymbol{\mathcal{F}(\psi_{gl}) \wedge \mathcal{G}(\neg\psi_{hz} \wedge \neg\psi_{vs})}$	**0.00**	**24.0**	$\mathcal{F}(\neg\psi_{\text{gl}}) \wedge \mathcal{G}(\neg\psi_{\text{hz}} \vee \neg\psi_{\text{vs}})$	3.0504×10^{-4}	33.3
7	$\mathcal{F}(\neg\psi_{\text{vs}}) \wedge \mathcal{G}(\psi_{\text{gl}} \vee \neg\psi_{\text{hz}})$	2.64×10^{-4}	25.0	$\mathcal{F}(\psi_{\text{gl}} \vee \psi_{\text{hz}}) \wedge \mathcal{G}(\neg\psi_{\text{vs}})$	3.0508×10^{-4}	29.2
8	$\mathcal{F}(\neg\psi_{\text{vs}}) \wedge \mathcal{G}(\psi_{\text{gl}} \vee \neg\psi_{\text{hz}})$	2.64×10^{-4}	27.1	$\mathcal{F}(\psi_{\text{gl}} \vee \psi_{\text{hz}}) \wedge \mathcal{G}(\neg\psi_{\text{vs}})$	3.0508×10^{-4}	21.9

ANALYSIS

Our method successfully recovered the underlying objectives of the target policies.

Ground-Truth Recovery: In the **CtF**, **Parking**, and **Robot Navigation** environments, where the target policy was trained on a known LTL formula, our search **successfully identified the exact ground-truth formula** as the best explanation.

Plausible Explanations: For the **Robot Navigation** task with a non-LTL target policy, our method found a reasonable explanation: *"Eventually, the robot reaches the goal or does not hit the vase. Globally, the robot does not enter a hazard."* This captures both the primary goal and the learned safety behavior.

Ablation Study: Removing any key component of our method (the search extension heuristics or the wKL weighting) caused the search to fail, confirming the importance of each part of the algorithm.

CONCLUSIONS

We have developed a robust method for generating LTL explanations of RL policies by comparing action distributions. This approach provides deeper insight into an agent's learned behavior than existing methods.

Limitations & Future Directions:

Scalability: The search is computationally intensive. Future work could explore neural network representations of LTL to improve efficiency.

Predicate Definition: The method currently requires user-defined predicates. Automating predicate discovery is a key next step.

Natural Language: Translating the final LTL formula into natural language would further improve interpretability.

ACKNOWLEDGEMENTS

This work was supported in part by ONR grant and JASSO Study Abroad Fellowship Program (Graduate Degree Program).

LINKS

Paper (IEEE L-CSS)

GitHub

ILLINOIS