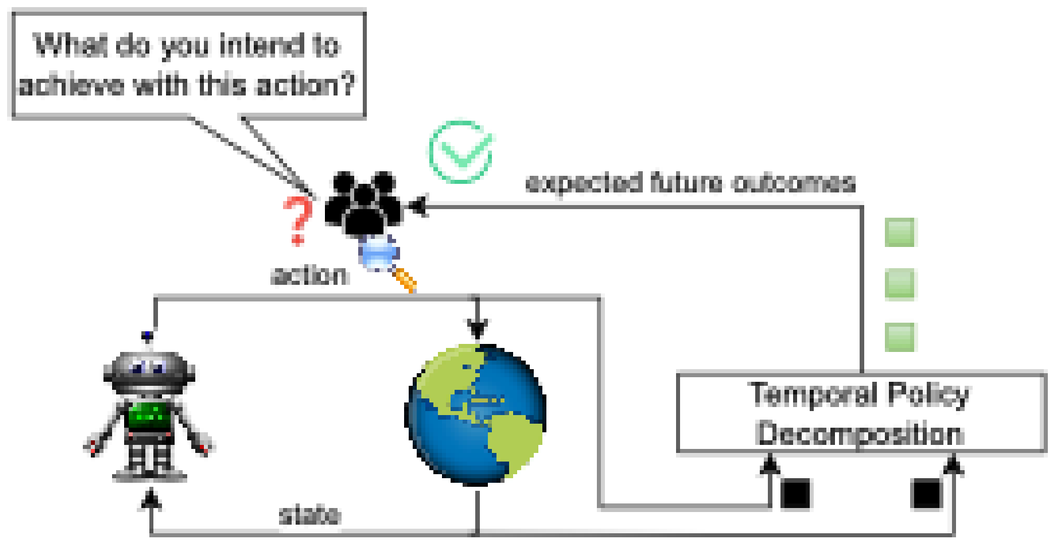
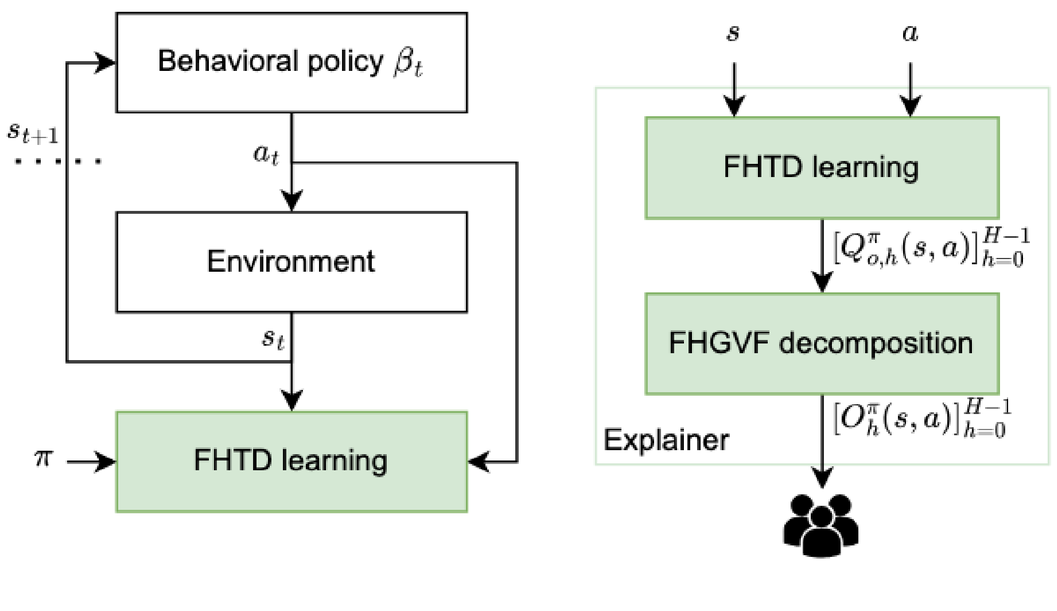
EXPLAINABLE REINFORCEMENT LEARNING VIA TEMPORAL POLICY DECOMPOSITION

Franco Ruggeri, KTH & Ericsson Research
Supervisors: Karl Henrik Johansson, Rafia Inam

WHAT DO YOU INTEND TO ACHIEVE?

- Reinforcement Learning (RL) involves sequential decision-making.
- An RL agent acts to maximize cumulative rewards.
- An RL action could be an intermediate step for a delayed reward.
- RL actions should be explained in terms of the future trajectory implied by the action.
- Existing eXplainable RL (XRL) methods focus on explaining the state-action mapping, neglecting the temporal dimension.





(a) Training (b) Inference

TEMPORAL POLICY DECOMPOSITION

1. Learn Fixed-Horizon Generalized Value Function [1] (off-policy)

$$\hat{G}_h = r_t + \gamma \hat{Q}_{h-1}(s_{t+1}, a_{t+1}),\ a_{t+1} \sim \pi(\cdot|s_{t+1})$$

$$\hat{Q}^{(t+1)}_h(s_t, a_t) \leftarrow \hat{Q}^{(t)}_h(s_t, a_t) + \alpha_{t,h}(s_t, a_t)[\hat{G}_h - \hat{Q}^{(t)}_h(s_t, a_t)]$$

2. Decompose FHGVFs into Expected Future Outcomes (EFOs):

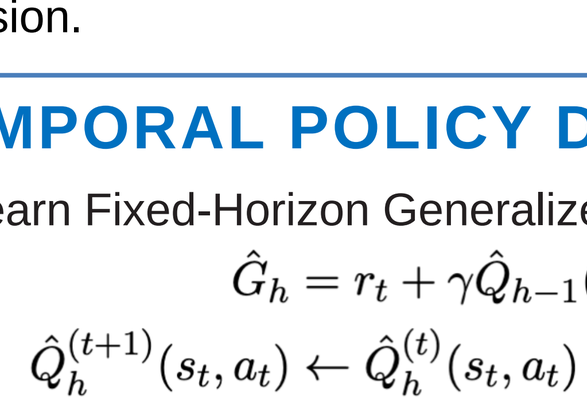
$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & \gamma & 0 & \cdots & 0 \\ 1 & \gamma & \gamma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \gamma & \gamma^2 & \cdots & \gamma^{H-1} \end{bmatrix} \begin{bmatrix} O_0(s,a) \\ O_1(s,a) \\ O_2(s,a) \\ \vdots \\ O_{H-1}(s,a) \end{bmatrix} = \begin{bmatrix} Q^{\pi}_{o,0}(s,a) \\ Q^{\pi}_{o,1}(s,a) \\ Q^{\pi}_{o,2}(s,a) \\ \vdots \\ Q^{\pi}_{o,H-1}(s,a) \end{bmatrix}$$

$$O^{\pi}_h(s,a) = \mathbb{E}\left[o(s_h, a_h, s_{h+1}) | s_0 = s, a_0 = a, \pi\right]$$

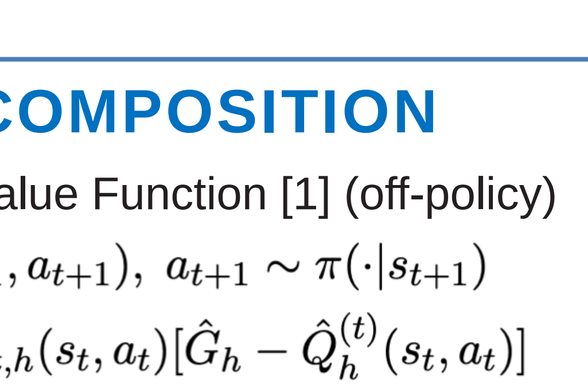
$$Q^{\pi}_{o,H}(s,a) = \mathbb{E}\left[\sum_{h=0}^{H-1} \gamma^h o_h \middle| s_0 = s, a_0 = a, \pi\right]$$

EXPLANATIONS

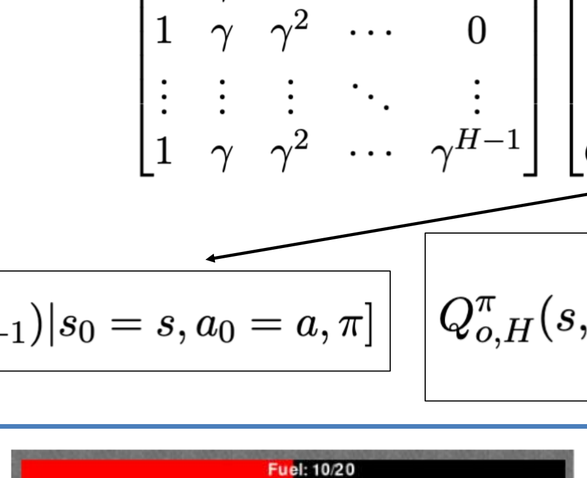
- Probability of future events: If the outcome function is an event (binary), the EFO is the probability of the event in a future time step.
- Expected reward components: If the outcome function is a part of the reward, EFOs provide a reward decomposition, extending [2] with the temporal dimension.
- Contrastive explanations: Comparing EFOs for different state-action pairs provides contrastive explanations (e.g., “Why is moving south preferred over action 2?”).



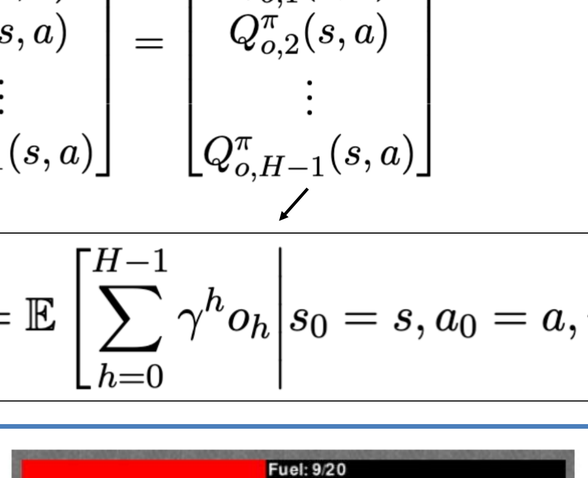
(a) Environment state (fuel 10)



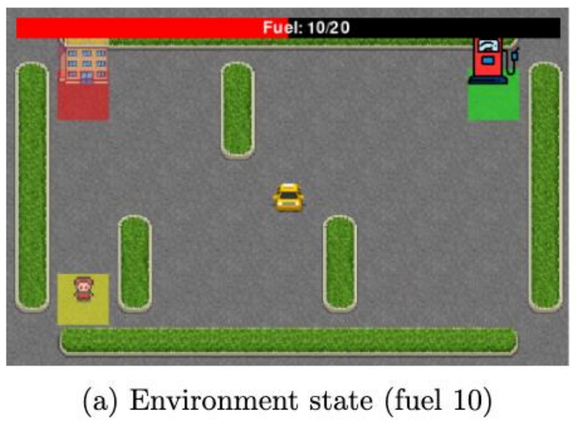
(b) Environment state (fuel 9)



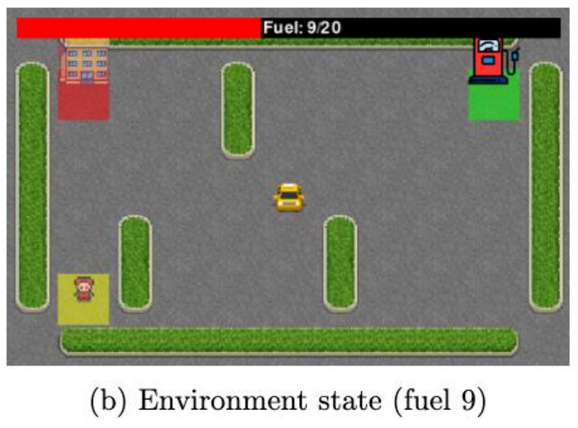
(c) Probabilities of future events (fuel 10)



(d) Probabilities of future events (fuel 9)



(e) Expected reward components (fuel 10)



(f) Expected reward components (fuel 9)

CONCLUSIONS

- Temporal Policy Decomposition [3] is an XRL method that explains individual actions by predicting EFOs
- Results in tabular problems show accurate predictions, which are necessary for having reliable explanations.
- In future work, we will experiment with continuous problems (deep neural networks).

[1] K. De Asis, A. Chan, S. Pitis, R. Sutton, and D. Graves, “Fixed-Horizon Temporal Difference Methods for Stable Reinforcement Learning,” *AAAI*, vol. 34, no. 04, pp. 3741–3748, Apr. 2020, doi: 10.1609/aaai.v34i04.5784.
[2] Z. Juozapaitis, A. Koul, A. Fern, M. Erwig, and F. Doshi-Velez, “Explainable reinforcement learning via reward decomposition,” in *IJCAI/ECAI Workshop on explainable artificial intelligence*, 2019.
[3] F. Ruggeri, A. Russo, R. Inam, K. H. Johansson, “Explainable Reinforcement Learning via Temporal Policy Decomposition”, arXiv, https://arxiv.org/abs/2501.03902.