Aligning Agent Policies with Preferences: Human-Centered Interpretable Reinforcement Learning

AI agents are increasingly developed for high-stakes decision-making, like finance and education. These decisions are captured by a policy, which defines the agent's behavior across situations and contexts. A natural choice for training these policies is reinforcement learning (RL) [6], but achieving strong performance in such complex settings typically requires representing policies with expressive function approximators [2, 5]. While effective, these representations are often not interpretable, hindering our ability to understand and collaborate with these agents [4]. Many desirable attributes of an interpretable policy, such as simplicity or alignment with institutional values, require human feedback. Yet existing methods typically collect such feedback only after training is complete, missing the opportunity to *inform* the learning process itself. Consequently, an unaddressed challenge in interpretable RL is to enable AI agents to integrate preference feedback into policy generation.

To address this gap, we propose a novel framework to align interpretable policies with human feedback during training. Illustrated at a high level in Figure 1, our framework interleaves preference learning with an evolutionary algorithm, using updated preference estimates to guide the generation of better-aligned policies, and using newly-generated policies to query users to refine the preference model. Evolutionary algorithms enable the exploration of the full space of policies; however, it is intractable to maintain separate preference estimates—like win rates or utility values—for each individual policy in this infinite space. To handle this challenge, we propose to represent policies as feature vectors consisting of a finite set of meaningful attributes. For example, among a set of policies with similar performance, some may be more intuitive or more amenable to human intervention. To maximize the value of each user query, we employ a novel filtering technique to avoid presenting policies that are dominated in all dimensions, as repeated selections of clearly superior policies provides little information.

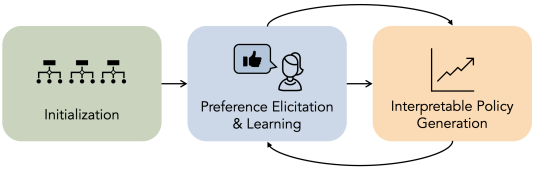

Figure 1: Overview of PASTEL, a novel algorithm for interpretable RL. Users provide feedback on interpretable models, and feedback-informed preference estimates guide policy generation.

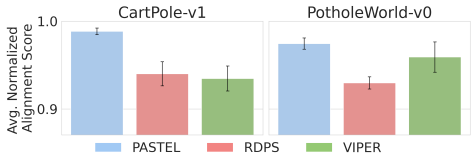

Figure 2: PASTEL creates more aligned interpretable policies.

We validate our method with experiments using decision-tree-structured policies [3, 7], as they are widely considered to be interpretable. We leverage synthetic preference data on two RL environments: CartPole and PotholeWorld [8]. As shown in Figure 2, PASTEL produces substantially more preference-aligned decision-tree policies than both VIPER [1] and RDPS in both environments. We also show that it requires fewer preference queries to produce such policies and is more robust to preference noise. By bridging the gap between training RL agents and evaluating their explanations, we believe our work opens new avenues for developing more interpretable, user-centered RL systems.

References

[1] O. Bastani, Y. Pu, and A. Solar-Lezama. Verifiable reinforcement learning via policy extraction. *NeurIPS*, 2018.

[2] M. Carroll et al. Uni [mask]: Unified inference in sequential decision problems. *NeurIPS*, 2022.

[3] J. Chen et al. Rgmdt: Return-gap-minimizing decision tree extraction in non-euclidean metric space. *NeurIPS*, 2024.

[4] E. M. Kenny, M. Tucker, and J. Shah. Towards interpretable deep reinforcement learning with human-friendly prototypes. In *The Eleventh International Conference on Learning Representations*, 2023.

[5] V. Mnih et al. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[6] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[7] P. Tambwekar and M. Gombolay. Towards reconciling usability and usefulness of policy explanations for sequential decision-making systems. *Frontiers in Robotics and AI*, 2024.

[8] N. Topin, S. Milani, F. Fang, and M. Veloso. Iterative bounding MDPs: Learning interpretable policies via non-interpretable methods. In *AAAI*, 2021.