Reward Trees for Interpretable Reinforcement Learning from Human Feedback

T. Bewley[1], J. Lawry[1], A. Richards[1], R. Craddock[2], and I. Henderson[2]

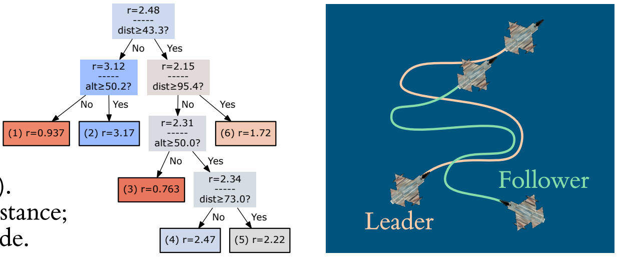
[1]University of Bristol, Bristol, United Kingdom

[2]Thales, United Kingdom

For a reinforcement learning (RL) agent to reliably achieve a goal or desired behaviour, this objective must be encoded as a reward function. However, manual reward design is widely understood to be challenging, with mis-specification liable to yield undesirable, unsafe and variable outcomes. For this reason, there has been growing interest in enabling RL agents to learn their reward functions from normative feedback provided by humans. Such RL from human feedback (RLHF) methods have shown promise from a technical perspective, but an oft-unquestioned aspect of the approach creates a roadblock to practical applications: reward learning typically uses black-box neural networks, which resist human scrutiny and interpretation. For advocates of explainable AI (XAI), this is a problematic state of affairs. The XAI community is vocal about the safety and accountability risks of opaque learning algorithms, but an inability to interpret even the objective that an agent is optimising places us in yet murkier epistemic territory, in which an understanding of the causal origins of learnt behaviour, and its alignment with human preferences, is virtually unattainable. The importance of interpretability for RLHF has been highlighted in surveys, and some post hoc analysis has been applied to learnt rewards to gain some insight into feature influence, but to our knowledge, there have been no efforts to make the reward function intrinsically interpretable (loosely speaking, human-readable) by constraining its functional form.

We have developed an RLHF algorithm that learns intrinsically interpretable reward functions from human preferences over candidate agent behaviours. Specifically, it yields tree-structured reward functions (*reward trees*), formed of independent components associated with disjoint subsets of the state-action space, and defined hierarchically as a binary tree. The tree is incrementally refined as new preference labels arrive, and the traceability of these changes provides a powerful mechanism for monitoring and debugging. Reward trees afford both diagrammatic and geometric visualisation, textual description as a rule set in disjunctive normal form, and the efficient computation of feature importance metrics, all of which provide insight into the mechanisms and trends of agent learning.

In this talk, I will motivate the use of interpretable models in RLHF, before descibing our reward tree learning algorithm. I will then summarise the experiments that we have performed to date, which are the subject of two research papers. In the first paper [1], we evaluate our algorithm on four benchmark RL problems using both synthetic and human feedback, and in both offline and online learning settings. We observe effective and sample-efficient learning of reward trees in each of these contexts, alongside some informative failure cases. In more recent work [2], we adapt and extend the method (including by integrating it with model-based RL agents), and compare it to neural network-based reward learning in a challenging aircraft handling domain. We find it to be broadly competitive on both quantitative metrics and qualitative assessments, with a proposed modification to tree growth yielding significant improvements.



Example of a reward tree (**left**) learnt for the task of following a leader aircraft (**right**). High reward learnt for maintaining close distance; low reward for dropping below a safe altitude.

[1] T. Bewley and F. Lecue, "Interpretable preference-based reinforcement learning with tree-structured reward functions," in *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 118–126, 2022.

[2] T. Bewley, J. Lawry, A. Richards, R. Craddock, and I. Henderson, "Reward learning with trees: Methods and evaluation," *arXiv preprint arXiv:2210.01007*, 2022.