# HLG: Bridging Human Heuristic Knowledge and Deep Reinforcement Learning for Optimal Agent Performance

Extended Abstract

Bin Chen
University of South Australia
Adelaide, Australia
bin.chen@mymail.unisa.edu.au

Zehong Cao
University of South Australia
Adelaide, Australia
jimmy.cao@unisa.edu.au

## ABSTRACT

Training an optimal policy in deep reinforcement learning (DRL) remains a significant challenge due to the pitfalls of inefficient sampling in dynamic environments with sparse rewards. In this paper, we proposed a Human Local Guide (HLG) incorporating high-level human knowledge and local policies to guide DRL agents to achieve optimal performance. HLG deployed the heuristic rules from human knowledge in differential decision trees and then injected them into neural networks, which can continuously improve the suboptimal global policy till the optimal level. Our developed HLG includes action guides based on a policy-switching mechanism and adaptive action guides inspired by an approximate policy evaluation scheme through a perturbation model to optimise policy further. Our proposed HLG outperforms PPO and PROLONET with at least 25% improvement in training efficiency and exploration capability based on MinGrid environments with sparse reward signals. This implies that HLG has a significant potential to continuously assist the DRL agent in achieving optimal policy in dynamic and complex environments.

## KEYWORDS

Deep Reinforcement Learning; Local Guide; Human Knowledge; Training Efficiency; Differential Decision Trees

**ACM Reference Format:**
Bin Chen and Zehong Cao. 2024. HLG: Bridging Human Heuristic Knowledge and Deep Reinforcement Learning for Optimal Agent Performance: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

## 1 INTRODUCTION

Deep reinforcement learning (DRL) faces sampling efficiency problems [4] with sparse rewards in dynamic environments [7] because it requires a massive amount of interactions with environments under partial observability. Due to this inefficient data sampling issue, it is not easy to train a DRL agent to achieve optimal performance even after millions of steps [8–10].

Several prevailing methods have sought to alleviate this issue by leveraging pre-trained models [19] or harnessing guidance from human expertise [11, 17, 18]. However, these methods assume access to a guide that would be efficient in all the state space of the dynamic environment, defined as a global guide. Considering the increasing complexity of dynamic environments, such global information becomes too challenging to grasp [7]. Policy transferred from the pre-trained model or human demonstrations may fail in dynamic environments and lead to sub-optimal agent performance. This limitation is further exacerbated when agents intrinsically struggle to exploit state space with partial observability [15]. Nonetheless, salient stimuli can often still be extracted [16] by humans, regardless of the complexity of the environment. For instance, humans may build heuristic [2] such as *'defensive'* policies to avoid the agent trap into unsafe states, or *'attractive'* policies guide the agent towards the critical states that dominate access to the final reward.

In this work, we propose a novel human-in-the-loop DRL method, **H**uman **L**ocal **G**uide (HLG) that does not require frequent interactions between the DRL agent with dynamic environments, and mitigate the issue from sample inefficiency. The HLG is designed to translate high-level human knowledge into a local policy, and the DRL agent can leverage it with the approximate policy iteration scheme. The HLG injected human knowledge represented by heuristic rules into differentiable decision trees and thus transformed into policy networks [6] as the trainable local guide. The trainable local guide can generate local policies and not only guide the DRL agent at the initial training stage but also further optimise the local policy to provide continuous guidance in the remaining training stage. In the following sections, we briefly describe the HLG framework presents a human-in-the-loop method for solving dynamic environments and shows an experimental study.

## 2 DEEP REINFORCEMENT LEARNING WITH HUMAN LOCAL GUIDE (HLG)

This section briefly introduces a novel framework that harnesses high-level human knowledge to construct a local guide, aiding the DRL agent in optimising its global policy. Our proposed framework comprises a Human Local Guide (HLG) that provides two forms of guidance for a DRL agent.

### 2.1 From Human Knowledge to HLG

Humans can swiftly make decisions by transforming high-level knowledge into heuristic rules [2] represented using decision trees [18]. However, decision trees cannot provide continuous optimisation in combination with deep neural networks. Therefore, the

*Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand.* © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

HLG leverages Differential Decision Trees (DDTs) [5] to represent human knowledge for local state space that can warm up local guide policy. The local guide encodes the human local knowledge from DDTs into DRL policies by utilising DDTs as function approximations for the DRL policy networks [14]. The local guide contains a set of heuristic rules designed from high-level human knowledge. Our method towards translating heuristic rules into DDTs involves a procedure that retains the interpretability of the rules while allowing for the gradient-based optimisation inherent to DDTs. The local guide contains a set of heuristic rules designed from high-level human knowledge. It inputs feature data $\mathbf{x} = \{x_1, ..., x_n\}$ yields based on partial observation within the local state space $S_g$ and generates a guiding policy $\pi_g$. Each rule has the form of:

- *Rule L*: IF $x_1$ meets $D_1$ $(\neg D_1)$ and $x_2$ meets $D_2(\neg D_2)$ and ... $x_n$ meets $D_n(\neg D_n)$ THEN Action is $left(right)$

The $D_n$ is the decision criteria corresponding to the feature $x_n$. Our method towards translating heuristic rules into DDTs involves a procedure that retains the interpretability of the rules while allowing for the gradient-based optimisation inherent to DDTs.

### 2.2 Enhencing DRL with Local Action Guide (LAG)

A simple and intuitive way to build a local guide is to switch between the global and local policies by switch mechanism [3]. An indicator function $\rho$ is needed to provide the criteria for switching. The overall policy after $n$ iterations can be formulated as follows:

$$\pi^n_{LAG}(s) = \begin{cases} a^s_g & if \rho(s) \geq \rho^-, \\ \pi^n_\theta(s) & \text{otherwise.} \end{cases} \tag{1}$$

LAG would benefit from a performance jump start but could not outperform a suboptimal guide, so this local guide is only suitable for optimal pre-trained local policies.

### 2.3 Enhencing DRL with Adaptive Local Action Guide (ALAG)

While LAG can provide good initialisation for global policy, it does not enable the local guide to be continuously optimised to provide optimal performance. To obtain an adaptive improving policy, we develop a perturbation model $\Psi_\Phi(s, a^s_g, \Phi)$ which produces an adjustment to action $a^s_g$ in the range $[-\Phi, \Phi]$. The perturbation model allows agents to increase the diversity of guide policy actions $a^s_g$, thus exploring more states with salient stimuli rather than being strictly limited to safe actions $a^s_g$. The global policy $\pi_{ALAG}$ at $n$ can be formulated as:

$$\pi^n_{ALAG}(s) = \begin{cases} a^s_g + \beta^n_{ALAG}\Psi(s, a^s_g, \Phi) & if \rho(s) \geq \rho^-, \\ \pi^n_\theta(s) & \text{otherwise.} \end{cases} \tag{2}$$

where $\beta^n_{ALAG}$ is the weight of the perturbation model $\Psi_\Phi$. At the beginning of training, $\beta^n_{ALAG}$ should be set close to 0 to improve the initial performance of the local guide and then gradually increase to 1 to introduce a well-trained perturbation model.

The size of the parameter $\Phi$ can play a role in balancing safety and exploration. The threshold parameter $\Phi$ needs to be chosen by humans according to the environmental characteristics. The threshold parameter $\Phi$ should be small if the environment is high risk. If the environment does not contain risky states, then the threshold parameter $\Phi$ can be increased to get more action space that improves global policy. The perturbation model updating can be formulated as follows:

$$\Psi^{n+1}_\phi \leftarrow \arg\max_\phi \mathbb{E}_{s,a'} \left[\phi(s) Q^{n+1}(s, a^s_g + a^{'})\right] \tag{3}$$

## 3 EXPERIMENTS

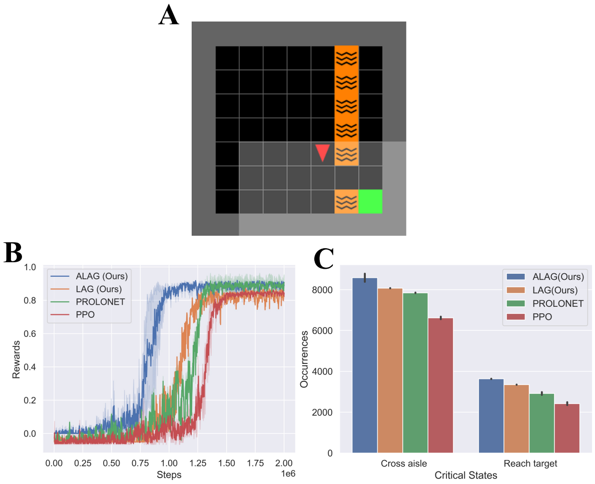

**Figure 1: Crossing task (A) in MiniGrid and performance of training efficiency (B) as well as exploration capability (C) in this task.**

We outline the experimental tasks designed based on the discrete control environment (MiniGrid [1] as shown in **Figure 1-A**). Task in MiniGrid with sparse rewards and partially observed setting. The red triangle represents an agent who can observe light grey shading regions and coloured objects, including lava and goal. In this study, we compare four methods across our experimental tasks. The first two methods are our ALAG and LAG, injected with the above human rules. We set the baseline method PROLONET [13] that encoded the same human rules. We also set PPO [12] as the baseline since the other three methods are all implemented based on PPO in the work. They are used to verify whether our local guides with human rules can improve training efficiency and exploration capability, as the results are shown in **Figure 1-B, and -C**, which indicates the best performance of our method.

## ACKNOWLEDGMENTS

This study is supported by Enterprise Research Scholarship UniSA Funding, partially supported by the UniSA Enterprise Fellowship and the Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA) Fellowship DE220100265. We are also grateful to the AAMAS program chairs and reviewers for their valuable comments, based on which our work was effectively revised.

## REFERENCES

[1] Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo de Lazcano, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. 2023. Minigrid & Miniworld: Modular & Customizable Reinforcement Learning Environments for Goal-Oriented Tasks. *CoRR* abs/2306.13831 (2023).

[2] Gerd Gigerenzer and Wolfgang Gaissmaier. 2011. Heuristic decision making. *Annual Review of Psychology* 62 (2011), 451–482.

[3] Sean Gillen, Marco Molnar, and Katie Byl. 2020. Combining deep reinforcement learning and local control for the acrobot swing-up and balance task. In *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 4129–4134.

[4] Zhiyu Huang, Jingda Wu, and Chen Lv. 2022. Efficient deep reinforcement learning with imitative expert priors for autonomous driving. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[5] Kelli D Humbird, J Luc Peterson, and Ryan G McClarren. 2018. Deep neural network initialization with decision trees. *IEEE Transactions on Neural Networks and Learning Systems* 30, 5 (2018), 1286–1295.

[6] Andrew Ilyas, Logan Engstrom, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. 2020. A Closer Look at Deep Policy Gradients. In *International Conference on Learning Representations*.

[7] Feibo Jiang, Kezhi Wang, Li Dong, Cunhua Pan, Wei Xu, and Kun Yang. 2020. AI driven heterogeneous MEC system with UAV assistance for dynamic environment: Challenges and solutions. *IEEE Network* 35, 1 (2020), 400–408.

[8] Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. 2020. Reinforcement learning with augmented data. *Advances in Neural Information Processing Systems* 33 (2020), 19884–19895.

[9] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533.

[10] Shubham Pateria, Budhitama Subagdja, Ah-hwee Tan, and Chai Quek. 2021. Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–35.

[11] Karl Pertsch, Youngwoon Lee, Yue Wu, and Joseph J Lim. 2022. Guided Reinforcement Learning with Learned Skills. In *Conference on Robot Learning*. PMLR, 729–739.

[12] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[13] Andrew Silva and Matthew Gombolay. 2021. Encoding human domain knowledge to warm start reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 5042–5050.

[14] Andrew Silva, Matthew Gombolay, Taylor Killian, Ivan Jimenez, and Sung-Hyun Son. 2020. Optimization methods for interpretable differentiable decision trees applied to reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1855–1865.

[15] Rodrigo Toro Icarte, Ethan Waldie, Toryn Klassen, Rick Valenzano, Margarita Castro, and Sheila McIlraith. 2019. Learning reward machines for partially observable reinforcement learning. *Advances in Neural Information Processing Systems* 32 (2019).

[16] Anne M Treisman and Garry Gelade. 1980. A feature-integration theory of attention. *Cognitive Psychology* 12, 1 (1980), 97–136.

[17] Min Wang, Xingzhong Wang, Wei Luo, Yixue Huang, and Yuanqiang Yu. 2022. Accelerating Deep Reinforcement Learning Under the Guidance of Adaptive Fuzzy Logic Rules. In *2022 Prognostics and Health Management Conference (PHM-2022 London)*. IEEE, 350–359.

[18] Peng Zhang, Jianye Hao, Weixun Wang, Hongyao Tang, Yi Ma, Yihai Duan, and Yan Zheng. 2020. KoGuN: accelerating deep reinforcement learning via integrating human suboptimal knowledge. *arXiv preprint arXiv:2002.07418* (2020).

[19] Zhuangdi Zhu, Kaixiang Lin, Anil K Jain, and Jiayu Zhou. 2023. Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).