Improvement and Evaluation of the Policy Legibility in Reinforcement Learning

Demonstration Track

Yanyu Liu
Xiamen University
Xiamen, China
yanyuliu@stu.xmu.edu.cn

Yifeng Zeng*
Northumbria University
Newcastle, United Kingdom
yifeng.zeng@northumbria.ac.uk

Biyang Ma
Minnan Normal University
Zhangzhou, China
mby@mnnu.edu.cn

Yinghui Pan*
Shenzhen University
Shenzhen, China
panyinghui@szu.edu.cn

Huifan Gao
Xiamen University
Xiamen, China
huifangao@stu.xmu.edu.cn

Xiaohan Huang
University of Liverpool
Liverpool, United Kingdom
xiaohan.huang20@gmail.com

ABSTRACT

When we work with intelligent agents, such as fighting a battle with other agents in computer games, it is difficult to achieve seamless collaboration if we can't figure out what the agents are doing. Especially in a complex problem domain, the agents are well trained and their actions could be too sophisticated to be comprehended by humans. In this article, we propose a novel reward shaping mechanism to improve the legibility of reinforcement learning that is used to train agents' policies. More importantly, we develop an interactive system to seek for users' evaluation of the policy legibility and show performance of the new learning approach.

KEYWORDS

Legibility; Reinforcement Learning; Reward Shaping

ACM Reference Format:

Yanyu Liu, Yifeng Zeng*, Biyang Ma, Yinghui Pan*, Huifan Gao, and Xiaohan Huang. 2023. Improvement and Evaluation of the Policy Legibility in Reinforcement Learning: Demonstration Track. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

In many practical scenarios, such as loading cargoes onto shelves with robots in a warehouse, or blowing up enemy *Ancients* with AI teammates in *DotA* games, agents and humans need to collaborate with each other in order to achieve a common goal. However, it would be challenging if humans/agents cannot figure out what their collaborators are doing. Their interaction will become difficult and even lead to mission failure. Hence it becomes necessary for agents to convey their intention timely to assist humans in predicting what they are doing as shown in Fig. 1, and increases human's confidence in interacting with the agents [15, 17, 18]. In other words, the agents need to perform in a legible way so that humans could understand their intentions during the action execution.

The previous research has shown its possibility of improving the policy legibility through different action choices. Dragan *et al.* [8] proposed a framework for quantifying legibility, which essentially exploited the assumption that an observer expects the agent to be rational and act efficiently or justifiably. Furthermore, they proposed a gradient optimization technique to generate legible robotic arm motion [5–7]. In reinforcement learning (RL), the existing work has succeeded in manually influencing the policy training process through policy shaping [3, 9], reward shaping [4, 11, 13]. Bied *et al.* [2] proposed the integration of observers into the RL framework through reward shaping, and designed the additional rewards from different perspectives, e.g. goal distance, cost of the observed trajectory in comparison with cost of the optimal trajectory, or a legibility function[8]. Persiani *et al.* [14] modelled the agent and observers as two equivalent Bayesian Networks (BN), and took the distance between the two BNs as the regularization term in the reward computation.

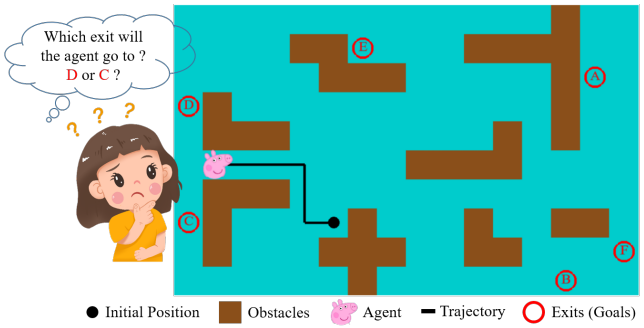


Figure 1: A multi-exits maze. The agent is initialized randomly in the maze and given a target exit, which is unknown to the observer. The easier the observer could predict the agent's intention (exit), the more legible the agent's policy is.

In this article, we propose a new way of generating legible policies for intelligent agents through a new reward shaping mechanism. From the observer's perspective, we use information entropy to measure the observer's uncertainty about the agent's goal, which acts as the reward shaping function in RL. By doing this, we encourage the agent to execute the actions that would reduce the entropy in a fast way. In particular, we implement an interactive system to seek for users' evaluation about the improved legibility in RL, which would inspire the legibility research in intelligent agents.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

2 INFORMATION GAIN BASED LEGIBILITY

We develop the legibility in a general RL framework that is defined with a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, R, \gamma)$, in which $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow P(s'|s, a)$ describes the state transition probability. A reward function $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the feedback signal for tuning agent's policy. A discount factor γ balances the immediate and future reward. In traditional RL, the agent's objective is to learn a policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ that maximizes the discounted cumulative reward. We adopt a general RL algorithm, namely the Q-Learning, to learn the value of an action in a particular state. Given the policy π, the expected discounted cumulative reward for state s and action a can be represented with the $Q(s, a)$. In every learning step, the agent selects the action a based on the current policy, receives the reward R, gets to the new state s' from s, and the the Q-value is updated with a *Bellman* equation:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[R + \gamma \times \max_{a' \in \mathcal{A}} Q(s',a') - Q(s,a)] \quad (1)$$

where α is the learning rate.

In this article, we introduce the novel Information Gain based Legibility (IGL) reward shaping into the Q-learning. Equation 2 shows the new reward R_L where R is the original reward signal, L is the legibility value for shaping the rewards, and β is a trade-off factor to accommodate their scale difference.

$$R_L = R + \beta \times L \quad (2)$$

In order to learn a legible policy, we design the function L that enables agents to take more legible actions by giving them higher reward when the actions let the observer more easily predict the agents' goals. To evaluate prediction uncertainty, we propose the legibility function L in Eq. 3

$$L = \mathcal{H}(\tau_t) - \mathcal{H}(\tau_t, a) \quad (3)$$

where $\mathcal{H}(\tau_t)$ is a measurement of information entropy that represents the uncertainty, from the observer's viewpoint, in predicting the agent's true goal given the agent's trajectory τ_t at time t. $\mathcal{H}(\tau_t, a)$ is the observer's uncertainty after the agent executes the action a following the trajectory τ_t. A low value of $\mathcal{H}$ means that the observer is confident about his/her prediction of the agent's goal. Intuitively, the larger $\mathcal{H}(\tau_t) - \mathcal{H}(\tau_t, a)$ (i.e. more entropy is reduced), indicates that the action a conveys more information to the observer about the agent's intention.

Subsequently, we implement $\mathcal{H}(\tau_t)$ and $\mathcal{H}(\tau_t, a)$ respectively in Eqs. 4-5. In both equations, we transform the observer's belief in each goal into the entropy calculation.

$$\mathcal{H}(\tau_t) = -\sum_{g \in \mathcal{G}} b(g|\tau_t) log(b(g|\tau_t)) \quad (4)$$

$$\mathcal{H}(\tau_t, a) = \sum_{s_{t+1} \in \mathcal{S}} T^o(s_{t+1}|s_t, a) \mathcal{H}(\tau_t \circ s_{t+1}) \quad (5)$$

where $b(g|\tau_t)$ is the conditional probabilities of the observer's prediction in the agent's goal g, and $\mathcal{G}$ is the set of all potential goals of agent.

Apparently, the accuracy of $b(\cdot)$ directly contributes to the legibility computation - even the RL convergence. In principle, any function which maps a policy to the agent's beliefs can be used to update b [10, 12].

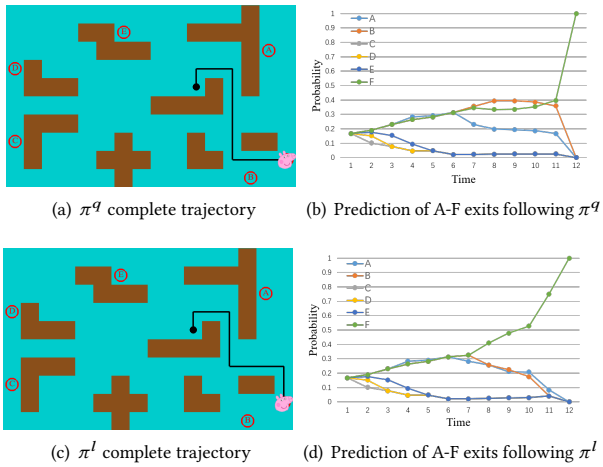
We choose the Bayesian method in the belief update. Inspired by the work of Baker *et al.* [1], we update the posterior beliefs in Eq. 6.

$$b(g|\tau_t \circ s_{t+1}) = \frac{T^o(s_{t+1}|a_t, s_t)\pi^o(s_t, a_t|g)}{\sum_{g' \in \mathcal{G}} T^o(s_{t+1}|a_t, s_t)\pi^o(s_t, a_t|g')} b(g|\tau_t) \quad (6)$$

where T^o is the transition probability in the observer's mind, and π^o is the legible policy recognized by the observer. It assumes that agent's actions can be seen by the observer.

3 EXPERIMENTS AND DEMONSTRATION

We conduct the experiments in a maze-like multi-goals environment in Fig. 1. We use a solid black line to indicate the trajectory. We firstly train the policy π^q with the Q-Learning as a baseline, and the legible policy π^l with the reward shaping. For simplicity we set the $\pi^o = \pi^q$ in updating the observer's belief in Eq. 6. Figure 2



(a) π^q complete trajectory (b) Prediction of A-F exits following π^q

(c) π^l complete trajectory (d) Prediction of A-F exits following π^l

Figure 2: Comparison of the trajectory legibility and prediction from the Q-learning and IGL methods

demonstrates the complete trajectories from the two policies. In Step 7 to 9 of the π^q trajectory (Fig. 2(a)), the observer is very likely to misjudge that the agent will go to exit B and does not realize that the agent's real goal is exit F until Step 11. Such behavior can obviously surprise the observer. In contrast, the π^l trajectory (Fig. 2(c)) convinces the observer that the agent's goal in Step 8. We can find in Fig. 2(d) the probability that the observer predicts agent is going to the exit F rises rapidly after Step 8, compared to the prediction probability of F when the plain Q-learning is used to generate the trajectory π^q in Fig. 2(b). In addition, we implement an interactive system [1] [2] that allows an observer to evaluate the legibility of the trained policies in this environment and will seek for the live evaluation from the AAMAS audience.

4 CONCLUSION

We propose a new reward shaping function to improve the RL legibility and develop one interactive evaluation system to demonstrate the performance of the new legibility research. Using the IGL function improves the learned policy, which is also verified in humans' evaluation. The demonstration system will interact with the audience to evaluate the policy legibility. The interaction data could be used to further improve the legibility formulation. Our work expects to elicit further research on legible plans in multiagent systems [16] and contribute to explainable AI research.

[1] Video: https://youtu.be/Ow6hFLs3O2U

[2] Download: https://pan.baidu.com/s/1irGrO$_V$f8YQX9cfA − y71WQ?pwd = 7dly

5 ACKNOWLEDGMENTS

This work is supported in part by the National Natural Science Foundation of China (Grants No. 61836005, 62176225 and 62276168). Yinghui is also supported by the Natural Science Foundation of Guangdong Province (2023A1515010869).

REFERENCES

[1] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. 2009. Action understanding as inverse planning. *Cognition* 113, 3 (2009), 329–349.
[2] Manuel Bied and Mohamed Chetouani. 2020. Integrating an observer in interactive reinforcement learning to learn legible trajectories. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 760–767.
[3] Thomas Cederborg, Ishaan Grover, Charles L Isbell, and Andrea L Thomaz. 2015. Policy shaping with human teachers. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.
[4] Sam Michael Devlin and Daniel Kudenko. 2012. Dynamic potential-based reward shaping. In *Proceedings of the 11th international conference on autonomous agents and multiagent systems*. IFAAMAS, 433–440.
[5] Anca Dragan and Siddhartha Srinivasa. 2013. Generating legible motion. (2013).
[6] Anca D Dragan. 2017. Robot planning with mathematical models of human state and action. *arXiv preprint arXiv:1705.04226* (2017).
[7] Anca D Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S Srinivasa. 2015. Effects of robot motion on human-robot collaboration. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 51–58.
[8] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 301–308.
[9] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems* 26 (2013).
[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[11] Yujing Hu, Weixun Wang, Hangtian Jia, Yixiang Wang, Yingfeng Chen, Jianye Hao, Feng Wu, and Changjie Fan. 2020. Learning to utilize shaping rewards: A new approach of reward shaping. *Advances in Neural Information Processing Systems* 33 (2020), 15931–15941.
[12] Shuwa Miura and Shlomo Zilberstein. 2021. A unifying framework for observer-aware planning and its complexity. In *Uncertainty in Artificial Intelligence*. PMLR, 610–620.
[13] Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, Vol. 99. 278–287.
[14] Michele Persiani and Thomas Hellström. 2022. Policy Regularization for Legible Behavior. *arXiv preprint arXiv:2203.04303* (2022).
[15] Alessandra Sciutti, Martina Mara, Vincenzo Tagliasco, and Giulio Sandini. 2018. Humanizing human-robot interaction: On the importance of mutual understanding. *IEEE Technology and Society Magazine* 37, 1 (2018), 22–29.
[16] Miura Shuwa and Zilberstein Shlomo. 2020. Maximizing Plan Legibility in Stochastic Environments. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*. 1931–1933.
[17] Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. 2021. Explainable embodied agents through social cues: a review. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 3 (2021), 1–24.
[18] Zhang Zhang, Yifeng Zeng, Wenhui Jiang, Yinghui Pan, and Jing Tang. 2023. Intention Recognition for Multiple Agents. *Information Sciences* DOI: 10.1016/j.ins.2023.01.066 (2023).