DSSV 2023
Data Science, Statistics & Visualisation
ECDA 2023
European Conference on Data Analysis

Evaluating the Risk Alignment of Preference-Based Reinforcement Learning Agents

M. Schweizer[1] and A. Geyer-Schulz[1]

[1]Karlsruhe Institute of Technology, Department of Department of Economics and Management, Karlsruhe, Germany

April 26, 2023

Preference-based reinforcement learning (PbRL), also known as Reinforcement Learning from Human Feedback (RLHF), has recently gained popularity as a tool for fine-tuning agent policies, for example in `ChatGPT` and `InstructGPT` [1]. More broadly, PbRL is a promising approach to AI alignment, which is the challenge of building intelligent agents that share and effectively pursue their user's preferences. A crucial aspect of these preferences, in domains such as healthcare, autonomous driving, and finance, is the user's attitude towards risk. So in order to achieve genuine AI alignment, an agent must act in accordance with the user's risk preference. In short, the agent should be risk-aligned.

This work first develops the concept of risk alignment for PbRL and inverse reinforcement learning, relating it to risk-sensitive reinforcement learning. On the empirical side, we present a set of experiments in simple gridworld environments that can be used to evaluate the properties of PbRL agents with respect to risk alignment and apply these experiments to the well-known PbRL agent by [2].

Based on these empirical results and theoretical considerations, we discuss shortcomings of current state-of-the-art PbRL agents. For example, the collected preference data may not contain risk information, due to the particular preference query structure and the underlying reinforcement learning algorithm does not explicitly incorporate a notion of risk. This motivates a number of potential improvements for risk aligned PbRL, which conclude this work.

Keywords: Preference-Based Reinforcement Learning (PbRL); Reinforcement Learning from Human Feedback (RLHF); Reinforcement Learning from Human Preferences (RLHP); Risk-Sensitivity; AI Alignment

[1] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, and Others, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 27730–27744, Curran Associates, Inc., 2022.

[2] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.