Reward Hacking Mitigation

Joe Gaucho
UC Santa Barbara
jgaucho@ucsb.edu

Joe Gaucho
UC Santa Barbara
jgaucho@ucsb.edu

Abstract

Choosing good reward functions in reinforcement learning (RL) is notoriously difficult. Oftentimes, the true reward function is very sparse, as in a game of chess that gives a reward signal only when the agent wins. In other scenarios, such as preference optimization for large language models (LLMs) using algorithms like reinforcement learning from human feedback (RLHF) [2], the true reward function—alignment to human preferences—is impossible to specify. As a result, RL techniques typically employ proxy rewards, which provide finer-grained feedback loops and are easier to learn. However, these proxy rewards can be misspecified; RL agents that exploit misspecifications in the proxy reward function can exhibit undesirable and potentially harmful behaviors. According to [4], this type of behavior—where an agent attains a high proxy reward but does not accomplish the human-intended goal—is referred to as *reward hacking*. In this work, we investigate the problem of reward hacking and propose a novel approach that improves the approximation of the true reward function by incorporating feedback from Vision-Language Models (VLMs) into the training loss.

1 Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.

2 Background

Example citations [1] [3] [4]

Unlabeled Section

Lorem

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[2] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023.

[3] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models, 2022.

Preprint. Under review.

[4] Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking, 2022.