Reward Hacking Mitigation

Joe Gaucho
UC Santa Barbara
jgaucho@ucsb.edu

Joe Gaucho
UC Santa Barbara
jgaucho@ucsb.edu

Abstract

Reward hacking is a significant concern in the training of advanced reinforcement learning (RL) agents. In complex environments and tasks, it is challenging to specify a perfect reward function that fully captures the intended human behaviors across all edge cases. RL agents that exploit misspecifications in the reward function can exhibit undesirable and potentially harmful behaviors. 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] [2] [3]

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

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

Preprint. Under review.