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 [5], 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

Reinforcement learning has led to breakthroughs in areas such as robotics, game-playing, and LLMs. However, one of its fundamental challenges is the design of reward functions that effectively capture the human-intended objectives of a task. In many cases, the true reward function is either sparse or difficult to specify explicitly, necessitating the use of proxy rewards. Use of such proxies, however, introduces the concept of reward hacking, where an agent discovers unintended strategies to maximize reward without accomplishing the underlying goal.

Reward hacking can manifest in both benign and harmful ways. In some cases, agents discover novel but unintended strategies that still achieve high rewards, such as finding new methods for robot locomotion. However, it is more problematic when agents exploit bugs, manipulate physics engines, cheat, or even engage in deceptive behavior [3]. Mitigating reward hacking is essential for deploying RL in high-trust settings, where unintended behaviors could compromise safety, reliability, or ethical standards.

In this work, we propose an approach that leverages *Vision-Language Models (VLMs)* to mitigate reward hacking at training time. VLMs, which process both visual and textual information, have shown strong generalization capabilities in tasks requiring multimodal reasoning. We explore how feedback from VLMs can be integrated into the RL training pipeline to improve alignment between proxy rewards and the true underlying objective. By incorporating VLM-based signals into the training loss, we aim to reduce instances of reward hacking and enhance the overall robustness of RL agents.

Preprint. Under review.

2 Background

Example citations [1] [4] [5]

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] Joel Lehman, Jeff Clune, Dusan Misevic, Christoph Adami, Lee Altenberg, Julie Beaulieu, Peter J. Bentley, Samuel Bernard, Guillaume Beslon, David M. Bryson, Patryk Chrabaszcz, Nick Cheney, Antoine Cully, Stéphane Doncieux, Fred C. Dyer, Kai Olav Ellefsen, Robert Feldt, Stephan Fischer, Stephanie Forrest, Antoine Frénoy, Christian Gagné, Leni K. Le Goff, Laura M. Grabowski, Babak Hodjat, Frank Hutter, Laurent Keller, Carole Knibbe, Peter Krcah, Richard E. Lenski, Hod Lipson, Robert MacCurdy, Carlos Maestre, Risto Miikkulainen, Sara Mitri, David E. Moriarty, Jean-Baptiste Mouret, Anh Nguyen, Charles Ofria, Marc Parizeau, David P. Parsons, Robert T. Pennock, William F. Punch, Thomas S. Ray, Marc Schoenauer, Eric Schulte, Karl Sims, Kenneth O. Stanley, François Taddei, Danesh Tarapore, Simon Thibault, Westley Weimer, Richard A. Watson, and Jason Yosinski. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *CoRR*, abs/1803.03453, 2018.

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

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