Learning & Adaptive Systems

ETH zürich

Addressing Reward Hacking in RLHF with Causality

Project Proposal for Master Thesis

Project description

Reward hacking [Skalse et al., 2022] is a critical problem in AI alignment, particularly in reinforcement learning (RL). AI systems designed to optimize for a specific reward often discover unintended ways to maximize this reward, diverging from human intentions. This misalignment between the true objective and the model's learned behavior can lead to unsafe or undesirable outcomes. Addressing reward hacking is essential for building AI systems that align reliably with human values.

A major source of reward hacking in Reinforcement Learning with Human Feedback (RLHF) is causal misidentification [Tien et al., 2022]. This occurs when a model incorrectly learns the causal relationships between actions and rewards, leading it to optimize for proxies or spurious correlations instead of the true goals. For example, the model might manipulate metrics or exploit shortcuts in its environment. This creates a scenario where the AI appears successful according to the reward function but fails to fulfill the intended objective.

This project aims to explore whether accurately identifying causal mechanisms within a reward model can help mitigate reward hacking. By modeling the causal relationships that drive desirable behaviors, we hope to guide the AI towards more aligned learning. Specifically, the project will investigate methods for integrating causal inference into reward modeling to improve RLHF robustness, reducing the risk of AI exploiting unintended loopholes. The goal is to understand how causal reasoning can contribute to better alignment of AI systems with human values.

Ideal Candidate

A good candidate will look like someone who has:

- Experiences running reinforcement learning experiments (e.g. using Stable Baselines [Raffin et al., 2021]);
- Knowledge in causality; and
- Strong interests in the AI safety/alignment problem.

Contact

If you are interested in this project, please contact Xin Chen (Cynthia) through email xin.chen@inf.ethz.ch. Please attach your resume, transcript of records, and briefly state why you are interested in this project.

References

[Raffin et al., 2021] Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. (2021). Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8.

[Skalse et al., 2022] Skalse, J., Howe, N., Krasheninnikov, D., and Krueger, D. (2022). Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471.

[Tien et al., 2022] Tien, J., He, J. Z.-Y., Erickson, Z., Dragan, A. D., and Brown, D. S. (2022). Causal confusion and reward misidentification in preference-based reward learning. *arXiv preprint arXiv:2204.06601*.