Natural Language Interfaces and Reward Hacking in RL

UvA MSc AI - Thesis Proposal - Supervision: Niklas Höpner

Giulio Starace - 13010840

October 20, 2022

The problem

In Reinforcement Learning (RL), the goal of a particular autonomous agent is formalised in the form of a reward signal emitted by the environment to the agent. This reward signal is typically computed via some handcrafted reward function. However, handcrafted reward functions can be difficult to specify for more complex problems and environments, and can lead to undesired agent behaviour due to reward hacking [26, 34].

Why we want to solve it

Addressing the issue of reward misspecification is important because it is one of the many limiting factors that make RL difficult to apply. Furthermore, due to reward hacking, the issue can lead to undesired behaviour. The negative impacts of misbehaviour can be as simple as a model underperforming in production and as dire as causing safety concerns [16].

Current solutions and their shortcomings

Inverse Reinforcement Learning

Inverse Reinforcement Learning (IRL) [17, 25, 43] is the problem of extracting a reward function given observed expert behaviour (demonstrations). While promising and perhaps suitable for many problems, IRL has some limitations. For instance expert demonstrations are not always available and can be difficult to obtain. Furthermore, for many environments it is very difficult to determine the reward function from the demonstrations [2, 7]. Another limitation is that model performance may be limited to the performance of the experts from which it is learning [11, 12]. IRL is often also criticised for overlooking side-effects [22] and encouraging power-seeking [36]. Even if these issues were addressed, IRL does not necessarily address the overarching problem, as reward hacking has been observed in the IRL context as well [19].

In general, IRL is considered to be a subfield of ***imitation learning*** [1], where the goal is now to predict trajectories, given expert demonstrations. Imitation learning faces similar limitations to those of IRL.

Preference-based learning

Preference-based learning circumvents the need for demonstrations by using a more direct signal of human preferences. This includes, for example, directly asking users what they want via e.g. pairwise comparisons [4, 8, 30]. The main approach is to express preferences via pairwise comparison. This however can be limited in expressivity: consider a case in which two sub-optimal but complimentary correct trajectories are presented. Under pairwise comparison, there is no way to express the necessary granularity in preferences for this example. Another potential issue is that the expressed preferences may be different from the real preferences.

Proposed approach

Using advances in natural language processing, particularly in large language models (LLMs) [6, 31, 37] and prompting techniques [14, 28, 38], and inspired by their applications beyond a pure NLP context [10, 27, 29], we can develop a more natural interface between human and machine to specify goals and or rewards. This is after-all how humans communicate desired outcomes to each other. There already exist many recent works leveraging the expressivity of language models in an RL context [5, 9, 15, 18, 20, 24, 32, 33, 35, 39, 40, 42]. A number of NL-RL-hybrid environments and datasets [3, 13, 21, 23, 41] have accompanied many of these papers in the field. These works however mostly focus on their contributions to planning performance, learning efficiency and other more common RL metrics of success. Using techniques similar to those developed by [26] and taking inspiration from the recent works cited above, this work hopes to explore the question: **to what extent can natural language interfaces curtail the issue of reward hacking in RL?**

References

[1] Abbeel, P. and Ng, A.Y. 2004. Apprenticeship learning via inverse reinforcement learning. *Proceedings of the twenty-first international conference on Machine learning* (New York, NY, USA, Jul. 2004), 1.

[2] Amin, K. and Singh, S. 2016. Towards Resolving Unidentifiability in Inverse Reinforcement Learning. arXiv.

[3] Anderson, P. et al. 2018. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. (2018), 3674–3683.

[4] Biyik, E. and Sadigh, D. 2018. Batch Active Preference-Based Learning of Reward Functions. *Proceedings of The 2nd Conference on Robot Learning* (Oct. 2018), 519–528.

[5] Brooks, E. et al. 2022. In-Context Policy Iteration. arXiv.

[6] Brown, T.B. et al. 2020. Language Models are Few-Shot Learners.

[7] Choi, J. and Kim, K.-E. 2011. Inverse Reinforcement Learning in Partially Observable Environments. *Journal of Machine Learning Research.* 12, 21 (2011), 691–730.

[8] Christiano, P. et al. 2017. Deep reinforcement learning from human preferences. arXiv.

[9] Ding, Y. et al. 2022. Robot Task Planning and Situation Handling in Open Worlds. arXiv.

[10] Dosovitskiy, A. et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv.

[11] Evans, O. et al. 2015. Learning the preferences of bounded agents. (2015).

[12] Evans, O. et al. 2015. Learning the Preferences of Ignorant, Inconsistent Agents. arXiv.

[13] Fan, L. et al. 2022. MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge. arXiv.

[14] Gal, R. et al. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. arXiv.

[15] Gramopadhye, M. and Szafir, D. 2022. Generating Executable Action Plans with Environmentally-Aware Language Models. arXiv.

[16] Hendrycks, D. et al. 2022. Unsolved Problems in ML Safety. arXiv.

[17] Ho, J. and Ermon, S. 2016. Generative Adversarial Imitation Learning. *Advances in Neural Information Processing Systems* (2016).

[18] Huang, W. et al. 2022. Inner Monologue: Embodied Reasoning through Planning with Language Models. arXiv.

[19] Ibarz, B. et al. 2018. Reward learning from human preferences and demonstrations in Atari. arXiv.

[20] Jiang, Y. et al. 2022. VIMA: General Robot Manipulation with Multimodal Prompts. arXiv.

[21] Jiang, Y. et al. 2022. VIMA: General Robot Manipulation with Multimodal Prompts. Zenodo.

[22] Krakovna, V. et al. 2019. Penalizing side effects using stepwise relative reachability. arXiv.

[23] Liu, E.Z. et al. 2018. Reinforcement Learning on Web Interfaces Using Workflow-Guided Exploration. arXiv.

[24] Lu, Y. et al. 2022. Neuro-Symbolic Procedural Planning with Commonsense Prompting. arXiv.

[25] Ng, A.Y. and Russell, S. 2000. Algorithms for Inverse Reinforcement Learning. *In Proc. 17th International Conf. On Machine Learning* (2000), 663–670.

[26] Pan, A. et al. 2022. The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. arXiv.

[27] Ramesh, A. et al. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv.

[28] Reynolds, L. and McDonell, K. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. arXiv.

[29] Rombach, R. et al. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. (2022), 10684–10695.

[30] Sadigh, D. et al. 2017. *Active preference-based learning of reward functions*.

[31] Sanh, V. et al. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization.

[32] Shridhar, M. et al. 2021. CLIPort: What and Where Pathways for Robotic Manipulation. arXiv.

[33] Shridhar, M. et al. 2022. Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation. arXiv.

[34] Skalse, J. et al. 2022. Defining and Characterizing Reward Hacking. arXiv.

[35] Sumers, T.R. et al. 2021. Learning Rewards from Linguistic Feedback. arXiv.

[36] Turner, A.M. et al. 2021. Optimal Policies Tend to Seek Power. arXiv.

[37] Vaswani, A. et al. 2017. Attention Is All You Need.

[38] Wu, T. et al. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, Apr. 2022), 1–22.

[39] Yao, S. et al. 2022. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv.

[40] Yu, A. and Mooney, R.J. 2022. Using Both Demonstrations and Language Instructions to Efficiently Learn Robotic Tasks. arXiv.

[41] Zholus, A. et al. 2022. IGLU Gridworld: Simple and Fast Environment for Embodied Dialog Agents. arXiv.

[42] Zhou, L. and Small, K. 2020. Inverse Reinforcement Learning with Natural Language Goals. arXiv.

[43] Ziebart, B.D. et al. 2008. Maximum entropy inverse reinforcement learning. *Proceedings of the 23rd national conference on Artificial intelligence - Volume 3* (Chicago, Illinois, Jul. 2008), 1433–1438.