Natural Language Interfaces and Reward Hacking in RL

CHAI Internship - Research Proposal

Giulio Starace

November 10, 2022

The following is the research proposal for my AI master's thesis at the University of Amsterdam. The topic has been accepted. My internal supervisor is Niklas Höpner of the AMLab, who has seconded my application to CHAI. Part-time work in the form of a literature review and brainstorming has begun in early November, with full-time work commencing in February and an expected completion date of August 25^{th}, 2023. There will be a month-long pause in the work in December due to a course I will be attending full-time.

At CHAI, I hope to receive mentorship from experts in AI Safety (AIS). While my local supervision is of great value, the staff at my institute are not very familiar with AIS. I hope that my CHAI mentor would be able to guide me in making the right mental connections and finding appropriate citations that I may have otherwise missed. I also hope to receive more classical mentorship in the form of additional perspectives and creative approaches to the problem. Ultimately the goal is to produce a piece of research worthy of peer review and publication and to connect with more people in AIS. AIS is the direction I would like to pursue in my career.

Finally, I should note that I am generally curious about alternative approaches to Reinforcement Learning that address the issue of reward hacking. I am also interested in human-AI interface design, currently seemingly dominated by prompting which is where most of my experience lies outside of AI safety. I developed the proposal below because it captured both interests while remaining flexible in terms of what can be contributed and to potential topic pivots. I should note that both my supervisor and I are open to adapting the topic to something similar that may be more suitable for CHAI mentorship.

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 [27, 36].

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, 26, 46] 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 [38]. Regardless of these issues, IRL does not necessarily address the overarching problem, as reward hacking has also been observed in this context [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, 23, 31]. 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 *how* is difficult to express.

Proposed approach

Using advances in natural language processing, particularly in large language models (LLMs) [6, 32, 39] and prompting techniques [14, 29, 41], and inspired by their applications beyond a pure NLP context [10, 28, 30], 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, 25, 34, 35, 37, 40, 42, 43, 45]. A number of NL-RL-hybrid environments and datasets [3, 13, 21, 24, 44] 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 [27] and [23] 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?**

Potential outcomes and focus points

This proposal has mostly motivated the problem from the perspective of Reinforcement Learning. In this sense, one of the potential outcomes to focus on is an alternative method for addressing reward misspecification and hacking in RL.

However, the research could also pivot towards investigating ways of improving the ability of language models in understanding human instructions. This would be similar to the work of [33], with perhaps more attention to the safety implications. Prompts currently dominate interfaces through which humans can guide capabilities, but can we develop better and safer interfaces? Human-AI interface design may be a fruitful area of AI alignment research.

A sub-problem that needs addressing in integrating language feedback into an RL policy is that of grounding the language to the environment. A contribution in this area could also be made, although it is not immediately clear how this directly contributes to safety.

Of course, the holy grail would be contributions to all of these areas in a unified solution. By casting a wider net at the beginning, the hope is that at least one of these can be achieved.

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.F. et al. 2017. Deep Reinforcement Learning from Human Preferences. *Advances in Neural Information Processing Systems* (2017).

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

[10] Dosovitskiy, A. et al. 2022. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. (Mar. 2022).

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

[12] Evans, O. et al. 2016. Learning the preferences of ignorant, inconsistent agents. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (Phoenix, Arizona, Feb. 2016), 323–329.

[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. *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, Dec. 2018), 8022–8034.

[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] Lee, K. et al. 2021. PEBBLE: Feedback-Efficient Interactive Reinforcement Learning via Relabeling Experience and Unsupervised Pre-training. *Proceedings of the 38th International Conference on Machine Learning* (Jul. 2021), 6152–6163.

[24] Liu, E.Z. et al. 2022. Reinforcement Learning on Web Interfaces using Workflow-Guided Exploration. (Feb. 2022).

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

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

[27] Pan, A. et al. 2022. The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. (Feb. 2022).

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

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

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

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

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

[33] Scheurer, J. et al. 2022. Training Language Models with Language Feedback. arXiv.

[34] Shridhar, M. et al. 2021. CLIPort: What and Where Pathways for Robotic Manipulation. (Nov. 2021).

[35] Shridhar, M. et al. 2022. Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation. (Sep. 2022).

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

[37] Sumers, T.R. et al. 2021. Learning Rewards From Linguistic Feedback. *Proceedings of the AAAI Conference on Artificial Intelligence*. 35, 7, 7 (May 2021), 6002–6010. DOI:https://doi.org/10.1609/aaai.v35i7.16749.

[38] Turner, A.M. et al. 2022. Optimal Policies Tend To Seek Power. (Jan. 2022).

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

[40] Watkins, O. et al. 2021. Teachable Reinforcement Learning via Advice Distillation. *Advances in Neural Information Processing Systems* (2021), 6920–6933.

[41] 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.

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

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

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

[45] Zhou, L. and Small, K. 2021. Inverse Reinforcement Learning with Natural Language Goals. *Proceedings of the AAAI Conference on Artificial Intelligence*. 35, 12, 12 (May 2021), 11116–11124. DOI:https://doi.org/10.1609/aaai.v35i12.17326.

[46] 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.