Natural Language Interfaces for Specification Learning

UvA MSc AI - Thesis Proposal

Giulio Starace - 13010840

October 12, 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 [14].

Why We Want To Solve It

Relying on handcrafted reward functions can be tedious, requiring at times ample domain knowledge and mental effort. Furthermore, after design, the reward function has to be manually implemented as part of the agent's environment. Finally, handcrafted reward functions may suffer from bias and human error, leading to subpar or undesired performance of our models. Generally, these are symptoms signaling difficulty in scaling and generalisation. In the case of undesired model performance, this has safety implications [11].

Current Solutions and their Shortcomings

Inverse Reinforcement Learning

Inverse Reinforcement Learning (IRL) [12, 13, 22] is the problem of extracting a reward function given observed expert behaviour (demonstrations). While promising and perhaps suitable for many problems, IRL presents some limitations:

- Expert demonstrations are not always available and can be difficult to obtain
- For many environments it is very difficult to determine the reward function from the demonstrations.
  - There is some research addressing this issue [2, 5].
- Model performance may be limited to the performance of the experts from which it is learning.
  - There is some research addressing this issue [8, 9].
- Natural intelligent agents (e.g. humans) don't always need expert demonstrations to learn a reward function, so this is indicative of a lack of generalisation.

IRL has a considerable overlap with ***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 [3, 6, 18].

- Expression of preferences via pairwise comparison can be limited.
- Expressed preferences may be different from real preferences.

Proposed Approach

Using advances in natural language processing, particularly in large language models (LLMs) [4, 19, 20] and prompting techniques [10, 16, 21], and inspired by their applications beyond a pure NLP context [7, 15, 17], 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.

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] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[21] 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.

[22] 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.