Define Agentic Preferences

This project relates to the Theoretical Foundations of Reward Learning research agenda. You can find more information about this research agenda here (link), including a broader description of the aims and background assumptions behind this research. This project would be mainly theoretical/mathematical.

Overview:

The aim of this research project is to formulate a mathematical definition of how "agentic" a given reward function is, and to investigate the properties of this definition.

Background:

Some people have the intuition that we can distinguish between "agentic" and "non-agentic" preferences, where agentic preferences give rise to convergent instrumental goals, but non-agentic preferences do not. For example, the task of creating as many paperclips as possible may be viewed as being quite agentic, whereas the task of solving a given mathematical problem is quite non-agentic. It would be useful if this distinction could be formalised mathematically. Given such a definition, we may then be able to develop methods that could steer a reward learning algorithm (such as RLHF) away from agentic preferences and towards non-agentic preferences, or it may suggest methods for determining if a given reward model is likely to induce agentic behaviour or not.

Methodology:

There are many ways to attempt to quantify how agentic a reward function is. For example, in a Markov Decition Process (MPD), the value $Q(s, a)$ of taking a particular action a in a given state s can be decomposed as

$$Q(s, a) = \mathbb{E}_{S' \sim T(s,a)} \left[R(s, a, A') + \gamma \cdot V(S') \right],$$

where R is the reward function maximised by the agent, T is the transition function of the MDP, γ is the discount factor of the agent, and $V(s)$ describes the expected cumulative discounted reward of the agent conditional on starting in state s. We may then observe that if $\gamma \cdot V(S')$ typically dominates $R(s, a, A')$ in the equality above, then the agent will typically pick its actions based on long-term considerations rather than short-term considerations, and vice versa. Perhaps this observation can be used to quantify whether or not R induces "agentic" preferences? Other alternatives include simply quantifying how sparse R is, etc. Given a candidate definition, we can then test experimentally if it seems to fit our intuitive judgements regarding how agentic a given reward function is. Moreover, given a formal definition, we can also attempt to derive additional results about these reward functions. For example, are most reward learning methods likely to produce reward functions that encode agentic preferences, or non-agentic preferences?