1 Composing Belief States in RL

1.1 Motivation

Current safety alignment is critically reliant on *post*-training. Post-training, in turn, is largely based on reinforcement learning. Moreover, this reliance has increased with the advent of inference-time scaling and is likely to increase further.

For our interpretability tools to be relevant, they therefore have to deal with this paradigm. There are however, no clearly understood models of reinforcement learning.

1.2 Basic idea

We want to study how belief states change under reinforcement learning. Although there are a number of different setups in which to do this, the one that seemed to be most concrete to me is studying compositionality under RL. In particular, given a pre-trained model that has learned different skills, we want to understand how they are combined together to form new skills under RL. Our basic question is:

"Can we detect when RL is composing belief states into something novel vs. merely reweighting them?"

1.3 Simple example

A very simple example could be logical operations. The IMPLIES conditional is equivalent to a combination of NOT and AND operations. This means we could study a case in which:

1. We pre-train a model on a HMM simultating noisy AND, and probe for the belief state.
2. We pre-train a model on the HMM simulating noise NOT, and probe for the belief state. We should expect that the
3. We make the reward function the results of applying IMPLIES and RL train the

This example is too simple to actually work, so a slightly more complicated case could be:

1. Pre-train 1: Train on Mess3
2. Pre-traing 2: Train on another Mess process or RRXOR.
3. Post-train: Train on a value function that requires the composition of both.

1.4 Safety relevance

There are a number of safety relevant applications that could arise from this toy model.

1. Capability auditing:
   Can we find belief state signatures for when a model acquires a genuinely new capability vs when it is eliciting existing capabilities?

2. Alignment robustness/Emergent misalignment:
   Emergent misalignment arises because alignment seems to be a general concept rather than particular to each misalignment instance. This may mean that we can identify misalignment as a 'factored' belief state that can be easily accessed. If this is true, it should make emergent misalignment easy to turn on/off (already the case via steering vectors, but we want to see if we can find a mechanistic understanding of why it is easy to access). We can then study whether other RL processes can make misalignment harder to access.

3. Model Personas:
   Can we think of personas as different elements that can be composed via RL? If so, can we study how RL changes the model persona?

1.5 Alternative settings for belief states in RL

1. Chain-of-thought monitoring: Can we use belief states to help detect when a model is doing unfaithful reasoning?
   Do we have a good toy model for chain of thought reasoning?

2. Steganography: Similar to the above, can we use belief states to help us identify steganography?

3. Rare capabilities in RL: We could study a factored belief process in which one of the factors appears only when a certain token is observed for the other process. If we allow the model to select which token it sees, it should be able to observe the rarer process. We can use this as a model for how RL can access rare behaviour in LLMs.