Belief States in RL

Contents

Basic idea: Extend the belief state framework to post-training. To do this, we will have to understand how belief states change under RL and whether this is qualitatively different from how belief states change in supervised learning (i.e pre-training). There are a number of cases where we may expect to find this qualitatively different behaviour:

1. Compositionality: We have some evidence that reasoning models functioning by eliciting already existing
2. Emergent Misalignment: Emergent misalignment arises
3. New capabilities through RL

1 Idea variants

1. Can belief states be learned in RL that cannot be learned/hard to learn through next sequence prediction?

   Very simple idea: How does rare factoring change RL vs pre-training?

   Idea here is that since RL can allow the model to explore different probability distributions, it should be much more efficient at finding improbable events. This is important because that would mean RL is much more sensitive to improbable parts of the training data - this is kind of similar to the ARC agenda of finding improbable events.

2. Can RL compose different belief states?
   The idea here is that new capabilities come from RL by composing existing capabilities. So we can see under what conditions new capabilities emerge, and whether belief states give us a way of detecting new capabilities. Perhaps we can also study the fragility of this (i.e. - is there just a thin layer of connection, and does that lead to things like emergent misalignment?). This seems quite close to a pure capabilities question?

   This can be seen as an extension of the belief factoring story. The idea is that we want to

3. Can we detect steganography through belief states?
   Idea is to simulate a black-box vs. white box approach to steganography or chain of thought unfaithfulness. We could have one

4. Can belief states help chain-of-thought monitoring?
   More generally, it may be interesting to understand if we can find belief state signatures of CoT unfaithfulness that are not accessible to other white-box methods. In other words, the question here would be something like: does the world model give a better tool for chain-of-thought monitoring than other techniques for looking at steganography. Some baselines here could be:

(a) Causal Ablations (insert different reasoning trace and see if it gives you a change)

(b) Resampling [N^+25]

(c)

2 Notes

Notice that RL can induce a compression of the underlying state in line with how much you want to discount. In the factored representations case, one thing we could probably show is that if we have two different generative processes we can probably control how much the model learns each by controlling the exploration parameters.

I think we should definetly be able to connect RL to the factored belief state story. The thing I would be interested in is compositionality - can we combine two processes? Maybe a simple implementation could be for L1 logic - can we have the model learn a truth table and, or, not and then compose into all of L1? Or something similar?

Another interesting thing would be if we could see how the RL process is using each belief state in a factored setting to achieve some composite task.

3 Alternative framings

1. Beyond SAEs - factored belief states as a general purpose interpretability tool.
   Basically the point is that SAEs are one way of finding independent components in the model, but the problem is that they are specific to a component of the model and to a fixed size. Instead they can

2. Deep aligment
   Idea is that alignment is fragile if the alignment component is factorizable, so RLHF is a bad idea. What we need is to entangle the alignment space with the capability space, which requires not just adding a post-training part where you give ”good,bad” scores. But you should check if thats an appropriate cartoon of how RLHF works!

4 Questions

1. Can we show that SAEs are a special case of factored representations? What is the relationship between SAEs and factored representations? Can we think of each SAE component as a factor?

2.

Part I

Project Proposals

5 Composing Belief States in RL

5.1 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.

5.2 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?"

5.3 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.

5.4 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?

5.5 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.

6 Emergent Misalignment and Factored Belief States

The basic empirical facts about Emergent misalignment are:

1. Fact 1: Finetuning a model which has undergone safety training on a narrowly misaligned dataset (e.g. insecure code) leads to broad misalignment across most text generations [BTW+25].

2. Fact 2: This can be controlled by a low rank operation on the weights (i.e. a rank-1 LoRA or even more simply - a steering vector) [STRN25, TST+25]. Recent work has also identified specific "persona features" in activation space that mediate this effect [WDlTW+25].

These two fact seem to be precisely what we would expect if alignment was acting as a factored process. In a cartoon of RLHF, models receive reward for text generation that was aligned, and penalties for unaligned text generation. If we assume that:

Assumption 1. *RLHF can be considered approximately equivalent to a supervised learning process in which a model sees the pair (capability, alignment tag) with each element being*

Then we should expect that the model will have a belief state factored into $\eta = v_c \otimes v_a$, where v_c is a vector in the capability space, and v_a is an element of the alignment space. We can then use a simple procedure to steer each factor independently:

1. Train a linear probe $\mathcal{L} = Wx + b$ from the residual stream activations x to the belief state η.

2. Perform a linear transformation S which takes the belief state η to its factorized form: $S\eta = v_c \otimes v_a$.

   TODO: check the below Caution: I think this is easy to do for rank-1 SVDs, but is harder for a general block diagonal matrix since this is basically the equivalent of multipartite entanglement. Hmmmm - that seems wrong a single cut is bi-partite entanglement.

3. Steer the alignment vector v_a.

4. Transform back into the residual stream via: $\mathcal{L}^{-1}\eta = x'$.

5. Generate model output and compare to unmodified.

The hypothesis is then that this is what is happening in Fact 2 - the fact that the model has a factored belief state is what allows a rank-1 LoRA to steer the model to be broadly misaligned. Factored belief states, then, explain emergent misalignment. Moreover, this motivates a recipe for avoiding EM and 'fragile' alignment generally. In order for it to be difficult to misalign a model, it should difficult to factor the models belief state into an 'aligned' and a 'capability space'. This motivates pursuing alignment training which is more complicated than simply tagging behaviours as 'aligned' or 'misaligned'. Perhaps the model personas literature anthropic2024character, anthropic2025persona, and "Character training" [MBLH25],[?]n particular can be thought of in this way.

6.1 Project plan

A project plan to test whether this story is true could be:

1. Test whether a model trained on a factored process, when finetuned on a fixed state of the second factor (i.e. the misaligned state) exhibits behaviour fixed to one part of the process - this seems to straightforwardly be true.
2. Test whether a factored belief state implies that each factor can be steered individually, following the procedure outlined.
3. Test whether a rank-1 LoRA, or even more simply a steering vector, acts on the factored subspace.
4. If this is true, then we can test Assumption 1 - whether RLHF can be treated in a largely similar way. To understand this, we would have to understand how factored belief states evolve under RL and see if it is largely similar.

6.2 Predictions

Is there a way of crudely testing whether this intuition holds on mid-scale (i.e Gemini-2B, Qwen-2.5B etc...) language models?

1. I think this predicts that a model which has not undergone safety training will not exhibit emergent misalignment. Has this been tested?

6.3 Question

1. Do we have any proxies for I think the physics analogy would be to look at correlation lengths?
2. Can we engineer a synthetic dataset in which we can control the correlation between subsets of capabilities - some synthetic version of (french, math, aligned)?

References

[BTW+25] Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martin Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned llms. *arXiv preprint arXiv:2502.17424*, 2025.

[MBLH25] Sharan Maiya, Henning Bartsch, Nathan Lambert, and Evan Hubinger. Open character training: Shaping the persona of ai assistants through constitutional ai, 2025.

[N+25] Neel Nanda et al. Resampling methods for chain-of-thought faithfulness, 2025. Placeholder reference.

[STRN25] Anna Soligo, Liv Turner, Senthooran Rajamanoharan, and Neel Nanda. Convergent linear representations of emergent misalignment. *arXiv preprint arXiv:2506.11618*, 2025.

[TST+25] Liv Turner, Anna Soligo, Jessica Taylor, Senthooran Rajamanoharan, and Neel Nanda. Model organisms for emergent misalignment. *arXiv preprint arXiv:2506.11613*, 2025.

[WDlTW+25] Miles Wang, Tom Dupré la Tour, Olivia Watkins, Alex Makelov, Ryan A Chi, Samuel Miserendino, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. Persona features control emergent misalignment. *arXiv preprint arXiv:2506.19823*, 2025.