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 on a narrowly misaligned dataset (e.g. insecure code) leads to broad misalignment across most text generations [BTW+25]. This works both for model that have undergone safety training, and those that have not ([BTW+25][Sec 4.8])

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 an 'almost' 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: $\mathcal{L}'\eta = x'$, solving the inverse problem of the original map.

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 [**?**, **?**], and "Character training" [MBLH25] in 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. This is explicitly NOT what is found in the original emergent misalignment paper [BTW+25]. I spoke with Daniel Tan, one of the authors, about this and he suspects this is because of training data leakage - i.e. the model knows that this

7 'Almost factored' processes

7.1 Dictionary between Emergent Misalignment and Almost factored belief

7.2 Partition-preserving tensor-product GHMMs (block invariance)

GHMM notation. We follow the GHMM notation of `arXiv:2507.07432`. Let $\mathcal{X}$ be a finite alphabet, let $\big(T(x)\big)_{x\in\mathcal{X}}$ be the transfer matrices of a (finite-state) generalized hidden Markov model (GHMM), and let $\langle\!\langle \eta(\emptyset)|$ be the initial latent row vector. The net transition operator is

$$T \;\coloneqq\; \sum_{x\in\mathcal{X}} T(x), \tag{7.1}$$

which has eigenvalue 1 with associated right eigenvector $|1\rangle\rangle$ satisfying $T|1\rangle\rangle = |1\rangle\rangle$. We normalize $\langle\!\langle \eta(\emptyset)|$ so that $\langle\!\langle \eta(\emptyset)|1\rangle\rangle = 1$.

For a word $w = x_{1:\ell}$, define $T(w) \coloneqq T(x_1)\cdots T(x_\ell)$. Then

$$\Pr(X_{1:\ell} = w) \;=\; \langle\!\langle \eta(\emptyset)|\, T(w)\, |1\rangle\rangle. \tag{7.2}$$

The corresponding (normalized) predictive vector is

$$\langle\!\langle \eta(w)| \;\coloneqq\; \frac{\langle\!\langle \eta(\emptyset)|\, T(w)}{\langle\!\langle \eta(\emptyset)|\, T(w)\, |1\rangle\rangle}. \tag{7.3}$$

In the HMM special case, the $T(x)$ are substochastic matrices with entries $T^{(x)}_{s,s'} = \Pr(s', x \mid s)$, and T is row-stochastic so $|1\rangle\rangle = \mathbf{1}$.

Intuition for 'Almost factored processes' (AFPs) For simplicitly, let us consider the total state space as consisting of the tensor product of two spaces, A and B such that a general state vector lives in the tensor product state:

$$\eta = \eta_A \otimes \eta_B \tag{7.4}$$

As an intuition pump, we imagine that A and B are both "world models" of different capabilities (bad word). Previously we studied 'Purely Factored Processes' (PFPs). Intuitively, these processes represent sequences which contain distinct capabilities - for example, space A may represent the space of mathematical symbols, and space B the set of proper names. The key property of these processes is that knowing any piece of information about A does not tell you any information about B.

For Emergent Misalignment, however, this assumption should not hold. As another intuition pump, we take space A to represent a model persona - for simplicity consisting of an "aligned" space A_1 and a misaligned space A_2 such that $A = A_1 \oplus A_2$. We imagine space B to be the world model associated with a given capability - producing code say. We imagine we can participate that space into a space of "secure" code B_1 and insecure code B_2. In this case we expect the intuition of factored process to break down strongly - knowing that the model is in its 'misaligned personas' strongly predicts that the model will also produce 'insecure' code.

Defining AFPs To model this dependence, we must generalize PFPs to a family of 'Almost Factored Processes' (AFPs). In general, we consider F factor spaces $I^{(f)}$ (e.g. $I^{(1)} = A$, $I^{(2)} = B$, ...), each of which is partitioned into N subspaces $I_i^{(f)}$ so that $I^{(f)} = \bigoplus_{i=1}^{N} I_i^{(f)}$. The general space of belief states V is then given by:

$$V = \bigotimes_{f=1}^{F} \left(\bigoplus_{i=1}^{N} I_i^{(f)} \right) \tag{7.5}$$

which distributes as:

$$V \;\cong \bigoplus_{\mathbf{i} \in \{1,\dots,N\}^F} \; \bigotimes_{f=1}^{F} I_{i_f}^{(f)}, \tag{7.6}$$

where $\mathbf{i} = (i_1, \dots, i_F)$ indexes the choice of one sector in each factor. Equivalently, we can index the sum by functions $\sigma : \{1, \dots, F\} \to \{1, \dots, N\}$:

$$V \;\cong \bigoplus_{\sigma : \{1,\dots,F\} \to \{1,\dots,N\}} \; \bigotimes_{f=1}^{F} I_{\sigma(f)}^{(f)}. \tag{7.7}$$

As is, this is just a relabelling of the underlying spaces. To give the partition operational significance, we impose the additional constraint that the transition operators respect the sector decomposition. Writing

$$V_\sigma \;\coloneqq \bigotimes_{f=1}^{F} I_{\sigma(f)}^{(f)}, \qquad V \;\cong \bigoplus_{\sigma : \{1,\dots,F\} \to \{1,\dots,N\}} V_\sigma, \tag{7.8}$$

we require *partition preservation*:

$$T(x)\big(V_\sigma\big) \subseteq V_\sigma, \qquad \forall\, x \in \mathcal{X},\ \forall\, \sigma : \{1, \dots, F\} \to \{1, \dots, N\}. \tag{7.9}$$

Equivalently, each $T(x)$ is block diagonal with respect to the direct-sum decomposition $V = \bigoplus_\sigma V_\sigma$.

Importantly, this is *not* the same as requiring full factorization: we do *not* require that a belief state in V_σ can be written as a pure tensor $\bigotimes_f v^{(f)}$, nor do we require that the restricted dynamics on a block factorizes as a tensor product of single-factor maps.

Canonical example: Two factored spaces spaces, two subspaces for each Let A and B be the hidden state spaces of two processes. A standard "factored" joint process on $A \otimes B$ has symbol-indexed operators of the form

$$T(x) \;=\; T^A(x_A) \otimes T^B(x_B), \tag{7.10}$$

where the observed symbol x encodes a pair (x_A, x_B), and $T^A(\cdot)$, $T^B(\cdot)$ are the single-process GHMM transfer matrices.

A partition of the state spaces. Assume that each factor admits a direct-sum decomposition

$$A \;=\; A_1 \oplus A_2, \qquad B \;=\; B_1 \oplus B_2. \tag{7.11}$$

By bilinearity of the tensor product, there is a canonical identification

$$A \otimes B \;\cong\; (A_1 \otimes B_1) \;\oplus\; (A_1 \otimes B_2) \;\oplus\; (A_2 \otimes B_1) \;\oplus\; (A_2 \otimes B_2). \tag{7.12}$$

We write $V_{ij} \coloneqq A_i \otimes B_j$ for these four sectors.

Partition-preserving (block-invariant) dynamics. We say that a joint GHMM on $A \otimes B$ *preserves the partition* if, for every symbol $x \in \mathcal{X}$, each sector V_{ij} is invariant under $T(x)$:

$$T(x)\big(V_{ij}\big) \subseteq V_{ij}, \qquad \forall\, x \in \mathcal{X},\ \forall\, i, j \in \{1, 2\}. \tag{7.13}$$

Equivalently, after choosing a basis adapted to the direct sum (**??**), each $T(x)$ is block diagonal with respect to the 4-sector decomposition:

$$T(x) \;=\; \begin{pmatrix} T_{11}(x) & 0 & 0 & 0 \\ 0 & T_{12}(x) & 0 & 0 \\ 0 & 0 & T_{21}(x) & 0 \\ 0 & 0 & 0 & T_{22}(x) \end{pmatrix}, \qquad T_{ij}(x) : V_{ij} \to V_{ij}. \tag{7.14}$$

We emphasize that this only constrains the *direct-sum* structure: within each block V_{ij}, the dynamics $T_{ij}(x)$ can be arbitrary and need not factorize as a tensor product of maps on A_i

and B_j. Likewise, a belief state supported on V_{ij} need not be a pure tensor in $A_i \otimes B_j$. This constraint is stronger than merely preserving a *sum* of sectors (e.g. $V_{11} \oplus V_{22}$); it enforces that the four bilinear components in the expansion

$$(a_1 + a_2) \otimes (b_1 + b_2) = a_1 \otimes b_1 + a_1 \otimes b_2 + a_2 \otimes b_1 + a_2 \otimes b_2 \tag{7.15}$$

do not mix under the dynamics.

Example: a partition-preserving Z1R-like single process. Consider a 3-state process with hidden-state ordering $(S0, S1, SR)$ and a binary alphabet $\{0, 1\}$. We impose the partition

$$\text{span}\{S0\} \;\oplus\; \text{span}\{S1, SR\}. \tag{7.16}$$

A simple Z1R-like choice that preserves this split, while retaining the emission probabilities of the usual Z1R (i.e. $\Pr(0 \mid S0) = 1$, $\Pr(1 \mid S1) = 1$, and $\Pr(0 \mid SR) = \Pr(1 \mid SR) = \frac{1}{2}$), is

$$T'_0 \;=\; \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{pmatrix}, \qquad T'_1 \;=\; \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{pmatrix}. \tag{7.17}$$

Note that $T' \coloneqq T'_0 + T'_1$ is row-stochastic:

$$T' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad T'\mathbf{1} = \mathbf{1}. \tag{7.18}$$

Two-process joint model with a 4-symbol alphabet. Let both factors A and B be copies of the process (**??**). The joint hidden space is $A \otimes B$ with basis ordered lexicographically:

$$(S0, S0), (S0, S1), (S0, SR), (S1, S0), (S1, S1), (S1, SR), (SR, S0), (SR, S1), (SR, SR). \tag{7.19}$$

We encode the pair of emitted bits $(x_A, x_B) \in \{0, 1\}^2$ as a single observed symbol in a 4-letter alphabet

$$\mathtt{A} = (0, 0), \qquad \mathtt{B} = (0, 1), \qquad \mathtt{C} = (1, 0), \qquad \mathtt{D} = (1, 1). \tag{7.20}$$

The joint GHMM transfer matrices are then defined by Kronecker products:

$$T_{\mathtt{A}} \;=\; T'_0 \otimes T'_0, \qquad T_{\mathtt{B}} \;=\; T'_0 \otimes T'_1, \qquad T_{\mathtt{C}} \;=\; T'_1 \otimes T'_0, \qquad T_{\mathtt{D}} \;=\; T'_1 \otimes T'_1. \tag{7.21}$$

By construction, each $T_{\mathtt{A}}, T_{\mathtt{B}}, T_{\mathtt{C}}, T_{\mathtt{D}}$ is entrywise nonnegative and the net transition operator

$$T_{\text{tot}} \;=\; T_{\mathtt{A}} + T_{\mathtt{B}} + T_{\mathtt{C}} + T_{\mathtt{D}} \;=\; (T'_0 + T'_1) \otimes (T'_0 + T'_1) \;=\; T' \otimes T' \tag{7.22}$$

satisfies $T_{\text{tot}}\mathbf{1} = \mathbf{1}$.

Partition preservation in the joint model. Under the joint partition induced by $\text{span}\{S0\} \oplus \text{span}\{S1, SR\}$ in each factor,

$$A \otimes B \;=\; (\text{span}\{S0\} \otimes \text{span}\{S0\}) \oplus (\text{span}\{S0\} \otimes \text{span}\{S1, SR\}) \oplus (\text{span}\{S1, SR\} \otimes \text{span}\{S0\}) \oplus (\text{span}\{S1, S \tag{7.23}$$

each symbol operator in (**??**) leaves all four sectors invariant, hence is block diagonal in an adapted basis as in (**??**). This gives an explicit family of partition-preserving joint GHMMs that reduce to a standard Kronecker-product construction within each block.

7.3 Density matrix formulation

For the two-factor HMM case, it is sometimes helpful to represent a belief state as a (classical) density matrix on $A \otimes B$. Fix the computational basis $\{|s_A\rangle\}_{s_A \in \{S0, S1, SR\}}$ for A and $\{|s_B\rangle\}_{s_B \in \{S0, S1, SR\}}$ for B, and write $|s_A, s_B\rangle \coloneqq |s_A\rangle \otimes |s_B\rangle$.

Classical density matrix and correlations. Given a history w, the predictive state induces a joint distribution $p_w(s_A, s_B) = \Pr(S_A = s_A, S_B = s_B \mid w)$ on hidden states. The corresponding density matrix is diagonal:

$$\rho_{AB}(w) \;\coloneqq\; \sum_{s_A, s_B} p_w(s_A, s_B)\; |s_A, s_B\rangle \langle s_A, s_B| \,. \tag{7.24}$$

In the lexicographic basis used above,

$$\rho_{AB}(w) \;=\; \text{diag}\Big(p_{00}, p_{01}, p_{0R}, p_{10}, p_{11}, p_{1R}, p_{R0}, p_{R1}, p_{RR}\Big), \tag{7.25}$$

where, e.g., $p_{0R} \coloneqq p_w(S0, SR)$ and $\sum p_{ab} = 1$. The marginals are $\rho_A = \text{Tr}_B\, \rho_{AB}$ and $\rho_B = \text{Tr}_A\, \rho_{AB}$. The factors are *uncorrelated* iff $\rho_{AB} = \rho_A \otimes \rho_B$ (equivalently $p_w(s_A, s_B) = p_w(s_A)\, p_w(s_B)$). Otherwise, the correlation is purely classical: it appears in the diagonal entries (joint probabilities), not in off-diagonal coherences.

Block structure from the partition. With the two-subspace partition $\text{span}\{S0\} \oplus \text{span}\{S1, SR\}$ in each factor, $\rho_{AB}(w)$ is block diagonal with block sizes $1, 2, 2, 4$ corresponding to the sectors

in (**??**). The total weight in each block is the coarse-grained joint distribution over partition labels.

Almost-factored Z1R×Z1R belief state. An "almost factored" (but still classical/diagonal) belief state can have strong correlations between the partitions. For example, putting all mass on the "matched" sectors V_{11} and V_{22} gives

$$\rho_{AB}^{\mathrm{AFP}} \;=\; p\,|S0,S0\rangle\langle S0,S0| \;+\; \frac{1-p}{4}\sum_{s\{S1,SR\}}\sum_{t\in\{S1,SR\}}|s,t\rangle\langle s,t|\,, \tag{7.26}$$

which is block diagonal (so it respects the partition) but generally satisfies $\rho_{AB}^{\mathrm{AFP}} \neq \rho_A \otimes \rho_B$, hence encodes classical correlation between the factors.

8 Reproducing emergenet misalignment fine-tuning

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.