Partial Bayesian Models for Guaranteed Safe AI

Tyra Burgess, Michael Cohen

Composable World Models

Guaranteed Safe AI [1] needs a way to combine individual theories of how different parts of the world work into a combined statistical world model.

This isn't possible with standard Bayesian induction over the theories, since every theory needs to make predictions about every part of the world. (i.e. regardless of what you're conditioning on).

Partial Bayesian Models

Using partial Bayesian models means models can choose whether to make a prediction or not, rather than making one every timestep.

This lets you build weather models that are only 'active' if the prior observations mean the current prediction is about the weather, or politics models that are only active if the current prediction is about politics, etc.

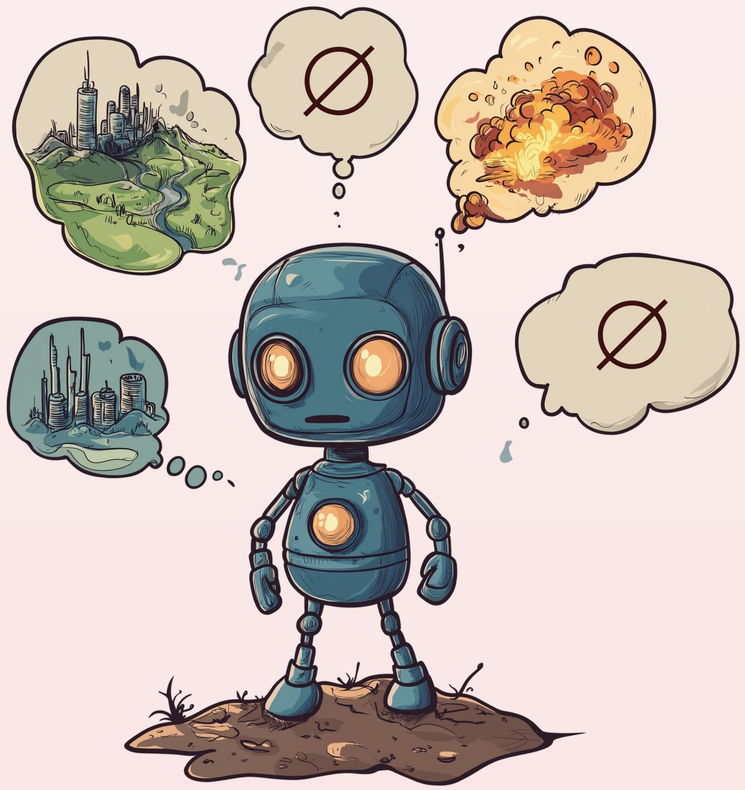
Total vs Partial

$$\tau : \text{List}\, \mathcal{Z} \to \text{Distribution over}\, \mathcal{Z}$$
$$\pi : \text{List}\, \mathcal{Z} \to \text{Maybe}\,(\text{Distribution over}\, \mathcal{Z})$$

What This Gets Us

Having a system for composable world modelling means:

- *Modular Safety*: Can break complex models into simpler, individually verifiable parts
- *Transparency*: Easier to audit, validate, and trust each component separately
- *Reusability*: Verified models can reused across different AI systems



Convergence to Accuracy

I've proven that doing induction over a set which contains an accurate partial model π* means your mixture predictions will converge to accuracy when π* is active.

Theorem 1

$$\mathbb{E}_\mu \left[\sum_{t=1}^{\infty} \mathbf{1}^{\pi^*}_{<t} \left(\mu_t(z_t) - \xi_t(z_t) \right)^2 \right] \leq \log \left(\frac{1}{w(\pi^*)} \right)$$

Bounding Harm

Recent results show that if you have a Bayesian oracle, you can upper bound the probability of an agent's action resulting in harmful outcomes by considering only a finite number of high-posterior models. [2]

However, this assumes that the true world model is in your hypothesis class. I am significantly relaxing that assumption using partial models and confirming similar results hold.

Theorem 2

The following holds with probability 1 - δ on the timesteps π* is active.

$$\mu_t(Y_t = 1) \leq \max_{\tau \in \mathcal{M}^{\alpha}_{\leq t}} \tau_t(Y_t = 1)$$

Next Steps

Now, I'm working to prove that the inequality converges towards equality, i.e. that the following holds for some B.

Theorem 3?

$$\mathbb{E}_\mu \left[\sum_{t=1}^{\infty} \mathbf{1}^{\pi^*}_{<t} \left(\mu_t(Y_t = 1) - \max_{\tau \in \mathcal{M}^{\alpha}_{\leq t}} \tau_t(Y_t = 1) \right)^2 \right] \leq B$$

References

[1] Dalrymple, D. et al. (2024, May 10). Towards Guaranteed Safe AI: a framework for ensuring robust and reliable AI systems. arXiv.org. https://arxiv.org/abs/2405.06624

[2] Bengio, Y., Cohen, M. K., Malkin, N., MacDermott, M., Fornasiere, D., Greiner, P., & Kaddar, Y. (2024, August 9). Can a Bayesian Oracle Prevent Harm from an Agent? arXiv.org. https://arxiv.org/abs/2408.05284