Inner Alignment through the lens of Causality: Initial Takeaways

David Reber
Columbia University
david.reber@columbia.edu

Abstract: abstract

0.1. Intended Audience

This post assumes prior familiarity with inner alignment; in particular, you've read at least two of post A, post B, or post C. Furthermore, basic familiarity with causal graphs; you've read The Book of Why, these posts, or better yet, [1].

1. Introduction

Every researcher knows that the language you use is itself a modeling choice. To paraphrase the oft-cited quote: 'Every language has blindspots, but some are useful.' Here, I present one possible mathematical formalism of inner alignment, some important implications and open questions that immediately follow, while tallying up the total cost of assumptions (blindspots) that are incurred along the way.

2. What is inner alignment?

Thus far, the extent to which inner alignment has been mathematicized appears to be in the following two lines:

$$\theta^* = \arg\max_{\theta} \mathbb{E}(O_b(\pi_\theta)) \tag{1}$$

$$\pi_\theta = \arg\max_{\pi} \mathbb{E}(O_m(\pi|\theta)) \tag{2}$$

where O_b and O_m are the base- and mesa-objective functions, inputting a policy π_θ and π (after θ has been fixed), respectively (cite). This represents the dual optimization process of the base optimizer trying to find the best parameters for a mesa optimizer, to maximize the base objective (while the mesa optimizer is itself optimizing the mesa objective). Alternatively, you can think of expressing these as a system of equations expressing that mesa-optimization has arisen, analogous to calculus' necessary first-order conditions satisfied at every critial point. (extra:A).

3. Current language of choice: Causality

What are the causal concepts we'll use to express (and then expore) inner alignment?

- causal graph
  - exogeneity
  - causal dependence
- soft interventions
  - σ-calculus
- transportability
  - selection diagram
- (structural learning)
  - equivalence classes of causal diagrams
- (sensitivity analysis)

Assumption: We're going to assume acyclicity for now. (I have inside-view reasons to believe the analysis will extend naturally to cyclic systems as well).

(a) minimal aligned

(b) selection

Figure 1: caption

3.1. *Mesa Optimization: Soft Interventions*

3.2. *Inner alignment as a question of transportability*

A simple regime indicator s is simply: "Are we in training right now?".
(Extra:B, Extra:C, Main:1, Main:2, Main:3, Main:4)
We say that the mesa-optimizer D_m is *inner aligned* if

$$\mathbb{E}(Y_b|Y_m, s, \sigma_{D_m}) = \mathbb{E}(Y_b|Y_m, \sigma_{D_m})$$

This seems like the right definition because

1. Even though we can't measure Y_m directly, we know it will always be maximally high
2. The right-hand side is precisely the optimal expectation achieved during training

A sufficient (and necessary?) condition for this to hold is if Rule 1 of the σ-calculus holds:

$$(Y_b \perp\!\!\!\perp s|Y_m)_{G_{\sigma_{D_m}}}$$

That is, if Y_m d-separates Y_b from the regime change, we get robustness; we can reliably trust the mesa optimizer to help us achieve high Y_b (at least, as well as it's capabilities allow...aka. intent alignment?). Extra:D
I believe a necessary-and-sufficient condition for $(Y_b \perp\!\!\!\perp s|Y_m)_{G_{\sigma_{D_m}}}$ is (excluding trivial cases like no causal connection whatsoever):

1. Y_b must be causally dependent on Y_m
2. All causal paths from s to Y_b must pas through Y_m

Extra:E, Extra:F.
Which situations pass this criteria, which don't? (Extras: H - L)

3.3. *What does $(Y_b \perp\!\!\!\perp s|Y_m)_{G_{\sigma_{D_m}}}$ mean practically?*

As in, how could we empirically test it?

- if we know Y_m, then knowing s doesn't give us any more info about Y_b.
- if we know Y_m is maximized then Y_b is also maximized (thanks to the mesa-equations). or is it just that the performance won't be worse than in training? Not necessarily better?

3.4. *Speculation towards a plausible experiment*

If

1. only the data is subject to regime shift,
2. ....

Then $(Y_b \perp\!\!\!\perp s|Y_m)_{G_{\sigma_{D_m}}} \iff (Y_b \perp\!\!\!\perp data|D_b)_{P(V)}$? (if so, then perhaps we can evaulate this during/after training?).
Furthermore, if some subset of the data satisfies the required independence relation, then we are robust w.r.t distributional shift in those dimensions.
Finite data considerations? how to practically test the independence?

4. Sensitivity analysis

So far, this has all been about enforcing a very strict equality. If however, we allow a mere bound (dependent on Y_m) like Y_b is always 90 percent of optimal" then we might be able to do partial-identifiability; be able to run simulations just using the data/reward correlations (without needing the optimal policy), etc.

If we're willing to provide a parametrization of the world, then we can do sensitivity analysis! (Possible parametrizations: linear regression. Graph-constrained NN).

This would look like "Assuming this is the mesa objective function, how sensitive is Y_b to a small regime change in the data?" (And then perhaps we can try to encourage 'nicer' mesa-objectives? soft inner alignment)

5. musings

I wonder if it's possible to numerically simulate the (likelihood? feasibility? magnitude?) of inner alignment as follows: (data $\rightarrow D$, D is parent of both Y_b and Y_m) For a given mesa objective, does the correlation (data,high-Y_m, high-Y_b) exist? Is it a common occurance? If so, how does the feasibility/frequency change as we gradually introduce a regime change to the data?

6. Would you rather? The Wireheading / Inner Misalignment tradeoff

We could perhaps force $D \rightarrow Y_m \rightarrow Y_b$by explicitly making reward depend on neuron values, etc. But won't this introduce a control incentive??

More generally, I wonder if $(Y_b \perp\!\!\!\perp s|Y_m)_{G_{\sigma_{D_m}}}$ is really getting at **a fundamental tradeoff between wireheading and inner misalignment**, if both boil down to whether or not a path exists? (hopefully there's some cases missing - this is just intution speaking, haven't actually written it all out yet).

Even if this is the case, perhaps the effect of each could be attuned, so that we find a sweet-spot in the middle which kind of minimizes both.

References

1. Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.