Corrigibility

Paul Christiano Follow
Jun 10, 2017 · 8 min read

(*Warning: rambling.*)

I would like to build AI systems which help me:

- Figure out whether I built the right AI and correct any mistakes I made
- Remain informed about the AI's behavior and avoid unpleasant surprises
- Make better decisions and clarify my preferences
- Acquire resources and remain in effective control of them
- Ensure that my AI systems continue to do all of these nice things
- ...and so on

We say an agent is *corrigible* (article on Arbital) if it has these properties. I believe this concept was introduced in the context of AI by Eliezer and named by Robert Miles; it has often been discussed in the context of narrow behaviors like respecting an off-switch, but here I am using it in the broadest possible sense.

In this post I claim:

1. A benign act-based agent will be robustly corrigible if we want it to be.
2. A sufficiently corrigible agent will tend to become more corrigible and benign over time. Corrigibility marks out a broad basin of attraction towards acceptable outcomes.

As a consequence, we shouldn't think about alignment as a narrow target which we need to implement exactly and preserve precisely. We're aiming for a broad basin, and trying to avoid problems that could kick out of that basin.

This view is an important part of my overall optimism about alignment, and an important background assumption in some of my writing.

1. Benign act-based agents can be corrigible

A benign agent optimizes in accordance with our preferences. An act-based agent considers our short-term preferences, including (amongst others) our preference for the agent to be corrigible.

If *on average* we are unhappy with the level of corrigibility of a benign act-based agent, then by construction it is mistaken about our short-term preferences.

This kind of corrigibility doesn't require any special machinery. An act-based agent turns off when the overseer presses the "off" button not because it has received new evidence, or because of delicately balanced incentives. It turns off because that's what the overseer prefers.

Contrast with the usual futurist perspective

Omohundro's The Basic AI Drives argues that "almost all systems [will] protect their utility functions from modification," and Soares, Fallenstein, Yudkowsky, and Armstrong cite as: "almost all [rational] agents are instrumentally motivated to preserve their preferences." This motivates them to consider modifications to an agent to remove this default incentive.

Act-based agents are generally an exception to these arguments, since the overseer has preferences about whether the agent protects its utility function from modification. Omohundro presents preferences-about-your-utility function case as a somewhat pathological exception, but I suspect that it will be the typical state of affairs for powerful AI (as for humans) and it does not appear to be unstable. It's also very easy to implement in 2017.

Is act-based corrigibility robust?

How is corrigibility affected if an agent is ignorant or mistaken about the overseer's preferences?

I think you don't need particularly accurate models of a human's preferences before you can predict that they want their robot to turn off when they press the off button or that they don't want to be lied to.

In the concrete case of an approval-directed agent, "human preferences" are represented by human responses to questions of the form "how happy would you be if I did a?" If the agent is considering the action a precisely because it is manipulative or would thwart the user's attempts to correct the system, then it doesn't seem hard to predict that the overseer will object to a.

Eliezer has suggested that this is a very anthropocentric judgment of "easiness." I don't think that's true—I think that given a description of a proposed course of action, the judgment "is agent X being misled?" is objectively a relatively easy prediction problem (compared to the complexity of generating a strategically deceptive course of action).

Fortunately this is the kind of thing that we will get a great deal of evidence about long in advance. Failing to predict the overseer becomes *less* likely as your agent becomes smarter, not more likely. So if in the near future we build systems that make good enough predictions to be corrigible, then we can expect their superintelligent successors to have the same ability.

(This discussion mostly applies on the training distribution and sets aside issues of robustness/reliability of the predictor itself, for which I think adversarial training is the most plausible solution. This issue will apply to any approach to corrigibility which involves machine learning, which I think includes any realistic approach.)

Is instrumental corrigibility robust?

If an agent shares the overseer's long-term values and is corrigible instrumentally, a slight divergence in values would turn the agent and the overseer into adversaries and totally break corrigibility. This can also happen with a framework like CIRL—if the way the agent infers the overseer's values is slightly different from what the overseer would conclude upon reflection (which seems quite likely when the agent's model is misspecified, as it inevitably will be!) then we have a similar adversarial relationship.

2. Corrigible agents become more corrigible/aligned

In general, an agent will prefer to build other agents that share its preferences. So if an agent inherits a distorted version of the overseer's preferences, we might expect that distortion to persist (or to drift further if subsequent agents also fail to pass on their values correctly).

But a corrigible agent prefers to build other agents that share *the overseer's* preferences—even if the agent doesn't yet share the overseer's preferences perfectly. After all, even if you only approximately know the overseer's preferences, you know that the overseer would prefer the approximation get better rather than worse.

Thus an entire neighborhood of possible preferences lead the agent towards the same basin of attraction. We just have to get "close enough" that we are corrigible, we don't need to build an agent which exactly shares humanity's values, philosophical views, or so on.

In addition to making the initial target bigger, this gives us some reason to be optimistic about the dynamics of AI systems iteratively designing new AI systems. Corrigible systems want to design more corrigible and more capable successors. Rather than our systems traversing a balance beam off of which they could fall at any moment, we can view them as walking along the bottom of a ravine. As long as they don't jump to a completely different part of the landscape, they will continue traversing the correct path.

This is all a bit of a simplification (though I think it gives the right idea). In reality the space of possible errors and perturbations carves out a low degree manifold in the space of all possible minds. Undoubtedly there are "small" perturbations in the space of possible minds which would lead to the agent falling off the balance beam. The task is to parametrize our agents such that the manifold of likely-successors is restricted to the part of the space that looks more like a ravine. In the last section I argued that act-based agents accomplish this, and I'm sure there are alternative approaches.

Amplification

Corrigibility also protects us from gradual value drift during capability amplification. As we build more powerful compound agents, their values may effectively drift. But unless the drift is large enough to disrupt corrigibility, the compound agent will continue to attempt to correct and manage that drift.

This is an important part of my optimism about amplification. It's what makes it coherent to talk about preserving benignity as an inductive invariant, even when "benign" appears to be such a slippery concept. It's why it makes sense to talk about reliability and security as if being "benign" was a boolean property.

In all these cases I think that I should actually have been arguing for corrigibility rather than benignity. The robustness of corrigibility means that we can potentially get by with a *good enough* formalization, rather than needing to get it exactly right. The fact that corrigibility is a basin of attraction allows us to consider failures as discrete events rather than worrying about slight perturbations. And the fact that corrigibility eventually leads to aligned behavior means that *if* we could inductively establish corrigibility, then we'd be happy.

This is still not quite right and not at all formal, but hopefully it's getting closer to my real reasons for optimism.

Conclusion

I think that many futurists are way too pessimistic about alignment. Part of that pessimism seems to stem from a view like "any false move leads to disaster." While there are some kinds of mistakes that clearly do lead to disaster, I also think it is possible to build the kind of AI where *probable* perturbations or errors will be gracefully corrected. In this post I tried to informally flesh out my view. I don't expect this to be completely convincing, but I hope that it can help my more pessimistic readers understand where I am coming from.

Postscript: the hard problem of corrigibility and the diff of my and Eliezer's views

I share many of Eliezer's intuitions regarding the "hard problem of corrigibility" (I assume that Eliezer wrote this article). Eliezer's intuition that there is a "simple core" to corrigibility corresponds to my intuition that corrigible behavior is *easy to learn* in some non-anthropomorphic sense.

I *don't* expect that we will be able to specify corrigibility in a simple but algorithmically useful way, nor that we need to do so. Instead, I am optimistic that we can build agents which learn to reason by human supervision over reasoning steps, which pick up corrigibility along with the other useful characteristics of reasoning.

Eliezer argues that we shouldn't rely on a solution to corrigibility unless it is simple enough that we can formalize and sanity-check it ourselves, even if it appears that it can be learned from a small number of training examples, because an "AI that seemed corrigible in its infrahuman phase [might] suddenly [develop] extreme or unforeseen behaviors when the same allegedly simple central principle was reconsidered at a higher level of intelligence."

I don't buy this argument because I disagree with implicit assumptions about how such principles will be embedded in the reasoning of our agent. For example, I don't think that this principle would affect the agent's reasoning by being explicitly considered. Instead it would influence the way that the reasoning itself worked. It's possible that after translating between our differing assumptions, my enthusiasm about embedding corrigibility deeply in reasoning corresponds to Eliezer's enthusiasm about "lots of particular corrigibility principles."

I feel that my current approach is a reasonable angle of attack on the hard problem of corrigibility, and that we can currently write code which is reasonably likely to solve the problem (though not knowably). I do not feel like we yet have credible alternatives.

I *do* grant that if we need to learn corrigible reasoning, then it is vulnerable to failures of robustness/reliability, and so learned corrigibility is not itself an adequate protection against failures of robustness/reliability. I could imagine other forms of corrigibility that do offer such protection, but it does not seem like the most promising approach to robustness/reliability.

I *do* think that it's reasonably likely (maybe 50–50) that there is some clean concept of "corrigibility" which (a) we can articulate in advance, and (b) plays an important role in our *analysis* of AI systems, if not in their construction.