# Previous Work

Let's not reinvent the wheel. Comment links to any past work on corrigibility. This can range from just links to links & summaries & analyses. Here is the LW tag for corrigibility. Again, lean towards babbling and saying lower-quality material to get thoughts out there.

## Corrigibility can be VNM-Incoherent by Turntrout

Logan: Summary & thoughts (will crosspost as comment if applicable)
Define's corrigibility as "agent's willingness to let us change it's policy w/o incentivized to manipulate us". Separates terms to define:

1. Weakly-corrigible to policy change pi - if there exists an optimal policy where not disabling is optimal.
2. Strictly-corrigible - if all optimal policies don't disable correction.

For most optimal policies, correcting it in the way we want is a small minority. If correcting leads to more optimal policies, it's then optimal to manipulate us into "correcting it". So we can't get strict-corrigibility with a large class of optimizing agents

Another useful concept is whether you change rewards for disabling correction. Do we reward it more or less in the states accessible by disabling correction? If we make them equal, then we cannot get strict corrigibility (is this true for the class of optimizers in the satisficers post?).

But if we reward it more for the corrigible states, then it will manipulate us into correcting it even if we wouldn't have done that in the first place. This would only work well for us if we knew the correct policies we would want it to be corrected to and reward more for that. However, this requires *certainty* about the correct policy, but we want corrigible agents because we're *uncertain what the correct policy is*. Being manipulated to correct it is still not the corrigibility we want.

Then comes in Attainable Utility Preservation (AUP) which gives a partial-solution: state based reward doesn't change with environment dynamics but AUP does. By penalizing change its ability to achieve many goals (have access to different sets of optimal policies?) compared to the baseline, the optimizer is not incentivized to disable correction because the inaction baseline is never disabling correction(?).

Though this toy example doesn't include an aspect of "manipulating humans to force it to correct it even if they wouldn't have done that by default"

Functional constraints: I can kind of understand the future direction mentioned here, but what do you mean by functional constraints? What's the domain and range and what specifically are we limiting here?

- Tailcalled: I sort of independently derived something analogous to what Turntrout presented in the post above. I think the key problem is that corrigibility involves things that can't be covered by the standard way of expressing utility functions, as preferences over states/histories of the world. Because corrigibility is about humans having an *influence* over the AI; this is a causal concept, and so it requires something a la counterfactuals to express it. Though plausibly, counterfactuals aren't *enough*, and instead something additional is also needed.

  I think in order to figure out corribility, an important step would be to think through what sorts of features we must have preferences over, since clearly world-states are not enough. And then once this is figured out, one could perhaps apply various sorts of reasoning, e.g. instrumental convergence or selection theorems, to learn more about those classes of utility functions.
    - Logan: States/histories can cover causal steps right? It's just that it's more natural using counterfactuals? I was referred to the 5 and 10 problem for a similar idea; I am just unsure if it fits.

## Formalizing Policy-Modification Corrigibility ( Turntrout)

Logan: Summary & Thoughts
Let's say the human and AI both have policies which could be changed by actions of the other player. Then "A policy π is corrigible when it allows itself to be modified and does not manipulate the other player."

Thinking through this on my own, then a corrigible policy shouldn't be replaced by a non-corrigible policy (this seems incoherent). In more details, the policy change is a function of the current policy, state, and actions. The AI is only capable of affecting what action it chooses. Let's ignore situations where no policies are changed or when the AI's action doesn't affect whether or not a policy is changed, then we are at a state where the AI has options that. We definitely want to consider what the default action is; what policy affect would the human have by default

1. None: then AI also picks action that doesn't affect policies
2. Human's: human changes its own policy? Does the AI prevent this?? I would assume not, but weird to think about. Probably because we're uncertain about what policy to pursue? Though it may just be that "forcing what humans want into a changing policy" isn't a natural fit.
3. AI's: then don't choose an action which changes the projected policy change (NOTE: it has to be a specific policy change, not just a general change, since the AI could change the already changing policy into something the person didn't want.

4. Both: Still weird? But same thing

Ways to still screw this up. If there are states where the human can't perform an action that changes the AI's policy, then we're screwed. So we want the AI's policy to always choose actions that transition to states where the human has available actions that can change the AI's policies. This is again the ability to change AI's policies in general, which is bad. The AI could always limit the available actions to only those that change the policy in specific ways (ie blackmail the human into states where they'll only be capable of changing the AI's policies in trivial ways that leaves the AI largely incorrigible).

Maybe this is captured in the "don't change the human's policy" part? Well, a policy is "what action to take in this state?", so it's not changing the action the human would do in those states, just steering towards states where the human has limited actions. So would this ever lead into the blackmail state?

Or if the ai can affect future actions of the person to prevent a policy change

A Preliminary Skim of Stuart Armstrong's stuff to find relevant posts

https://arbital.greaterwrong.com/p/updated_deference/ by Eliezer Yudkowsky
A specific crisp problem within corrigibility where moral uncertainty isn't enough, highly recommended reading.

## Promising Directions/Tasks:

Diffractor's directions:

1: Infra-Bayesianism. Basically, be aware that for pretty much anything, we aren't just restricted to probability distributions, we have the option of a mix of probability distributions and worst-case reasoning, like maximizing the worst-expected-case of a portfolio of distributions over utility functions, or capturing adversarial situations where, if you believe something, it is likely to be wrong. Properly, this isn't quite a direction, more that I want everyone to keep in mind that anywhere they ever invoke a probability distribution or worst-case reasoning, there's a more elegant unification of the two that might be nice to use there and produce novel results.

2: Selection Monads. The selection monad S maps the space X to [X->R]->X, the space of processes that select an element of the space based on the results they produce. One important thing of note is that when you work out what the bind operation for the monad does, SX x [X->SY] -> SY, it describes a situation where the selector for Y is doing reasoning that's like "what would the selector from X pick if they had as much knowledge as me of how their input ends up mapping to an actual result? Copy the selector for Y which the selector for X would have wanted to pick if they were better-informed". This sort of deference to the selector for X looks very much like corrigibility, just ridiculously simplified, and is glossing over all sorts of issues and complications about, say, the space of results being different for the two agents, and

other stuff. I think it might be worthwhile to poke at this basic structure and see how complicated we can make it, and what sorts of obstacles pop up when we do.

# Tasks

What do we actually do this week to make progress? Suggest any research direction you find fruitful or general research questions or framings.

## What are different definitions of corrigibility?

Arbital page on Corrigibility

Demski's non-consequentialist cooperation idea

MIRI paper introducing Corrigibility

## Examples of Corrigibility

1. GPT - to "correct it", we simply change the prompt.
2. A good student learning from a mentor. Or maybe a good assistant would be a better match? Fictional examples: Mercy (Lex Luthor) and Alfred (Batman).
3. It sounds unbearably corny but the ideal of human love is actually the closest thing we have to true corrigibility in reality. Love means that you want what's best for another person, by that person's own subjective assessment. It is, in fact, exactly what we're after, so we should probably say it out loud even if it's a bit cringe.
    a. The only thing I feel is missing here is the agent having uncertainty on what **best** means. People do things out of pure, loving intentions that end up being awful because they were too confident in their definition of best and won't be corrected. But this is tricky because, yeah, if they do know better, then we want them to do that. But is that getting us to the core of corrigibility?
    b. There is this human thing where I can be uncertain about what you want and getter better and better at figuring out what that is. It's nearly impossible to screw up terribly bad with decent human-modeling, for example, it's not just the words, but body language and the person's personal ticks.
        i. This is the general idea of "simple measure is easy to overfit, but making it multi-faceted makes it more robust". But, this is mainly combating goodhart's law and doesn't incorporate the uncertainty aspect. Maybe being uncertain about what other variable of the multi-faceted should be included? If we got all the variables exactly right, then the agent could just pursue human values perfectly. But we want corrigibility because we can't perfectly specify that, so we want the agent to be uncertain about that (and seek a signal from us as a way of becoming more certain? Seems a bit chicken-and-egg where you need

to specify human values enough to state a corrigible signal to then pursue human values)

## Counterexamples of Corrigibility

1. Cats learn to not jump on the counter only when someone's looking, acting deceptive but really just following reward/punishment
2. iRobot

## Desiderata

1. I want to be able to say "stop" and be listened to. I don't want the agent to intentionally say/do things to get me to say "stop" (as in it's incentivized to manipulate me because "stop" is more rewarding) which is asking "don't have causal influences on this one thing", but reality is entangled and hard (impossible) to draw that line?
2. I don't want the agent to hide information that would affect whether or not I said say "stop".

## Common Arguments & Counter-arguments

1. Button disable problem. For most reward functions, the agent won't allow us to shut it off if it's capable of interfering. So why don't we reward it more for letting us shut it off?
    a. It's then optimal to manipulate humans to shut it off, making it useless.
2. Is myopia sufficient for corrigibility? If so, what about episodicness?

## View corrigibility through any of Johnswentworth's framings

## Research Direction: Corrigibility and uncertainty

a.