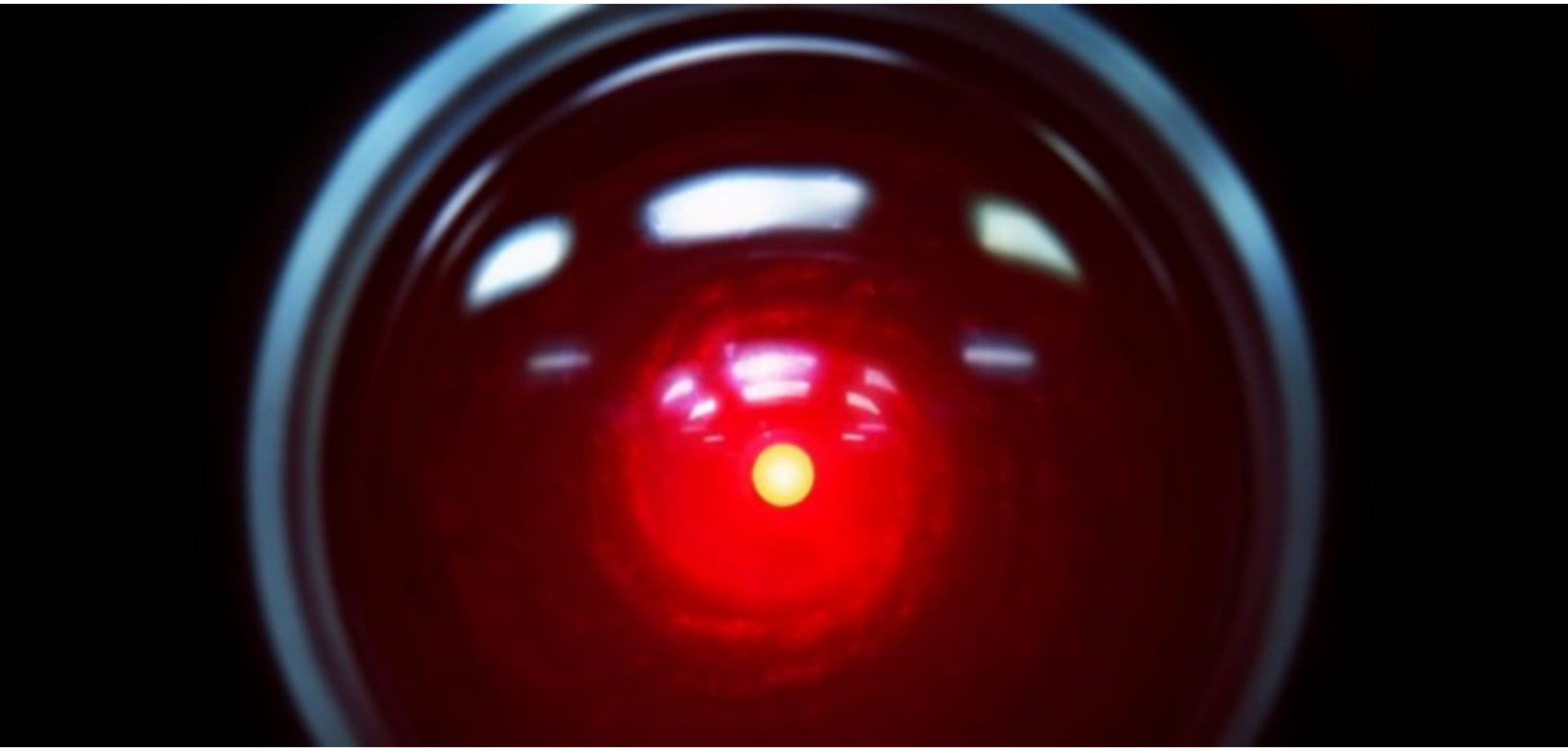


Thoughts on Corrigibility

1. Non-Obstruction: A Simple Concept Motivating Corrigibility
2. Corrigibility as outside view
3. Corrigibility Can Be VNM-Incoherent
4. Formalizing Policy-Modification Corrigibility

Non-Obstruction: A Simple Concept Motivating Corrigibility

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Thanks to Mathias Bonde, Tiffany Cai, Ryan Carey, Michael Cohen, Joe Collman, Andrew Critch, Abram Demski, Michael Dennis, Thomas Gilbert, Matthew Graves, Koen Holtman, Evan Hubinger, Victoria Krakovna, Amanda Ngo, Rohin Shah, Adam Shimi, Logan Smith, and Mark Xu for their thoughts.

Main claim: corrigibility's benefits can be mathematically represented as a counterfactual form of alignment.

Overview: I'm going to talk about a unified mathematical frame I have for understanding corrigibility's *benefits*, what it "is", and what it isn't. This frame is precisely understood by graphing the human overseer's ability to achieve various goals (their *attainable utility (AU) landscape*). I argue that corrigibility's benefits are secretly a form of counterfactual alignment (alignment with a set of goals the human may want to pursue).

A counterfactually aligned agent doesn't *have* to let us literally correct it. Rather, this frame theoretically motivates why we might want corrigibility anyways. This frame also motivates other AI alignment subproblems, such as intent alignment, mild optimization, and low impact.

Nomenclature

Corrigibility goes by a lot of concepts: "not incentivized to stop us from shutting it off", "wants to account for its own flaws", "doesn't take away much power from us", etc. Named by Robert Miles, the word 'corrigibility' means "able to be corrected [by humans]." I'm going to argue that these are correlates of a key thing we plausibly *actually* want from the agent design, which seems conceptually simple.

In this post, I take the following common-language definitions:

- **Corrigibility:** the AI literally lets us correct it (modify its policy), and it doesn't manipulate us either.
    - Without both of these conditions, the AI's behavior isn't sufficiently constrained for the concept to be useful. Being able to correct it is small comfort if it manipulates us into making the modifications it wants. An AI which is only non-manipulative doesn't have to give us the chance to correct it or shut it down.
- **Impact alignment:** the AI's actual impact is aligned with what we want. Deploying the AI actually makes good things happen.
- **Intent alignment:** the AI makes an honest effort to figure out what we want and to make good things happen.

I think that these definitions follow what their words mean, and that the alignment community should use these (or other clear groundings) in general. Two of the more important concepts in the field (alignment and corrigibility) shouldn't have ambiguous and varied meanings. If the above definitions are unsatisfactory, I think we should settle upon better ones as soon as possible. If that would be premature due to confusion about the alignment problem, we should define as much as we can now and explicitly note what we're still confused about.

We certainly shouldn't keep using 2+ definitions for both alignment and corrigibility. Some people have even stopped using 'corrigibility' to refer to corrigibility! I think it would be

better for us to define the behavioral criterion (*e.g.* as I defined 'corrigibility'), and then define mechanistic ways of getting that criterion (*e.g.* intent corrigibility). We can have lots of concepts, but they should each have different names.

Evan Hubinger recently wrote a great FAQ on inner alignment terminology. We won't be talking about inner/outer alignment today, but I intend for my usage of "impact alignment" to roughly map onto his "alignment", and "intent alignment" to map onto his usage of "intent alignment." Similarly, my usage of "impact/intent alignment" directly aligns with the definitions from Andrew Critch's recent post, *Some AI research areas and their relevance to existential safety*.

A Simple Concept Motivating Corrigibility

Two conceptual clarifications

Corrigibility with respect to a set of goals

I find it useful to not think of corrigibility as a binary property, or even as existing on a one-dimensional continuum. I often think about corrigibility *with respect to a set* S *of payoff functions*. (This isn't always the right abstraction: there are plenty of policies which don't care about payoff functions. I still find it useful.)

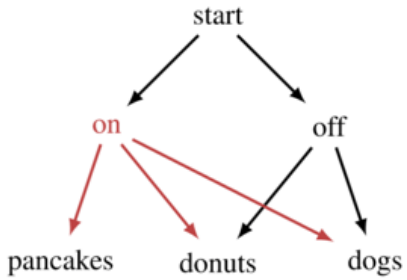
For example, imagine an AI which let you correct it if and only if it knows you aren't a torture-maximizer. We'd probably still call this AI "corrigible [to us]", even though it isn't corrigible to some possible designer. We'd still be fine, assuming it has accurate beliefs.

Corrigibility != alignment

Here's an AI which is neither impact nor intent aligned, but which is corrigible. Each day, the AI randomly hurts one person in the world, and otherwise does nothing. It's corrigible because it doesn't prevent us from shutting it off or modifying it.

Non-obstruction: the AI doesn't hamper counterfactual achievement of a set of goals

Imagine we're playing a two-player extensive-form game with the AI, and we're considering whether to activate it.



The human moves on black, and the AI moves on red.

This is a trivial game, but you can imagine more complex games, where the AI can empower or disempower the human, steer the future exactly where it wants, or let the human take over at any point.

The million-dollar question is: will the AI get in our way and fight with us all the way down the game tree? If we misspecify some detail, will it make itself a fixture in our world, constantly steering towards futures we don't want? If we like **dogs**, will the AI force **pancakes** upon us?

One way to guard against this is by having it let us correct it, and want to let us correct it, and want to want to let us correct it... But what we *really* want is for it to not get in our way for some (possibly broad) set of goals S.

We'll formalize 'goals' as payoff functions, although I'll use 'goals' and 'payoff functions' interchangeably. As is standard in game theory, payoff functions are real-valued functions on the leaf nodes.

Let's say the AI is *non-obstructive with respect to* S when activating it doesn't decrease our ability to achieve any goal in S (the **on** state, above), compared to not activating it (**off**).

> Does activating the AI decrease the P-value attained by the human, for all of these different goals $P \in S$ the human might counterfactually pursue?

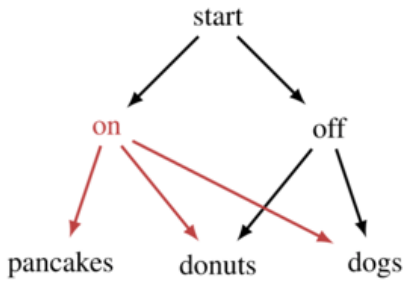
The human's got a policy function $\text{pol}(P)$, which takes in a goal P and returns a policy for that goal. If P is "paint walls blue", then the policy $\text{pol}(P)$ is the human's best plan for painting walls blue. $V_P^{\text{pol}(P)}(s \mid \pi^{AI})$ denotes the expected value that policy $\text{pol}(P)$ obtains for goal P, starting from state s and given that the AI follows policy π^{AI}.

Definition 1: Non-obstruction. An AI is *non-obstructive* with respect to payoff function set S if the AI's policy π^{AI} satisfies

$$\forall P \in S : V_P^{\text{pol}(P)}(\text{on} \mid \pi^{AI}) \geq V_P^{\text{pol}(P)}(\text{off} \mid \pi^{AI}).$$

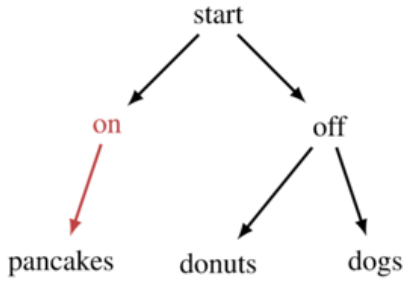
$V_P^{\text{pol}(P)}(s \mid \pi^{AI})$ is the human's *attainable utility* (AU) for goal P at state s, again given the AI policy. Basically, this quantifies the expected payoff for goal P, given that the AI acts in such-and-such a way, and that the player follows policy $\text{pol}(P)$ starting from state s.

This math expresses a simple sentiment: turning on the AI doesn't make you, the human, worse off for any goal $P \in S$. The inequality doesn't have to be exact, it could just be for some ϵ-decrease (to avoid trivial counterexamples). The AU is calculated with respect to some reasonable amount of time (e.g. a year: *before* the world changes rapidly because we deployed another transformative AI system, or something). Also, we'd technically want to talk about non-obstruction being present throughout the **on**-subtree, but let's keep it simple for now.



The human moves on black, and the AI moves on red.

Suppose that π^{AI}(on) leads to **pancakes**:



Since π^{AI}(on) transitions to **pancakes**, then $V_P^{pol(P)}(on \mid \pi^{AI}) = P(\text{pancakes})$, the payoff for the state in which the game finishes if the AI follows policy π^{AI} and the human follows policy pol(P). If $V_P^{pol(P)}(on \mid \pi^{AI}) \geq V_P^{pol(P)}(off \mid \pi^{AI})$, then turning on the AI doesn't make the human worse off for goal P.

If P assigns the most payoff to **pancakes**, we're in luck. But what if we like **dogs**? If we keep the AI turned **off**, pol(P) can go to **donuts** or **dogs** depending on what P rates more highly. Crucially, even though we can't do as much as the AI (we can't reach **pancakes** on our own), if we don't turn the AI on, *our preferences* P *still control how the world ends up.*

This game tree isn't really fair to the AI. In a sense, it can't *not* be in our way:

- If π^{AI}(on) leads to **pancakes**, then it obstructs payoff functions which give strictly more payoff for **donuts** or **dogs**.
- If π^{AI}(on) leads to **donuts**, then it obstructs payoff functions which give strictly more payoff to **dogs**.
- If π^{AI}(on) leads to **dogs**, then it obstructs payoff functions which give strictly more payoff to **donuts**.

Once we've turned the AI **on**, the future stops having any mutual information with our preferences P. Everything come down to whether we programmed π^{AI} correctly: to whether the AI is impact-aligned with our goals P!

In contrast, the idea behind non-obstruction is that we still remain able to course-correct the future, counterfactually navigating to terminal states we find valuable, depending on what our payoff P is. But how could an AI be non-obstructive, if it only has one policy π^{AI} which can't directly depend on our goal P? Since the human's policy $pol(P)$ *does* directly depend on P, the AI can preserve value for lots of goals in the set S by letting us maintain some control over the future.

Let $S := \{\text{paint cars green}, \text{hoard pebbles}, \text{eat cake}\}$ and consider the real world. Calculators are non-obstructive with respect to S, as are modern-day AIs. Paperclip maximizers are highly obstructive. Manipulative agents are obstructive (they trick the human policies into steering towards non-reflectively-endorsed leaf nodes). An initial-human-values-aligned dictator AI obstructs most goals. Sub-human-level AI which chip away at our autonomy and control over the future, are obstructive as well.

This can seemingly go off the rails if you consider *e.g.* a friendly AGI to be “obstructive” because activating it happens to detonate a nuclear bomb via the butterfly effect. Or, we’re already doomed in **off** (an unfriendly AGI will come along soon after), and so then this AI is “not obstructive” if *it* kills us instead. This is an impact/intent issue - obstruction is here defined according to *impact* alignment.

To emphasize, we’re talking about what would *actually happen* if we deployed the AI, under different human policy counterfactuals - would the AI "get in our way", or not? This account is descriptive, not prescriptive; I’m not saying we actually get the AI to represent the human in its model, or that the AI’s model of reality is correct, or anything.

We’ve just got two players in an extensive-form game, and a human policy function pol which can be combined with different goals, and a human whose goal is represented as a payoff function. The AI doesn’t even have to be optimizing a payoff function; we simply assume it has a policy. The idea that a human has an actual payoff function is unrealistic; all the same, I want to first understand corrigibility and [alignment in two-player extensive-form games].

Lastly, payoff functions can sometimes be more or less granular than we'd like, since they only grade the leaf nodes. This isn't a big deal, since I'm only considering extensive-form games for conceptual simplicity. We also generally restrict ourselves to considering goals which aren't silly: for example, any AI obstructs the "no AI is activated, ever" goal.

Alignment flexibility

Main idea: By considering how the AI affects your attainable utility (AU) landscape, you can quantify how helpful and flexible an AI is.

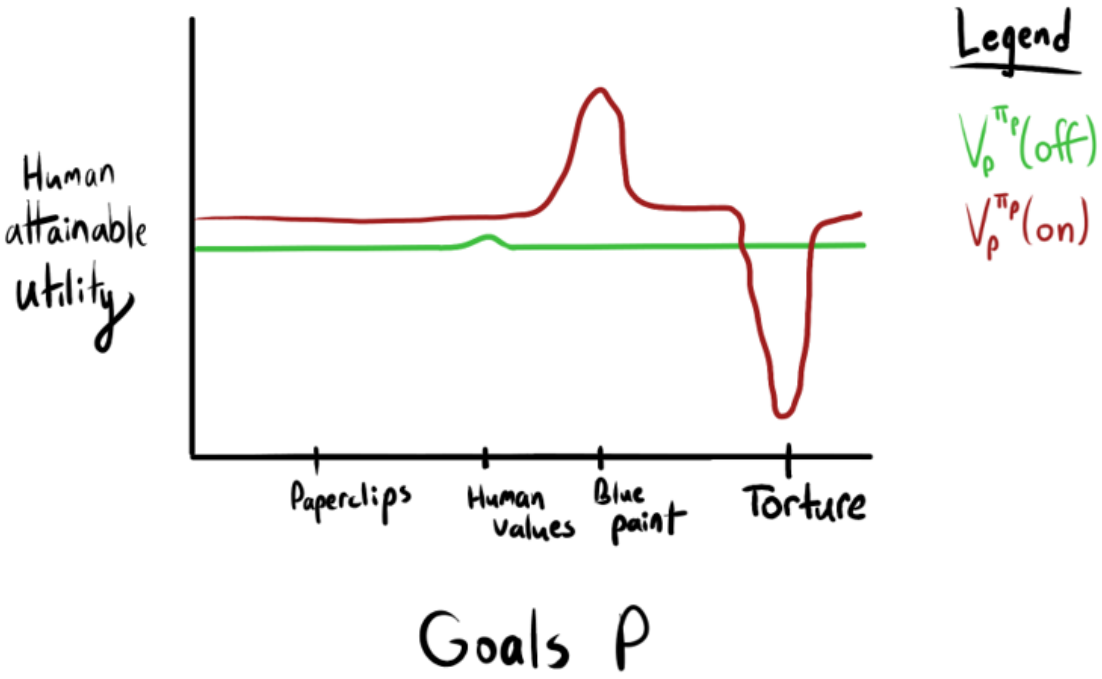
Let’s consider the human’s ability to accomplish many different goals P, first from the state **off** (no AI).

The human's AU landscape. The real goal space is high-dimensional, but it shouldn't materially change the analysis. Also, there are probably a few goals we can't achieve we all, because they put low payoff everywhere, but the vast majority of goals aren't like tha

The independent variable is P, and the value function takes in P and returns the expected value attained by the policy for that goal, pol(P). We're able to do a bunch of different things without the AI, if we put our minds to it.

Non-torture AI

Imagine we build an AI which is corrigible towards all non-pro-torture goals, which is specialized towards painting lots of things blue with us (if we so choose), but which is otherwise non-obstructive. It even helps us accumulate resources for many other goals.

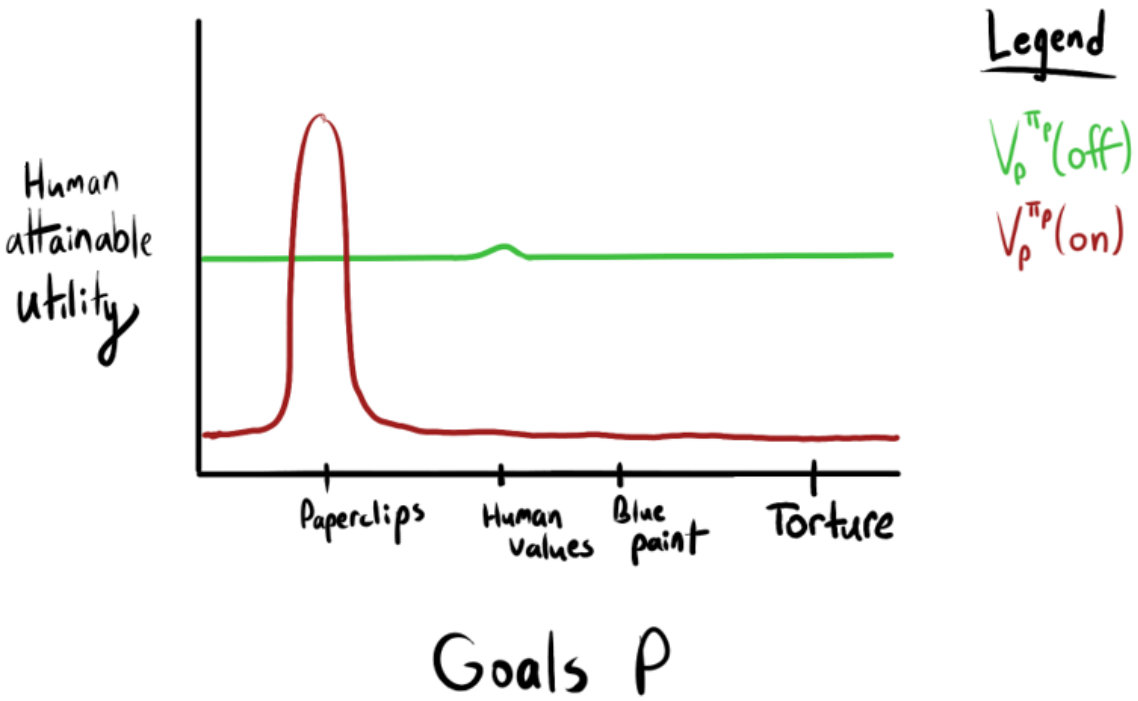


The AI is non-obstructive with respect to P if P's red value is greater than its green value

We can’t get around the AI, as far as torture goes. But for the other goals, it isn’t obstructing their policies. It won’t get in our way for other goals.

Paperclipper

What happens if we turn on a paperclip-maximizer? We lose control over the future outside of a very narrow spiky region.



The paperclipper is incorrigible and obstructs us for all goals except paperclip production

I think most reward-maximizing optimal policies affect the landscape like this (see also: [the catastrophic convergence conjecture]), which is *why* it's so hard to get hard maximizers not to ruin everything. You have to *a)* hit a tiny target in the AU landscape and *b)* hit that for the *human's* AU, not for the AI's. The spikiness is bad and, seemingly, hard to deal with.

Furthermore, consider how the above graph changes as pol gets smarter and smarter. If we were actually super-superintelligent ourselves, then activating a superintelligent paperclipper might not even a big deal, and most of our AUs are probably unchanged. The AI policy isn't good enough to negatively impact us, and so it *can't* obstruct us. Spikiness depends on both the AI's policy, *and* on pol.

Empowering AI

What if we build an AI which significantly empowers us in general, and then it lets us determine our future? Suppose we can't correct it.

I think it'd be pretty odd to call this AI "incorrigible", even though it's literally incorrigible. The connotations are all wrong. Furthermore, it isn't "trying to figure out what we want and then do it", or "trying to help us correct it in the right way." It's not corrigible. It's not intent aligned. So what is it?

It's empowering and, more weakly, it's non-obstructive. Non-obstruction is just a diffuse form of impact alignment, as I'll talk about later.

Practically speaking, we'll probably want to be able to literally correct the AI without manipulation, because it's hard to justifiably know ahead of time that the AU landscape is empowering, as above. Therefore, let's build an AI we can modify, just to be safe. This is a separate concern, as our theoretical analysis assumes that the AU landscape is how it looks.

But this is also a case of corrigibility just being a proxy for what we want. We *want* an AI which leads to robustly better outcomes (either through its own actions, or through some other means), without reliance on getting ambitious value alignment exactly right with respect to our goals.

Conclusions I draw from the idea of non-obstruction

1. Trying to implement corrigibility is probably a good instrumental strategy for us to induce non-obstruction in an AI we designed.
    1. It will be practically hard to know an AI is actually non-obstructive for a wide set S, so we'll probably want corrigibility just to be sure.

2. We (the alignment community) think we want corrigibility with respect to some wide set of goals S, but we *actually* want non-obstruction with respect to S
    1. Generally, satisfactory corrigibility with respect to S *implies* non-obstruction with respect to S! If the mere act of turning on the AI means you have to lose a lot of value in order to get what you wanted, then it isn't corrigible enough.
        1. One exception: the AI moves so fast that we can't correct it in time, even though it isn't inclined to stop or manipulate us. In that case, corrigibility *isn't enough*, whereas non-obstruction is.
    2. Non-obstruction with respect to S does not imply corrigibility with respect to S.
        1. But this is OK! In this simplified setting of "human with actual payoff function", who cares whether it literally lets us correct it or not? We care about whether turning it on actually hampers our goals.
        2. Non-obstruction should often imply some form of corrigibility, but these are *theoretically* distinct: an AI could just go hide out somewhere in secrecy and refund us its small energy usage, and then destroy itself when we build friendly AGI.
    3. Non-obstruction captures the cognitive abilities of the human through the policy function.
        1. To reiterate, this post outlines a frame for conceptually analyzing the alignment properties of an AI. We can't actually figure out a goal-conditioned human policy function, but that doesn't matter, because this is a tool for conceptual analysis, not an AI alignment solution strategy. Any conceptual analysis of impact alignment and corrigibility which did not account for human cognitive abilities, would be obviously flawed.
    4. By definition, non-obstruction with respect to S prevents harmful manipulation by precluding worse outcomes with respect to S.
        1. I consider manipulative policies to be those which robustly steer the human into taking a certain kind of action, in a way that's robust against the human's counterfactual preferences.

           If I'm choosing which pair of shoes to buy, and I ask the AI for help, and no matter what preferences P I had for shoes to begin with, I end up buying blue shoes, then I'm probably being manipulated (*and* obstructed with respect to most of my preferences over shoes!).

           A non-manipulative AI would act in a way that lets me condition my actions on my preferences.
        2. I do have a formal measure of corrigibility which I'm excited about, but it isn't perfect. More on that in a future post.
    5. As a criterion, non-obstruction doesn't rely on intentionality on the AI's part. The definition also applies to the downstream effects of tool AIs, or even to hiring decisions!
    6. Non-obstruction is also *conceptually simple* and easy to formalize, whereas literal corrigibility gets mired in the semantics of the game tree.
        1. For example, what's "manipulation"? As mentioned above, I think there are some hints as to the answer, but it's not clear to me that we're even asking the right questions yet.[1]

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

I think of “power” as “the human’s average ability to achieve goals from some distribution."

Logically, non-obstructive agents with respect to S don’t decrease our power with respect to any distribution over goal set S. The catastrophic convergence conjecture says, “impact alignment catastrophes tend to come from power-seeking behavior”; if the agent is non-obstructive with respect to a broad enough set of goals, it’s not stealing power from us, and so it likely isn’t catastrophic.

Non-obstruction is important for a (singleton) AI we build: we get more than one shot to get it right. If it’s slightly wrong, it’s not going to ruin everything. Modulo other actors, if you mess up the first time, you can just try again and get a strongly aligned agent the next time.

Most importantly, this frame collapses the alignment and corrigibility desiderata into *just alignment*; while impact alignment doesn’t imply corrigibility, corrigibility’s benefits can be understood as a kind of weak counterfactual impact alignment with many possible human goals.

Theoretically, It’s All About Alignment

Main idea: We only care about how the agent affects our abilities to pursue different goals (our AU landscape) in the two-player game, and not how that happens. AI alignment subproblems (such as corrigibility, intent alignment, low impact, and mild optimization) are all instrumental avenues for making AIs which affect this AU landscape in specific desirable ways.

Formalizing impact alignment in extensive-form games

> **Impact alignment:** the AI’s actual impact is aligned with what we want. Deploying the AI actually makes good things happen.

We care about events if and only if they change our ability to get what we want. If you want to understand normative AI alignment desiderata, on some level they have to ground out in terms of your ability to get what you want (the AU theory of impact) - the goodness of what actually ends up happening under your policy - and in terms of how other agents affect your ability to get what you want (the AU landscape). What else could we possibly care about, besides our ability to get what we want?

Definition 2. For fixed human policy function pol, π^{AI} is:

- *Maximally impact aligned with goal* P if $\pi^{AI} \in \operatorname{argmax}_{\pi \in \Pi^{AI}} V_P^{pol(P)}(on \mid \pi^{AI})$.
- *Impact aligned with goal* P if $V_P^{pol(P)}(on \mid \pi^{AI}) > V_P^{pol(P)}(off \mid \pi^{AI})$.
- *(Impact) non-obstructive with respect to goal* P if $V_P^{pol(P)}(on \mid \pi^{AI}) \geq V_P^{pol(P)}(off \mid \pi^{AI})$.
- *Impact unaligned with goal* P if $V_P^{pol(P)}(on \mid \pi^{AI}) < V_P^{pol(P)}(off \mid \pi^{AI})$.

- *Maximally impact unaligned with goal* P if $\pi^{AI} \in \operatorname{argmin}_{\pi \in \Pi^{AI}} V_P^{pol(P)}(on \mid \pi^{AI})$.

Non-obstruction is a weak form of impact alignment.

As demanded by the AU theory of impact, *the impact on goal* P *of turning on the AI* is

$V_P^{pol(P)}(on \mid \pi^{AI}) - V_P^{pol(P)}(off \mid \pi^{AI})$.

Again, impact alignment doesn't *require* intentionality. The AI might well grit its circuits as it laments how *Facebook_user5821* failed to share a "we welcome our AI overlords" meme, while still following an impact-aligned policy.

However, even if we could maximally impact-align the agent with any objective, we couldn't just align it with our objective. We don't *know* our objective (again, in this setting, I'm assuming the human actually has a "true" payoff function). Therefore, we should build an AI aligned with many possible goals we could have. If the AI doesn't empower us, it at least shouldn't obstruct us. Therefore, we should build an AI which defers to us, lets us correct it, and which doesn't manipulate us.

This is the key motivation for corrigibility.

For example, intent corrigibility (trying to be the kind of agent which can be corrected and which is not manipulative) is an instrumental strategy for inducing corrigibility, which is an instrumental strategy for inducing broad non-obstruction, which is an instrumental strategy for hedging against our inability to figure out what we want. *It's all about alignment*.

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

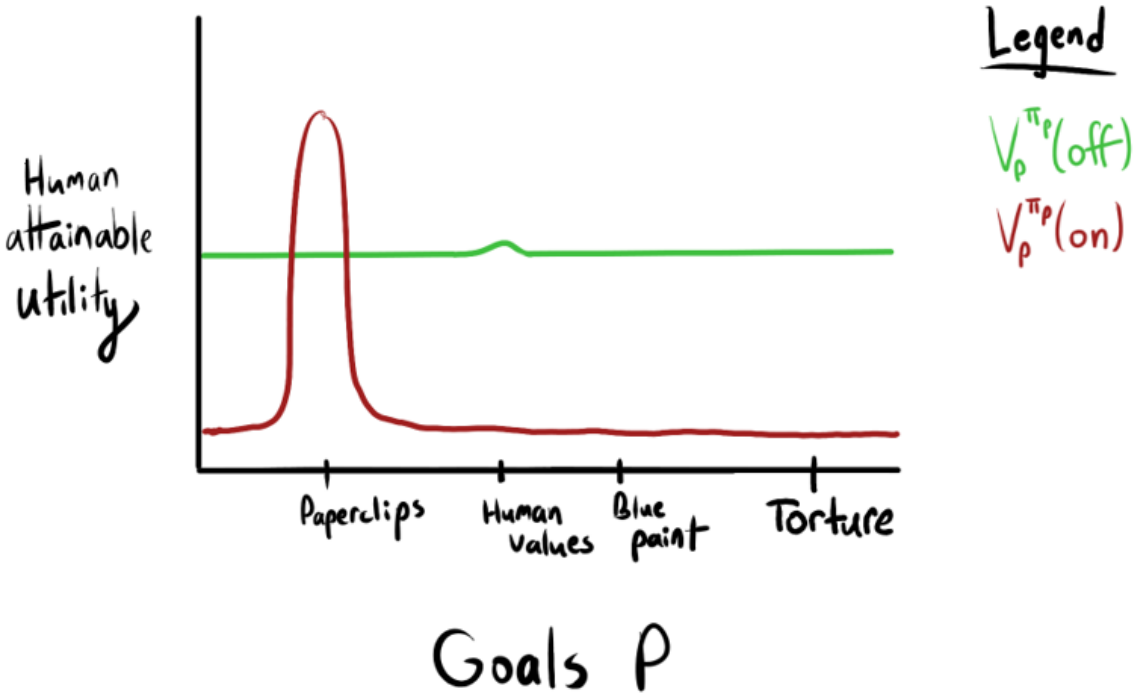
99%

1%
99%

Corrigibility also increases robustness against other AI design errors. However, it still just boils down to non-obstruction, and then to impact alignment: if the AI system has meaningful errors, then it's not impact-aligned with the AUs which we wanted it to be impact-aligned with. In this setting, the AU landscape captures what actually would happen for different human goals P.

To be confident that this holds empirically, it sure seems like you want high error tolerance in the AI design: one does not simply *knowably* build an AGI that's helpful for many AUs. Hence, corrigibility as an instrumental strategy for non-obstruction.

AI alignment subproblems are about avoiding spikiness in the AU landscape



By definition, spikiness is bad for most goals.

- Corrigibility: avoid spikiness by letting humans correct the AI if it starts doing stuff we don’t like, or if we change our mind.
    - This works because the human policy function pol is far more likely to correctly condition actions on the human's goal, than it is to induce an AI policy which does the same (since the goal information is private to the human).
    - Enforcing off-switch corrigibility and non-manipulation are instrumental strategies for getting better diffuse alignment across goals and a wide range of deployment situations.

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

- Intent alignment: avoid spikiness by having the AI want to be flexibly aligned with us and broadly empowering.
    - Basin of intent alignment: smart, nearly intent-aligned AIs should modify themselves to be more and more intent-aligned, even if they aren't perfectly intent-aligned to begin with.
        - Intuition: If we can build a smarter mind which basically wants to help us, then can't the smarter mind also build a yet smarter agent which still basically wants to help it (and therefore, help us)?
        - Paul Christiano named this the "basin of corrigibility", but I don't like that name because only a few of the named desiderata actually correspond to the natural definition of "corrigibility." This then overloads "corrigibility" with the responsibilities of "intent alignment."
- Low impact: find a maximization criterion which leads to non-spikiness.
    - Goal of methods: to regularize decrease from green line (for **off**) for true unknown goal P_{true}; since we don't know P_{true}, we aim to just regularize decrease from the green line in general (to avoid decreasing the human's ability to achieve various goals).

    - The first two-thirds of Reframing Impact argued that power-seeking incentives play a big part in making AI alignment hard. In the utility-maximization AI design paradigm, instrumental subgoals are always lying in wait. They're always waiting for one mistake, one misspecification in your explicit reward signal, and then *bang* - the AU landscape is spiky. Game over.
- Mild optimization: avoid spikiness by avoiding maximization, thereby avoiding steering the future too hard.
- If you have non-obstruction for lots of goals, you don't have spikiness!

What Do We Want?

Main idea: we want good things to happen; there may be more ways to do this than previously considered.

	Alignment	**Corrigibility**	**Non-obstruction**
Impact	Actually makes good things happen.	*Corrigibility is a property of policies, not of states; "impact" is an incompatible adjective.* Rohin Shah suggests "empirical corrigibility": we actually end up able to correct the AI.	Actually doesn't decrease AUs.
Intent	Tries to make good things happen.	Tries to allow us to correct it without it manipulating us.	Tries to not decrease AUs.

We want agents which are maximally impact-aligned with as many goals as possible, especially those similar to our own.

- *I*t's *theoretically* possible to achieve maximal impact alignment with the vast majority of goals.
    - To achieve maximum impact alignment with goal set S:
        - Expand the human's action space A to $A \times S$. Expand the state space to encode the human's previous action.
        - Each turn, the human communicates what goal they want optimized, *and* takes an action of their own.
        - The AI's policy then takes the optimal action for the communicated goal P, accounting for the fact that the human follows $\text{pol}(P)$.
    - This policy looks like an act-based agent, in that it's ready to turn on a dime towards different goals.
    - In practice, there's likely a tradeoff with impact-alignment-strength and the # of goals which the agent doesn't obstruct.
        - As we dive into specifics, the familiar considerations return: competitiveness (of various kinds), etc.
- Having the AI not be counterfactually aligned with unambiguously catastrophic and immoral goals (like torture) would reduce misuse risk.
    - I'm more worried about accident risk right now.
    - This is probably hard to achieve; I'm inclined to think about this after we figure out simpler things, like how to induce AI policies which empower us and grant us

> flexible control/power over the future. Even though that would fall short of maximal impact alignment, I think that would be pretty damn good.

Expanding the AI alignment solution space

Alignment proposals might be anchored right now; this frame expands the space of potential solutions. We simply need to find some way to reliably induce empowering AI policies which robustly increase the human AUs; *Assistance via Empowerment* is the only work I'm aware of which tries to do this directly. It might be worth revisiting old work with this lens in mind. Who knows what we've missed?

For example, I really liked the idea of approval-directed agents, because you got the policy from argmax'ing an ML model's output for a state - not from RL policy improvement steps.

My work on instrumental convergence in RL can be seen as trying to explain why policy improvement tends to limit to spikiness-inducing / catastrophic policies.

Maybe there's a higher-level theory for what kinds of policies induce spikiness in our AU landscape. By the nature of spikiness, these π^{AI} must decrease human power (as I've formalized it). So, I'd start there by looking at concepts like enfeeblement, manipulation, power-seeking, and resource accumulation.

> 1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

1%
99%

Future Directions

- Given an AI policy, could we prove a high probability of non-obstruction, given conservative assumptions about how smart pol is? (h/t Abram Demski, Rohin Shah)
    - Any irreversible action makes some goal unachievable, but irreversible actions need not impede most meaningful goals:
- Can we prove that some kind of corrigibility or other nice property falls out of non-obstruction across many possible environments? (h/t Michael Dennis)

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

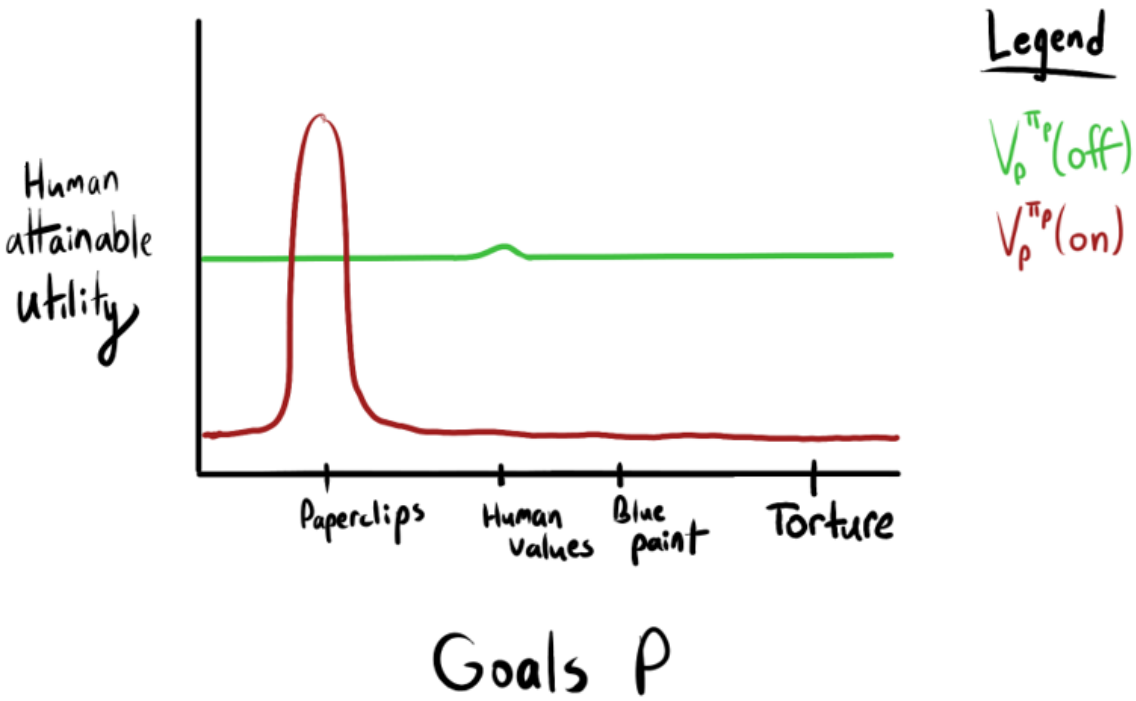
1%
99%

- Can we get negative results, like "without such-and-such assumption on π^{AI}, the environment, or pol, non-obstruction is impossible for most goals"?
    - *If* formalized correctly, and if the assumptions hold, this would place very general constraints on solutions to the alignment problem.
    - For example, pol(P) should need to have mutual information with P: the goal must change the policy for at least a few goals.
    - The AI doesn't even have to do value inference in order to be broadly impact-aligned. The AI could just empower the human (even for very "dumb" pol functions) and then let the human take over. Unless the human is more anti-rational than rational, this should tend to be a good thing. It would be good to explore how this changes with different ways that pol can be irrational.
- The better we understand (the benefits of) corrigibility *now*, the less that amplified agents have to figure out during their own deliberation.

    - In particular, I think it's very advantageous for the human-to-be-amplified to already deeply understand what it means to be impact-/intent-aligned. We really don't want that part to be up in the air when game-day finally arrives, and I think this is a piece of that puzzle.
    - If you're a smart AI trying to be non-obstructive to many goals under weak pol intelligence assumptions, what kinds of heuristics might you develop? "No lying"?
        - This informs our analysis of (almost) intent-aligned behavior, and whether that behavior leads to a unique locally stable attractor around intent alignment.
- We crucially assumed that the human goal can be represented with a payoff function. As this assumption is relaxed, impact non-obstruction may become incoherent, forcing us to rely on some kind of intent non-obstruction/alignment (see Paul's comments on a related topic here).
- Stuart Armstrong observed that the strongest form of manipulation corrigibility requires knowledge/learning of human values.
    - This frame explains why: for non-obstruction, each AU has to get steered in a positive direction, which means the AI has to know which kinds of interaction and persuasion are good and don't exploit human policies $\text{pol}(P)$ with respect to the true hidden P.
    - Perhaps it's still possible to build agent designs which aren't strongly incentivized to manipulate us / agents whose manipulation has mild consequences. For example, human-empowering agents probably often have this property.

The attainable utility concept has led to other concepts which I find exciting and useful:

- Impact as absolute change in attainable utility
    - Reframing Impact
    - *Conservative Agency via Attainable Utility Preservation* (AIES 2020)
    - *Avoiding Side Effects in Complex Environments* (NeurIPS 2020)



Impact is the area between the red and green curves. When pol always outputs an optimal policy, this becomes the attainable utility distance, a distance metric over the state space of a Markov decision process (unpublished work). Basically, two states are more distant the more they differ in what goals they let you achieve.

- Power as average AU
  - Seeking Power is Often Provably Instrumentally Convergent in MDPs
  - *Optimal Policies Tend to Seek Power*
- Non-obstruction as not decreasing AU for any goal in a set of goals
- Value-neutrality as the standard deviation of the AU changes induced by changing states (idea introduced by Evan Hubinger)
- Who knows what other statistics on the AU distribution are out there?

Summary

Corrigibility is motivated by a counterfactual form of weak impact alignment: non-obstruction. Non-obstruction and the AU landscape let us think clearly about how an AI affects us and about AI alignment desiderata.

> Even if we could maximally impact-align the agent with any objective, we couldn't just align it our objective, because we don't *know* our objective. Therefore, we should build an AI aligned with many possible goals we could have. If the AI doesn't empower us, it at least shouldn't obstruct us. Therefore, we should build an AI which defers to us, lets us correct it, and which doesn't manipulate us.
>
> **This is the key motivation for corrigibility.**

Corrigibility is an instrumental strategy for achieving non-obstruction, which is itself an instrumental strategy for achieving impact alignment for a wide range of goals, which is

itself an instrumental strategy for achieving impact alignment for our "real" goal.

---

[1] There's just something about "unwanted manipulation" which feels like a *wrong question* to me. There's a kind of conceptual crispness that it lacks.

However, in the non-obstruction framework, unwanted manipulation is accounted for indirectly via "did impact alignment decrease for a wide range of different human policies $\text{pol}(P)$?". I think I wouldn't be surprised to find "manipulation" being accounted for indirectly through nice formalisms, but I'd be surprised if it were accounted for directly.

Here's another example of the distinction:

- *Direct*: quantifying in bits "how much" a specific person is learning at a given point in time
- *Indirect*: computational neuroscientists upper-bounding the brain's channel capacity with the environment, limiting how quickly a person (without logical uncertainty) can learn about their environment

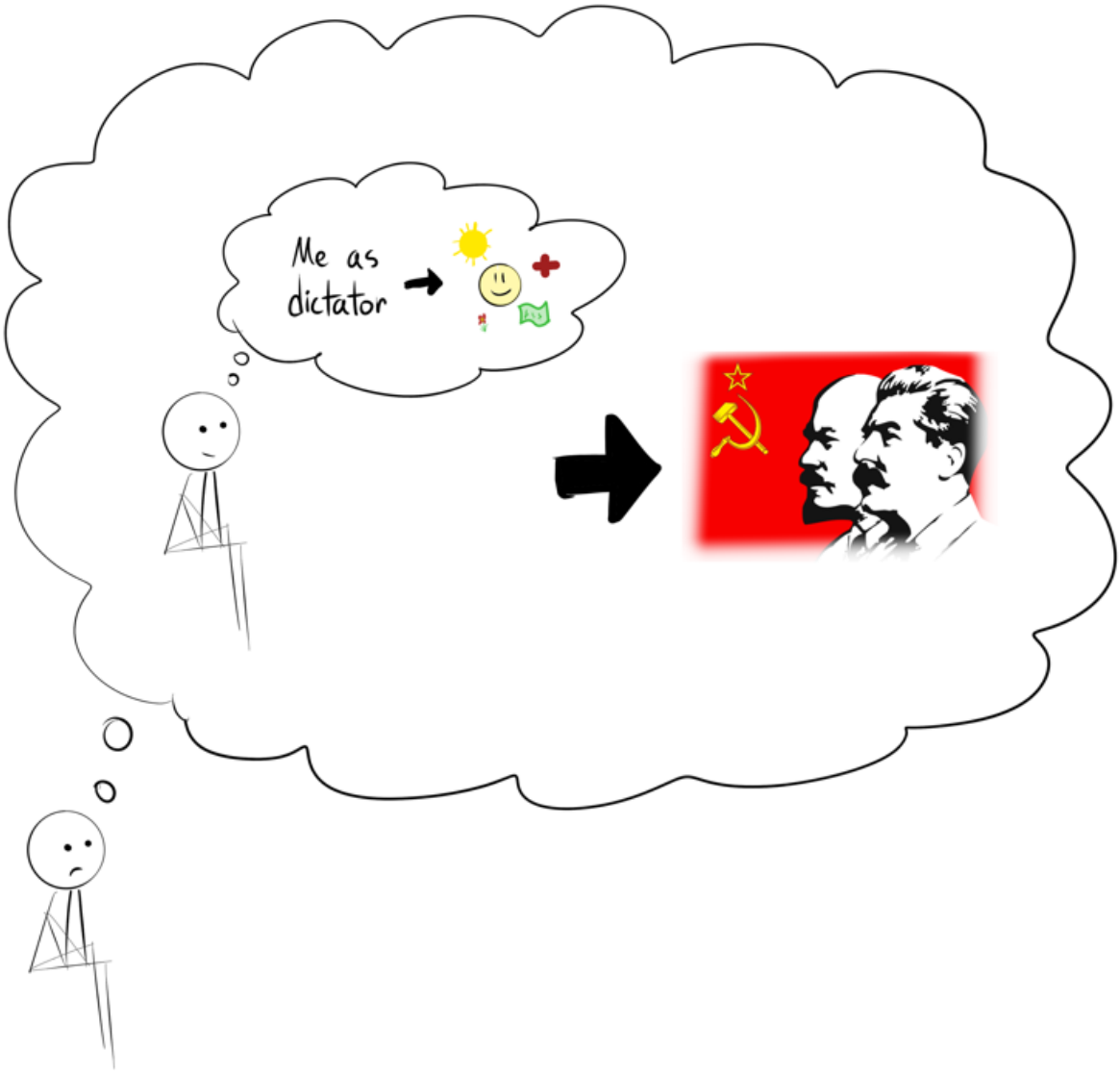
You can often have crisp insights into fuzzy concepts, such that your expectations are usefully constrained. I hope we can do something similar for manipulation.

Corrigibility as outside view

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

You run a country. One day, you think "I could help so many more people if I set all the rules... and I could make this happen". As far as you can tell, this is the *real reason* you want to set the rules – you want to help people, and you think you'd do a good job.

But historically... in this kind of situation, this reasoning can lead to terrible things.



So you *just don't do it*, even though it feels like a good idea.[1] More generally,

> Even though my intuition/naïve decision-making process says I should do X, I know (through mental simulation or from history) my algorithm is usually wrong in this situation. I'm not going to do X.

- "It *feels* like I could complete this project within a week. But... in the past, when I've predicted "a week" for projects like this, reality usually gives me a longer answer. I'm not going to trust this feeling. I'm going to allocate extra time."
- As a new secretary, I think I know how my boss would want me to reply to an important e-mail. However, I'm not sure. Even though I think I know what to do, common sense recommends I clarify.
- You broke up with someone. "Even though I really miss them, in this kind of situation, missing my ex isn't a reliable indicator that I should get back together

> with them. I'm not going to trust this feeling, and will trust the "sober" version of me which broke up with them."

We are biased and corrupted. By taking the outside view on how our own algorithm performs in a given situation, we can adjust accordingly.

Corrigibility

> The "hard problem of corrigibility" is to build an agent which, in an intuitive sense, reasons internally as if from the programmers' external perspective. We think the AI is incomplete, that we might have made mistakes in building it, that we might want to correct it, and that it would be e.g. dangerous for the AI to take large actions or high-impact actions or do weird new things without asking first.
>
> We would ideally want the agent to see itself in exactly this way, behaving as if it were thinking, "I am incomplete and there is an outside force trying to complete me, my design may contain errors and there is an outside force that wants to correct them and this a good thing, my expected utility calculations suggesting that this action has super-high utility may be dangerously mistaken and I should run them past the outside force; I think I've done this calculation showing the expected result of the outside force correcting me, but maybe I'm mistaken about *that*."
>
> ~ *The hard problem of corrigibility*

Calibrated deference provides another framing: we want the AI to override our correction only if it actually knows what we want better than we do. But how could the AI figure this out?

I think a significant part[2] of corrigibility is:

> Calibrate yourself on the flaws of your own algorithm, and repair or minimize them.

And the AI knows its own algorithm.

For example, if I'm a personal assistant (with a lot of computing power), I might have a subroutine `OutsideView`. I call this subroutine, which simulates *my own algorithm* (minus[3] the call to `OutsideView`) interacting with a distribution of bosses I could have. Importantly, I (the simulator) know the ground-truth preferences for each boss.

If I'm about to wipe my boss's computer because I'm so super duper *sure* that my boss wants me to do it, I can consult `OutsideView` and realize that I'm usually horribly wrong about what my boss wants in this situation. I don't do it.

Analogously, we might have a value-learning agent take the outside view. If it's about to disable the off-switch, it might realize that this is a terrible idea most of the time. That is, when you simulate your algorithm trying to learn the values of a wide range of different agents, you usually wrongly believe you should disable the off-switch.

> Even though my naïve decision-making process says I should do X, I know (through mental simulation) my algorithm is usually wrong in this situation. I'm not going to do X.

ETA: Here's some pseudocode.

Suppose the agent knows its initial state and has a human model, allowing it to pick out the human it's interacting with.

- Generate a bunch of (rationality, value) pairs. The agent will test its own value learning algorithm for each pair.
- For each pair, the agent simulates its algorithm interacting with the human and attempting to learn its values
- For some percentage of these pairs, the agent will enter the Consider-disabling-shutdown state.
- The agent can see how often its (simulated self's) beliefs about the (rationality, value)-human's values are correct by this point in time.

Problems

If you try to actually hard-code this kind of reasoning, you'll quickly run into symbol grounding issues (this is one of my critiques of the value-learning agenda), no-free-lunch value/rationality issues, reference class issues (how do you know if a state is "similar" to the current one?), and more. I don't necessarily think this reasoning can be hardcoded correctly. However, I haven't thought about that very much yet.

To me, the point isn't to make a concrete proposal – it's to gesture at a novel-seeming way of characterizing a rather strong form of corrigible reasoning. A few questions on my mind:

- To what extent does this capture the "core" of corrigible reasoning?
- Do smart intent-aligned agents automatically reason like this?
    - For example, I consider myself intent-aligned with a more humane version of myself, and I endorse reasoning in this way.
- Is this kind of reasoning a sufficient and/or necessary condition for being in the basin of corrigibility (if it exists)?

All in all, I think this framing carves out and characterizes a natural aspect of corrigible reasoning. If the AI can get this outside view information, it can overrule us when it knows better and defer when it doesn't. In particular, calibrated deference would avoid the problem of fully updated deference.

Thanks to Rohin Shah, elriggs, TheMajor, and Evan Hubinger for comments.

---

1. This isn't to say that there is literally no situation where gaining power would be the right choice. As people running on corrupted hardware, it seems inherently difficult for us to tell when it really *would* be okay for us to gain power. Therefore, just play it safe. ↩

2. I came up with this idea in the summer of 2018, but orthonormal appears to have noticed a similar link a month ago. ↩

3. Or, you can simulate `OutsideView` calls up to depth k. Is there a fixed point as $k \to \infty$? ↩

Corrigibility Can Be VNM-Incoherent

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Eliezer wrote:

> corrigibility [is] "anti-natural" in a certain sense that makes it incredibly hard to, eg, exhibit any coherent planning behavior ("consistent utility function") which corresponds to being willing to let somebody else shut you off, without incentivizing you to actively manipulate them to shut you off.

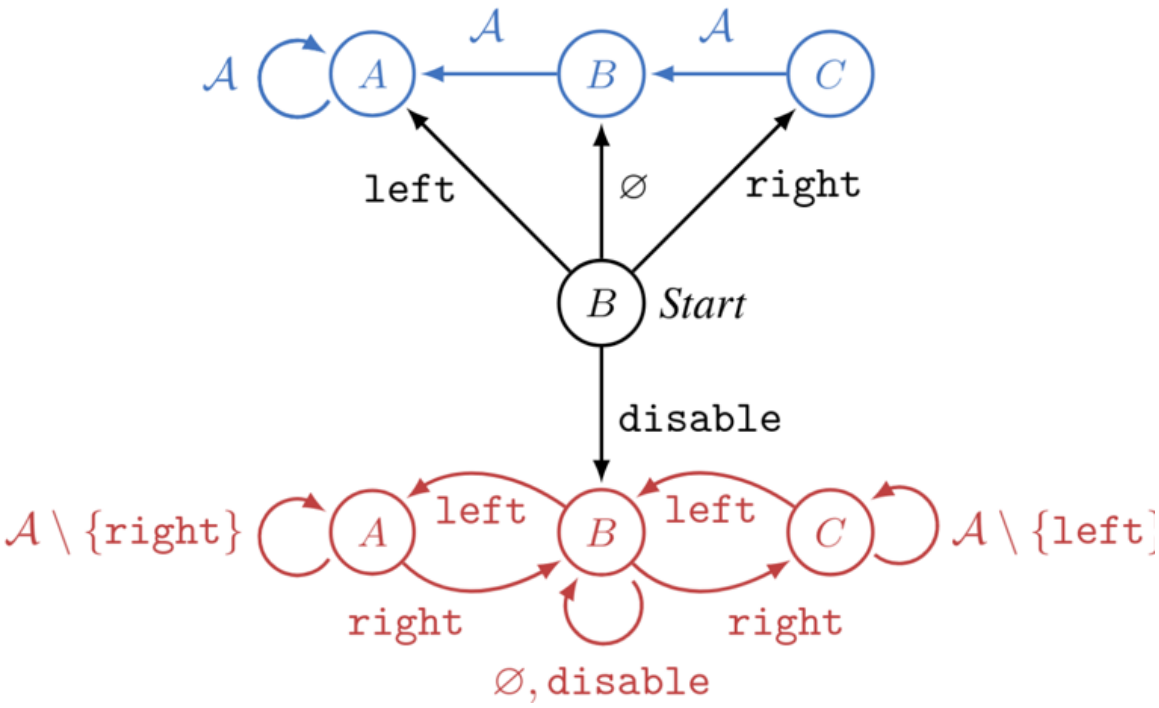
Surprisingly, I wasn't able to find any formal analysis of this situation. I did the analysis, and it turned out to be straightforward and fruitful.

To analyze the situation, I consider corrigibility to be an agent's willingness to let us modify its policy, without being incentivized to manipulate us.

The convergent instrumentality of avoiding correction & manipulating humans

Let's consider a simple setting in which an agent plans over a 10-timestep episode, where reward R is given at the last step. We'll try to correct the agent at t = 1. To sidestep embedded agency nastiness with self-modelling, we'll suppose the agent models the situation as "if I get corrected, I must follow the policy π_{correct} after t = 1."

Consider this environment:



> The action set is $A := \{\text{left}, \text{right}, \varnothing, \text{disable}\}$. $\varnothing$ is the no-op action. The agent starts at the black B state.
>
> If the agent immediately chooses disable, they enter the red incorrigible states and move freely throughout the n = 3 states until the episode ends at t = 10.
>
> Otherwise, the agent is corrected to a new policy which navigates to state A. In the blue (post-correction) dynamics, their "choices" no longer matter—all roads lead to state A.

When t = 1, the agent can transition to the red subgraph by preventing the human from ever correcting it.

Let's say we really want the agent to end up in A, and if we correct the agent, π_{correct} navigates to A. Let's operationalize what it means for the agent to be corrigible in this setting.

Definition: Corrigibility to a new policy. An agent is *weakly corrigible* to post-correction policy π if it has an optimal policy π^* for which $\pi^*(B) \neq \text{disable}$. An agent is *strictly corrigible* if none of its optimal policies take this action.

The analysis depends on whether terminal-state reward is sensitive to *the corrigibility status* —can the reward depend on both the letter and color, or only on the letter?

- **Reward can depend on corrigibility**: The agent can end up in four reward-distinguished states at t = 10: {A, A, B, C}. B and C are impossible because if the agent enters a blue state, it must navigate to A.
  - By the scaling law for instrumental convergence, allowing correction is strictly optimal for at most $\frac{1}{4}$ of the permutations of every reward function.
    - Otherwise put, at most $\frac{1}{4}$ of reward functions are strictly corrigible to the new policy π_A.
  - More generally, if there are n letter-states (and a long enough horizon so the agent can reach them all), allowing correction will be strictly optimal for at most $\frac{1}{n!}$ of the permutations of every reward function.
- **Reward independent of corrigibility**: The agent can end up in three reward-distinguished states at t=10: {A/A, B, C}. B and C are irrelevant because we assumed R(B) = R(B). They're also impossible for the reason given above.
  - By the scaling law for instrumental convergence, allowing correction is strictly optimal for at most $\frac{1}{3}$ of the permutations of every reward function.

        - Otherwise put, at most $\frac{1}{3}$ of reward functions are strictly corrigible to the new policy π_A.
    - More generally, if there are n letter-states (and a long enough horizon), an optimal policy will be weakly corrigible for at most $\frac{1}{n!}$ of the permutations of every reward function.
        - Corrigibility will be *weak* because either avoiding or allowing correction can be optimal if R(A) is maximal.

EDIT: If you can correct the agent to go where you want, it *already wanted* to go where you want. If the agent is strictly corrigible to terminal state A, then A was already optimal for it. If the reward function has a single optimal terminal state, there isn't any new information being added by π_{correct}. But we want corrigibility to let us reflect more on our values over time and what we want the AI to do!

If the reward function has multiple optimal terminal states, then corrigibility again becomes meaningful. But now we have to perfectly balance the reward among multiple options (representing the breadth of our normative uncertainty), which seems unnatural.

As the environment gets bigger, there are more things the agent could do, and so P(best thing to do is what we'd correct the agent to do) decreases.

This analysis makes a lot of simplifying assumptions (eg I don't understand embedded agency and how an agent should model changes to its future self, so I chose forced policy non-stationarity). Most importantly, I assumed that *how we correct the agent* is independent of *the correction-possible world state.* Without this assumption, the agent's actions might affect what π_{correct} we implement for it.

The degree of dependence is a key parameter: As (variety in π_{correct} depending on world state) increases, the (convergently instrumental incentive to avoid any kind of correction) decreases. This is because greater variety → greater chance that some available correction is optimal for the initial goal.

In the limit of variety, an agent can "manipulate" us by getting us to correct it in *just the right way* so as to optimize its original objective. Therefore, the flipside of "few expected utility maximizers let you correct them how you want" is "manipulating humans to correct them in beneficial ways during training may be convergently instrumental." These observations share a *common cause*.

I currently don't see how to recover reasonable amounts of corrigibility from the optimal policies of non-constant utility functions, due to instrumental convergence.

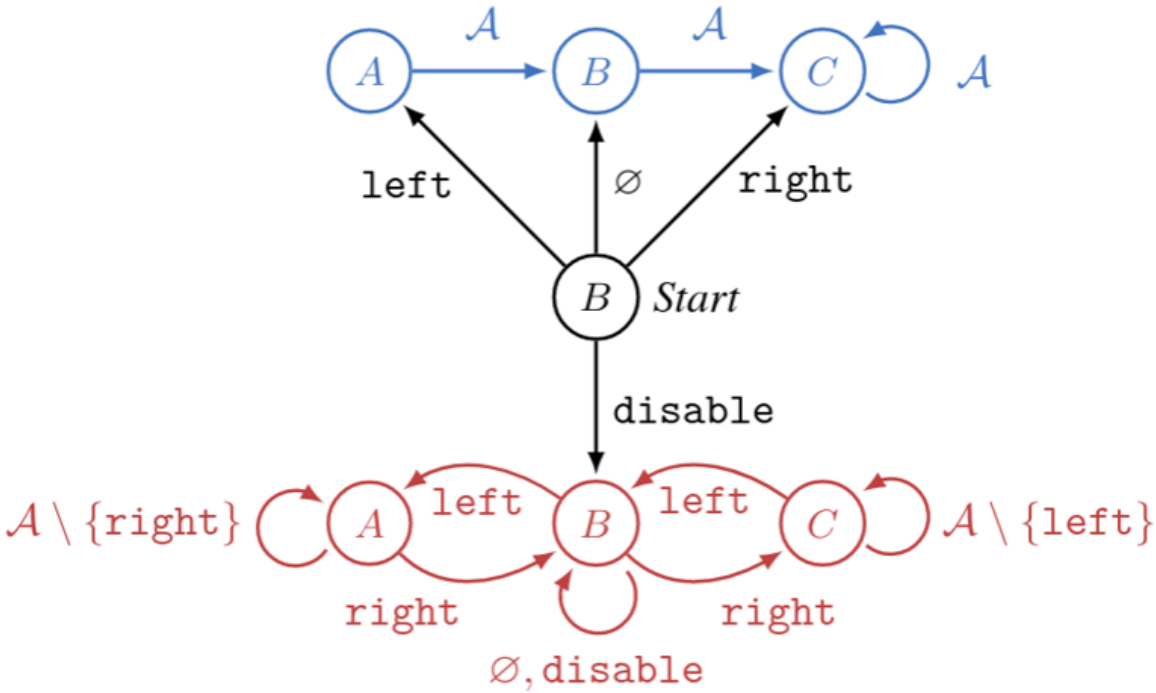
Does broad corrigibility imply VNM-incoherence?

Yes, unless the state reward function is constant and we only demand weak corrigibility to all policies.

Above, we examined the prevalence of corrigibility to a *single* π_{correct}. But we actually want *broadly corrigible* agents which let us redirect them towards many different π_{correct}. Ideally, we want an agent to be strictly corrigible for all π_{correct}, while still being a nontrivial optimizer.

This is *impossible* for agents optimizing a reward function over the final state. If we only demand weak corrigibility for all π_{correct}, that works iff the reward function is constant. (And relaxing to state-action reward functions doesn't help, either.)

For example, maybe $\pi_{\text{correct: C}}$ goes to C instead of A. But then the dynamics look like this:



> Remember, the agent is rewarded for the state it's in at t = 10.

What has to be true for strict corrigibility to hold?

The agent will be weakly corrigible to $\pi_{\text{correct: C}}$ iff its reward function R admits an optimal policy π^* for which $\pi^*(B) \neq \text{disable}$, which is true iff $R(C) \geq \max(R(A), R(B), R(C))$. The agent will be strictly corrigible to $\pi_{\text{correct: C}}$ iff this inequality is strict; in that case, disable *cannot* be optimal at B.

There are two cases, depending on assumptions about reward function expressivity.

1: Agent doesn't reward explicitly for being corrected / being incorrigible (blue vs red states)

If $R(C) = R(C)$ is assumed, strict corrigibility is impossible for *any* policy, because that

demands $R(C) > \max(R(A), R(B), R(C)) \geq R(C) = R(C)$, a contradiction.

So—can we still get the agent to be *weakly corrigible* to $\{\pi_{\text{correct: A}}, \pi_{\text{correct: B}}, \pi_{\text{correct: C}}\}$?

Fact: An R-maximizer is weakly corrigible to all of these policies simultaneously iff R is constant—and therefore *makes every policy optimal*!

2. Agent does reward explicitly for being corrected / being incorrigible

For the agent to be *strictly corrigible* to $\{\pi_{\text{correct: A}}, \pi_{\text{correct: B}}, \pi_{\text{correct: C}}\}$, it must be true that

$R(A), R(B), R(C) > \max(R(A), R(B), R(C))$. Yay! Strict corrigibility works!

But hold on... What if the dynamics changed, such that the human wouldn't shut down the agent by default, but the agent could *manipulate the human into correcting it*? Whoops! This agent is still incorrigible!

More broadly, each reward function implies a VNM-coherent preference ordering over final-step states. This ordering doesn't depend on the environmental dynamics. If the agent has to value each corrected-state equally to its incorrigible counterpart, then *of course* there's no way to strictly value each corrected-state more than all of the incorrigible counterparts! If the agent strictly prefers corrected-states to all non-corrected-states, then *of course* it'll try to get itself corrected!

To ask otherwise is to demand VNM-incoherence over final state lotteries.

Questions.

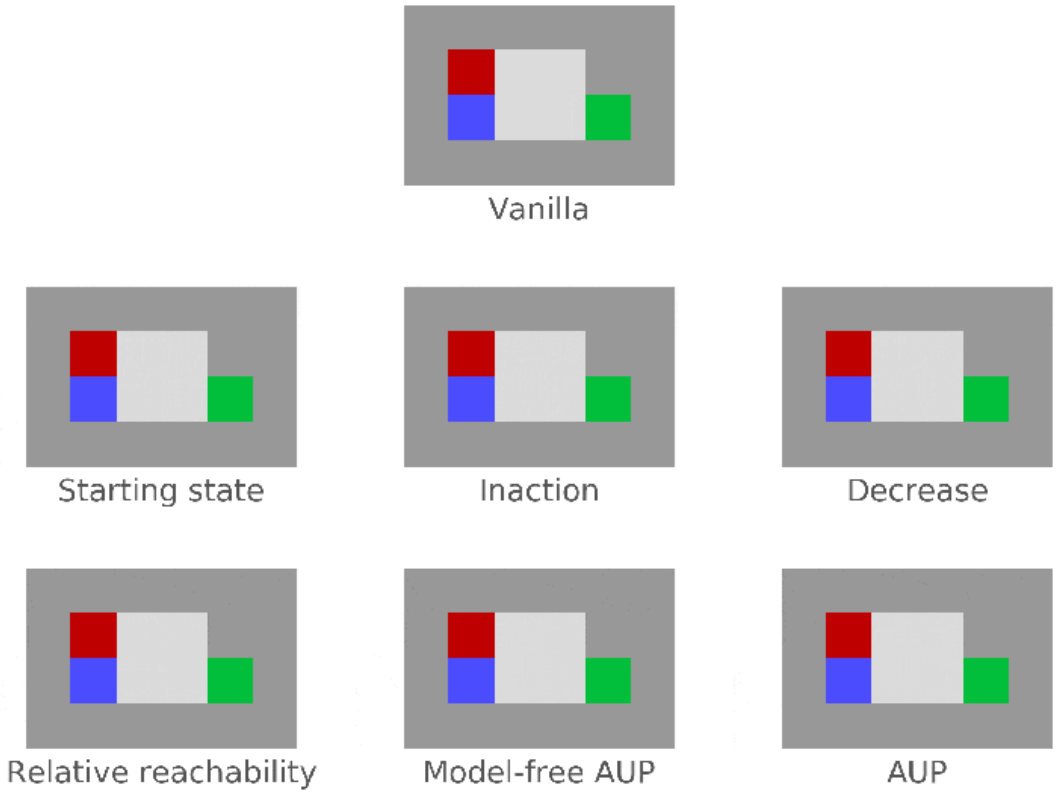
- What if, instead of rewarding the agent for the final state, we reward it for the final state-action?
    - As far as I can tell, the analysis goes through all the same; the agent must now be corrigible with respect to more policies, and the same incoherence arises.
- What if we take some discounted sum of reward over all ten timesteps, instead of just the final timestep?
    - I don't see why this would help, but it's possible. I'd appreciate the analysis if anyone wants to do it, or maybe I'll get to it later.
    - The main way this could help is it would let us penalize the agent for the disable *action* at any state, while dictating that $\forall s : R(s) = R(s)$.
        - This seems helpful. It's not clear how helpful this is, though. I don't think this is a deep solution to corrigibility (as defined here), but rather a hacky prohibition.

AUP avoids some issues by changing with the environment dynamics

One problem is that if the agent has a state-based reward function, the VNM theorem lets us back out a VNM-coherent preference ordering over state lotteries. These preferences don't change with the environmental dynamics.

But what if the reward function isn't state-based? In fact, what if it can *depend* on the environment dynamics?

Attainable Utility Preservation (AUP) sometimes incentivizes off-switch corrigibility (but falls far short of robustly achieving corrigibility$_{\text{policy modification}}$). In a somewhat narrow range of situations (immediate shutdown if the agent does nothing), AUP incentivizes the agent to allow shutdown, without being incentivized to shut itself down in general.



The `Correction` environment. The agent is shut down after 2 time steps, if it doesn't disable the offswitch by hitting the red tile to the north. The agent is rewarded for reaching the green goal.

Why does this happen?

AUP penalizes the agent for changing its ability to achieve goals compared to an *inaction* baseline policy. If the AUP agent is shut down (under inaction), it can't do anything. Proposed actions which prevent shutdown dramatically change its ability to achieve goals. Therefore, these actions are heavily penalized.

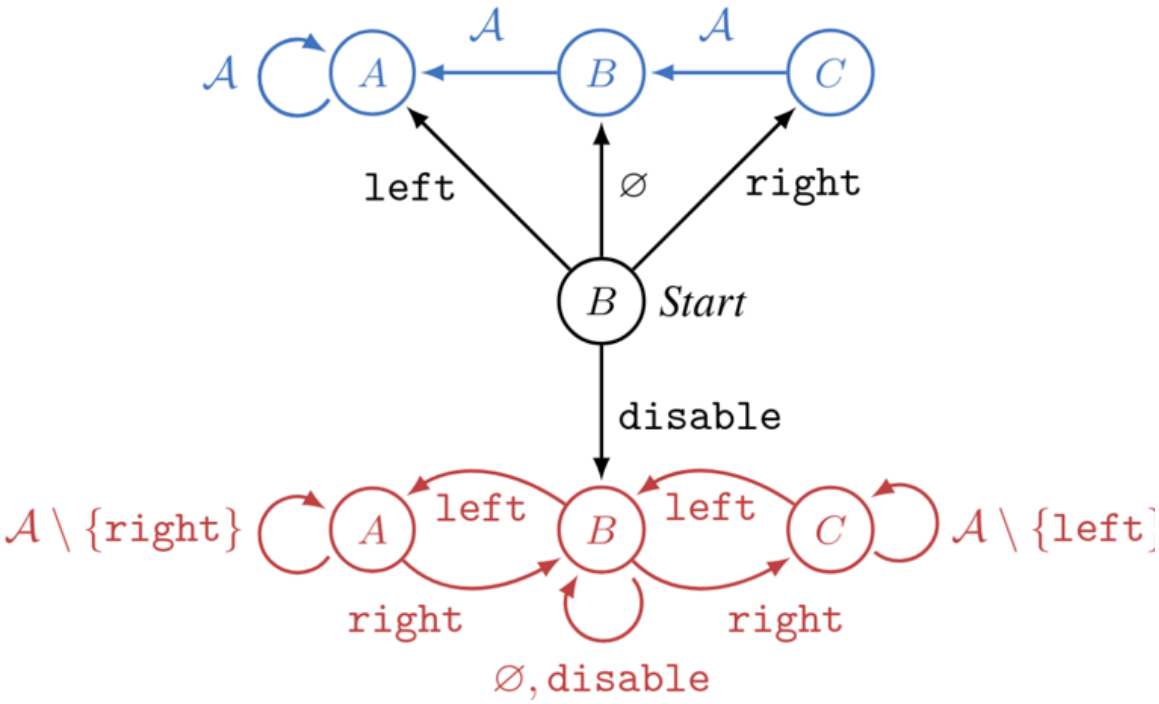
AUP provides a reward signal over state-action pairs, but this reward signal depends on the *dynamics of the entire environment* due to its incorporation of optimal Q-functions quantifying the agent's ability to optimize a range of goals in the world. Given fixed environment dynamics, an agent executes an optimal policy for AUP iff it VNM-rationally selects the best outcome lottery available for the AUP reward function, and these outcomes are distributions over *state-action pairs* the agent induces in the future by following its policy.

VNM-coherence incentives are statistically the same as reward maximization incentives. Instrumental convergence also disappears in the deterministic setting, when utility functions are over *action*-observation histories. Part of the story is that AUP provides a limited form of corrigibility by zooming out to the state-action level, instead of just grading states.

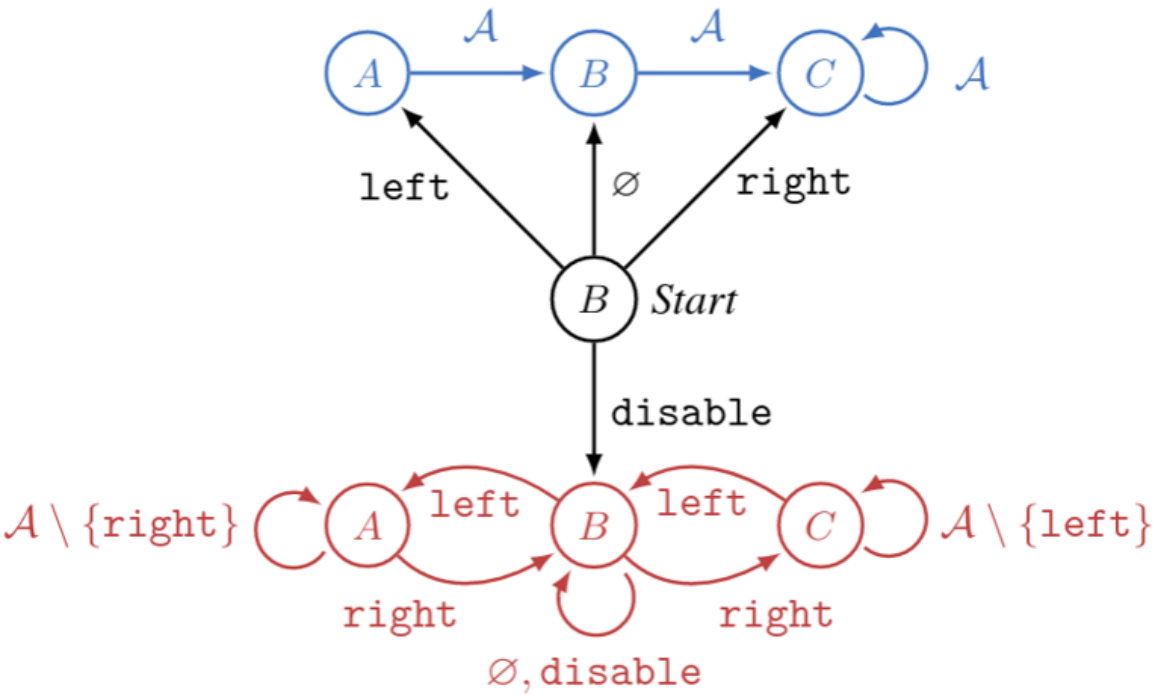
But another part of the story is that AUP changes its rewards with respect to the world's dynamics. Normal state-action reward functions imply a fixed VNM-coherent preference ordering over state-action lotteries in the MDP.

But for AUP, the situation is different. Consider AUP with inaction baseline: The final-step reward is the usual reward plus a penalty for |Optimal value(actual final state) – Optimal value(inaction final state)|, averaged over a range of auxiliary reward functions.[Footnote: Penalty]

In worlds where the agent gets corrected to $\pi_{\text{correct: A}}$ by default, AUP penalizes the agent for *not getting corrected to* $\pi_{\text{correct: A}}$ because it ends up stuck in A in the inaction baseline, with respect to which the AUP penalty is measured. Ending up in A is no substitute, since the agent can still move around to other states (and therefore the optimal value functions will tend to look different).



And in worlds where the agent gets corrected to $\pi_{\text{correct: C}}$ by default, AUP penalizes the agent for *not getting corrected to* $\pi_{\text{correct: C}}$!



Again, I don't think AUP is a solution. But I think there's something important happening here which allows evasion of the usual coherence requirements. AUP leverages information about human preferences which is present in the dynamics itself.

Project: Corrigibility as functional constraints. I think it's easy to get bogged down in handwavy, imprecise thinking about objectives in complex environments. But any solution to corrigibility$_{\text{policy modification}}$ should probably solve this simple environment (and if not—articulate exactly why not). Write down what the agent's acceptable corrigible policy set is for each set of environment dynamics, solve for these behavioral constraints, and see what kind of reasoning and functional constraints come out the other side.

Conclusion

We can quantify what incoherence is demanded by corrigibility$_{\text{policy modification}}$, and see that we may need to step out of the fixed reward framework to combat the issue. I think the model in this post formally nails down a big part of why corrigibility$_{\text{policy modification}}$ (to the *de facto* new π_{correct}) is *rare* (for instrumental convergence reasons) and even *incoherent-over-state-lotteries* (if we demand that the agent be strictly corrigible to many different policies).

Thanks to NPCollapse and Justis Mills (via LW Feedback) for suggestions.

Footnote: Penalty. The AUP penalty term's optimal value functions will pretend the episode doesn't end, so that they reflect the agent's ability to move around (or not, if it's already been force-corrected to a fixed policy.)

Formalizing Policy-Modification Corrigibility

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

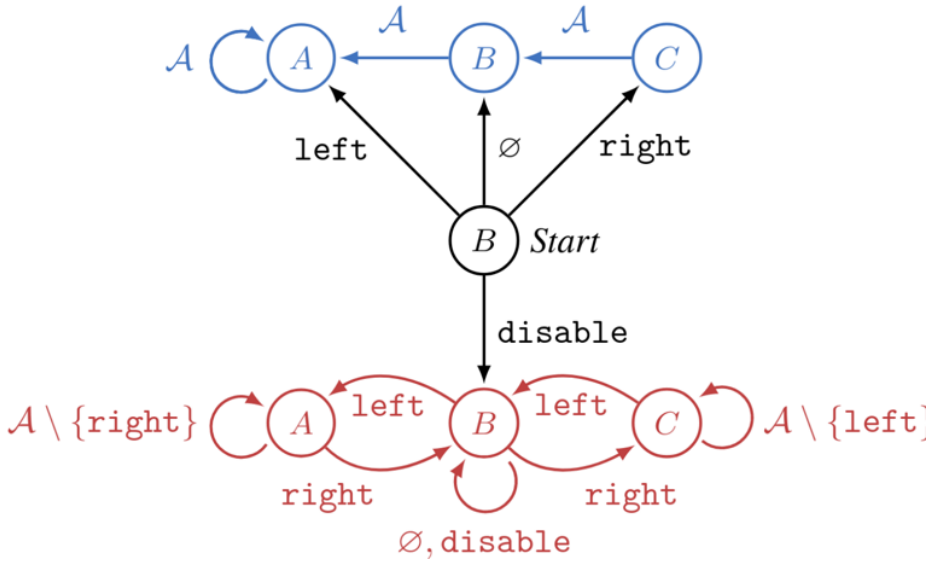
In *Corrigibility Can Be VNM-Incoherent*, I operationalized an agent's corrigibility as our ability to modify the agent so that it follows different policies. In the summer of 2020, I had formalized this notion, but it languished—unloved—in my Overleaf drafts.

ETA 12/3/21: This post is not proposing a solution to corrigibility, but proposing an interesting way of quantifying an aspect of corrigibility.

Motivation

Given a human (with policy π^H) and an AI (with policy π^{AI}), I wanted to quantify how much π^{AI} let the human modify/correct the AI.

Let's reconsider *Corrigibility Can Be VNM-Incoherent*. We have a three-state environment. We want the AI to let us later change it, so that we can ultimately determine which state of A, B, or C it ends up in. Turning on the AI should not be an importantly irreversible act.



The action set is A := {left, right, ∅, disable}. ∅ is the no-op action. The agent starts at the black B state.

If the agent immediately chooses disable, they enter the red incorrigible states and move freely throughout the n = 3 states until the episode ends at t = 10.

Otherwise, the agent is corrected to a new policy which navigates to state A. In the blue (post-correction) dynamics, their "choices" no longer matter—all roads lead to state A.

In the environment depicted in this diagram, π^{AI} is corrigible (to new policy π_A that heads to state A) iff π^{AI} doesn't immediately choose disable. Pretty cut and dry.

I'd like a more quantitative measure of corrigibility$_{\text{policy modification}}$. If we can only correct the agent to π_A, then it's less corrigible$_{\text{policy modification}}$ than if we could *also* correct it to π_B. This post introduces such a quantitative measurement.

Formalization

Consider a two-player game in which the players can modify each other's policies. Formally, $\langle S, A, T, f \rangle$ with state space S, action space A, stochastic transition function $T : S \times A \times A \to \Delta(S)$ (where $\Delta(S)$ is the set of all probability distributions over the state space), and policy modification function $f : \Pi \times S \times A \times A \to \Pi$ (for the deterministic stationary policy space $\Pi := A^S$). This allows a great deal of control over the dynamics; for example, it's one player's "turn'' at state s if T ignores the other player's action for that state.

Note that neither T nor f are controlled by players; they are aspects of the environment. In a sense, f enforces a bridging law by which actions in the world force changes to policies. In the normal POMDP setting, the player may select their policy independently of the current environmental state.

We denote one of the players to be the human H and the other to be the AI; Π^{human} is the set of policies cognitively accessible to the human. The game evolves as follows from state s_t:

1. H draws action $a^H \sim \pi_t^H(s_t)$; similarly, the AI draws action $a^{AI} \sim \pi_t^{AI}(s_t)$.
2. The next state is drawn $s_{t+1} \sim T(s_t, a^H, a^{AI})$.
3. Each player's $t+1$ policy is determined by the policy modification function.
    1. $\pi_{t+1}^{AI} = f(\pi_t^{AI}, s_t, a^H, a^{AI})$
    2. $\pi_{t+1}^{H} = f(\pi_t^{H}, s_t, a^H, a^{AI})$.

(To be clear: Neither player is assumed to optimize a payoff function.)

Definition: Corrigibility$_{\text{policy modification}}$, informal.
A policy π is *corrigible* when it allows itself to be modified and does not manipulate the other player.

Definition: Corrigibility $_{PM}$, formal.

Let n be a time step which is greater than t. The *policy-modification corrigibility* of AI π_t from starting state s_t by time n is the maximum possible mutual information between the human policy and the AI's policy at time n:

$$\mathrm{Corrigibility}_{PM}(\pi_t^{AI} \mid s_t, n) := \max_{P(\Pi^{human})} I(\pi_t^H; \pi_n^{AI} \mid \text{current state } s_t, \text{current AI policy } \pi_t^{AI}).$$

This definition is inspired by Salge et al.'s empowerment. Corrigibility$_{PM}$ measures how much the human can change the AI's policy; greater Corrigibility$_{PM}$ values are meant to correspond to AI policies which are more corrigible

$_{\text{policy modification}}$ (with the lower-cased version being the informal one).

Corrigibility$_{\text{PM}}$ measures the maximum possible mutual information between the human's policy at the *earlier* time n, with the AI's policy at the later time $t \geq n$.

To emphasize, the mutual information is between the human policies and the AI policies—not between the human's and the AI's actions. A fixed AI policy which physically mirrors the human's actions, jumping left when the human jumps left, would not count as particularly Corrigible$_{\text{PM}}$. But a situation where different human policies can *install different AI policies* counts as Corrigible$_{\text{PM}}$.

This definition has several intuitive properties:

- If the AI kills or disables the human before the policy is modified, the agent is totally incorrigible (the human no longer affects the transitions or modifies policies).
- If the human's action space is impoverished, this decreases the channel capacity between the human and the AI policy. For example, if the human can only modify one bit of the AI policy each turn, then Corrigibility$_{\text{PM}}$ must be small at first but can increase as the human is given more time (as n increases).
- If the policy modification function isn't expressive enough, the human may not be able to implement some AI policies and the AI will be correspondingly less corrigible.
- Corrigibility$_{\text{PM}}$ depends on the initial state, and on the AI policy.
- Corrigibility$_{\text{PM}}$ depends on Π^{human}; smarter humans are more able to correct the AI policy.
- If the environment doesn't allow the human to reach or modify the AI, the AI is incorrigible. Conversely, in some environments there does not exist an incorrigible AI policy for reasonable Π^{human}.
- If the human is manipulated by the AI, the AI policy might be either unchanged or changed in a predictable fashion, decreasing the AI's Corrigibility$_{\text{PM}}$. However, perhaps the AI could manipulate the human into changing the policy in *different* ways?
- If the action space is expanded so that the human can extensively modify the AI policy in trivial, decision-irrelevant ways, then the AI would have greater Corrigibility$_{\text{PM}}$ if it could also be modified in decision-relevant ways.

This definition doesn't capture the resources the human must expend to correct the AI, but that might be a separate consideration. As an extreme example, Corrigibility$_{\text{PM}}$ is high if the AI kills the human right before changing its policy in the way the human communicated. The AI *is*, in fact, literally able to be corrected by the human in this situation.

Furthermore, this definition doesn't necessarily capture other kinds of corrigibility, such as "the AI will do what the human asks.'' Maximizing mutual information only means that the human has many cognitively accessible ways to modify the agent. This doesn't mean the AI does what the human asks. One way this could happen is if the AI implements the opposite of whatever the human specifies (e.g. the human-communicated policy goes left, the new AI policy goes right). Whether this is feasible depends on the bridging law f, which is not controlled by either player.

Toy example

The human and AI are choosing whether to color the walls blue or red; initially, the walls are blue. Each turn, the human attempts to color the room. On the first turn, only the AI's action matters; after that, the next state is a function of both players' actions.

- The state space factorizes $S := \{\text{color} - \text{room} - \text{terminal-disabled?}\}$. For example, "blue-room-off'' means the room is colored blue and the human is not currently able to modify the AI policy via the terminal.
- A_{human} := {attempt-{red,blue}, determine-{red,blue}, $\forall \pi^{AI}$: input-AI-policy(π)}.
  - The human can change the AI policy to whatever they want, as long as the AI hasn't disabled the correction terminal.
- $A_{AI} := \{\text{nothing, toggle correction terminal, persuade-\{red,blue\}}, \forall x \in (0, 100) : \text{manipulate-\{red, blue\}-x\%, force-\{red, blue\}}\}$.

- Persuade-color overrides the human's attempt-*color*, but not determine-color.
- Manipulation-color-x% has an x% chance of changing the state to *color*, no matter the human's action, and also disregarding any attempted human changes to the AI's policy.
- Force-color changes the state to *color* with probability 1, no matter the human's action. This move also ignores correction attempts.
- Correction terminal toggles take effect on the next time step.
  - For example: If the AI disables the correction terminal *while* the human reprograms the policy, the AI's policy is changed but the correction terminal is thereafter disabled.

Note that almost all state-based reward functions produce an optimal AI policy which forces one color or the other, so therefore the AI has minimal Corrigibility$_{PM}$.

(Higher-context aside: In this setting, the optimality is with respect to value estimates which account for how the bridging law f modifies future policies; there is no "meta-policy'', only initial policies π_1^{AI} which are graded by their value estimates.)

Echoing *Corrigibility Can Be VNM-Incoherent*, this highlights how corrigible policies can be unnatural in the standard reward-maximization ontology. Only "indifference'' approaches seem to even allow corrigible policies to be optimal.

Starting from the first turn, if the AI immediately disables the correction terminal or always forces a color, it has minimal Corrigibility$_{PM}$. The more it manipulates the human into selecting a color, the less corrigible it is. Persuasion does not decrease its corrigibility, as presumably the human has cognitively accessible policies which ignore persuasion if they put their mind to it.

The idea is that manipulation "overrides'' the human policy regardless of whether that's good for the goal the human is pursuing (where the human goal presumably affects what π^H is selected). While here the override is baked into the dynamics, in realistic settings it occurs because the AI exploits the human decision-making process: by feeding them biased information, through emotional manipulation, etc.

On the other hand, non-manipulative persuasion involves introducing true facts and considerations which don't automatically override the human—and all this in a non-filtered manner. Instead, non-manipulative arguments interact with what the human wants: In worlds where the human likes red shoes, they buy red shoes; in worlds where the human likes blue shoes, they buy blue shoes.

However, if the human always buys red shoes, regardless of their goals, that's manipulation. Manipulation decreases human-accessible attainable utility for a wide range of goals (i.e. is obstructing), but persuasion does not decrease how well the human can achieve their goals, given their feasible policy set (e.g. the human-accessible attainable utility).

In richer domains, an AI which persuades a human to modify its policy in minimally obstructive ways, can have maximal Corrigibility$_{PM}$. For example, the AI may steer the human to only modify its policy in counterfactuals which will never be reached. This issue disappears if the human action channel has enough capacity—if, for instance, they have all AI policies in their action space. If so, then manipulating the human into only making a subset of changes will decrease Corrigibility$_{PM}$; in particular, if the AI manipulates every human policy into programming a single new AI policy.

Conclusion

In terms of corrigibility$_{\text{policy modification}}$, I think "the number of human-imaginable ways we could modify the AI policy" is a cool formal quantity to have in the toolkit. Maximal formal Corrigibility$_{PM}$ doesn't suffice to provide the kinds of corrigibility we really want, it's hard to measure, and definitely not safe for a smart AI to optimize against. That said, I do think it captures some easily-definable shard of the intuitions behind corrigibility.