Contents

1. Background

(TODO: we vs I?)

The Late 2021 MIRI conversions include discussion about the difficulty of alignment (don't worry, spending hours reading these isn't required for this post). One shared frame which wasn't immediately present in the discussion was a clean delineation of possible approaches to alignment and what they require. We claim that alignment techniques can be useful understood as deontological, utility focused, or capability restriction (or a mixture of these) and we'll be going through the challenges associated with constructing *sufficiently useful* and safe AI using these approaches. *(TODO: Not a fan of this wording, also better term than utility focused?)* After going through each of these approaches at a high level, I'll explain my take on the strategic implications of this analysis and attempt to craft a frame for analyzing alignment difficulty. We'll be focussing on X-risk, so we won't directly discuss failures which would 'only' result in large losses of life or economic damage. This is primarily a 'first principles' sort of analysis, though I'll be implicitly (and occasionally explicitly) referencing empirical work.

Epistemic status: exploratory. While many of the ideas stated here appear to be widely accepted in the alignment community, I'm a newcomer to the field trying

to cover a lot of ground. But everyone keeps telling me to be Very Ambitious and that alignment lacks strategic direction. So, uh, here goes an attempt at that I guess?

2. Prerequisites

The main prerequisite will be the sorts of concepts discussed in AGI safety from first principles.

We'll refer in more detail to:

- The idea of utility maximization and that coherent decisions imply consistent utilities
- Goodhart's law
- The concept of inner alignment
- Power seeking/instrumental convergence
- The current lack of understanding around deep learning generalization and transparency

Really, nothing else?

Well, other knowledge or ideas will hopefully be linked as necessary. Also, here are some posts which could be helpful to read (though I'm not sure if I would recommend reading them before or after reading this post):

- A discussion of using an objective framing or a generalization framing of inner alignment
- Model splintering: out of distribution behavior
- Reward splintering: model splintering on reward (really utility)

3. Assumptions and focuses

First of all, what is this *sufficiently useful* criteria we mentioned earlier? The criteria is that the alignment tax must be sufficiently small on the capability dimensions we care about. *(TODO: link to something on alignment tax)* And what is sufficiently small? And which dimensions? Well, I don't think we currently have a good understanding of this (as it requires predicting the future), but 2 typical models are:

1. Small enough that alignment can be enforced via governance without too much incentive for defection. This framing is probably more relevant in slow takeoff.
2. Small enough that an actor could use a lead in AI capabilities to accomplish a pivotal act *(TODO: this article isn't very focused, better one?)* safely before unaligned AIs are constructed. Note that under this framing, the 'capability dimensions we care about' are the ones which can be used to cause a pivotal act. If the alignment penalty makes all pivotal acts

impossible, then that technique is (approximately) worthless. This framing is more relevant in fast takeoff and the acceptable levels of alignment tax could depend on how far ahead the actor attempting to cause a pivotal act is. *(TODO: revise/cleanup language)*

For the remainder of this post, we'll abstract over this distinction in views, referencing different perspectives as necessarily.

But abstracting over everything results in a mess, so we'll make the following assumptions:

1. Unrestricted, superintelligent, and capable AGIs which act like long-term (bounded rationality) expected utility maximizers (aka Consequentialists) would cause an existential catastrophe if created with approaches reasonably similar to current ML. This is due to an inability to construct a human values utility function, an inability to perfectly align an agents utility function, Goodhart's law, and instrumental convergence.
2. Societal and government competence and coordination aren't very high (this informs how hard it is to enforce alignment through governance).

I won't make a case for why these are good assumptions here (because I'd guess most readers at least roughly already buy this). *(TODO: maybe make arguments in appendix? maybe link something?)* If you strongly disagree with these statements, well, maybe pick a different post to read.

We'll also mostly pretend AIs will be deep neural networks trained with SGD, but I wouldn't be surprised if this post generalizes.

4. Approaches

4.1 Deontological approaches

(TODO: maybe define/link deontology? I'm reluctant to pin down a short definition at the moment.)

4.1.1 Tool AIs are purely deontological AIs

Long-run expected utility maximization kills us all, so let's now consider AIs which don't care about optimizing their environments. Specifically we'll first consider AIs which have reduced agency: tool AIs. You may have noticed this appears in the deontological approaches section. That's because I claim that tool AIs (as typically described) are just *purely deontological* AIs. Agency is mostly just a set of capabilities coupled with (long-term) consequentialism. If wish to remove agency while keeping capabilities, we must remove consequentialism yielding a deontological AI. It may also be possible to reduce agency by removing some capabilities (such as self-modeling), this will be discussed in the section on restriction based approaches. Tool AIs are an extreme version of a deontological approach as they are *purely deontological*, but they serve as a good exhibit of the

weaknesses and safety advantages of deontological AIs as well as the challenges in constructing them.

(TODO: also serve as exhibit/lead in to some of main claims of post, maybe motivate better)

(TODO: maybe footnote here explaining that purely deontological AIs basically always have the features which are currently associated with the words 'tool AIs'. However, there is exception of very strange deontological principles like wanting to imitate a consequentialist etc... Maybe also description of tool vs process based task vs purely deontological and how uses of these terms has differed? Maybe put this footnote below where we discuss purely deontological AIs which basically act like agents? In general, I'm worried about language related confusion given how I first expand the notion of tools and then contract the notion of actually achievable purely deontological AIs below)

4.1.2 Purely Deontological AI, what is it good for

(TODO, ok maybe use actual title that references that this section is about safety properties...)

Given that different people use the term 'tool AI' in somewhat different ways, I will stick with the verbose *purely deontological* AI from here on.

Note that *purely deontological* AIs can be capable of modeling consequences, but they don't *care* about the consequences of their actions.[1] These means that *purely deontological* AIs can appear very agentic. For instance, consider a *purely deontological* AI which just cares about imitating the actions of a human. For a more absurd example, consider an AI which only cares about imitating what its actions would be if it here a expected value maximizer. For a competent imitator, this the *same* as being an expected value maximizer. So wait! Why have we bothered with defining this class of AIs if it practically includes expected value maximizers anyway!? Well, this come down to why the intentions of AIs matter at all. Intentions determine behavior when out of distribution for intelligent and robustly capable agents. *(TODO: maybe more here about how capability robustness probably fails safe instead of failing 'kill us all' so we shouldn't care as much about that for x-risk)* For example, consider some empirical observations of objective robustness failures in which agents 'care' about a correlated feature and then purse that feature when out of distribution instead of the reward from the original environment. There are also more arcane considerations like deceptive alignment which can leverage slight distributional differences into unsafety failures.[2] *(TODO: maybe justify that intentions determine generalization better or link something? Possibly I should just state this less confidently: I'm not that confident this is true in a deep sense.)*

[1] This may differ from how others use the term tool AI.

[2] Even if test inputs are within the training distribution (somehow), unless training is arbitrarily long there can still be safety concerns due to probabilistic treacherous turns.

So there can be purely deontological AIs which act like consequentialist agents in their training environments, but we generally expect them to act less like consequentialist agents on out of distribution inputs. In general, I would be Very Surprised if a highly capable, purely deontological AI caused existential catastrophe in a bid for power without having been trained explicitly to do so. It is isn't agency which scares us: it's generalizing agency.

In summary, purely deontological AIs are not existentially dangerous *by default*. They're in the same danger category as a nuclear bomb: they might kill a bunch of people, but you'll have to launch them first.

4.1.3 Constructing sufficiently useful purely deontological AIs

The next question we need to ask is how a *sufficiently useful* purely deontological AI can be constructed.

How could we train a purely deontological AI to do useful things? We could train to imitate or predict instead of optimizing for outcomes. Perhaps apply some iterative amplification or similar and boom, you've got an tool AI which do useful things.

Did you catch it?

The error in the above reasoning? Take a second and think through what's wrong before peeking.

Just because an AI is trained to imitate or predict doesn't mean it's guaranteed to be a purely deontological AI!

For instance, consider an AI trained to imitate a another AI which is a competent expected utility maximizer. It seems quite plausible that this imitator would itself just become an expected utility maximizer!

More generally, inner alignment is not guaranteed by all training procedures.

To be clear, I don't think this is a common misconception among people working on or thinking about alignment. However, it does seem like a potential trap, so I thought I would try to push readers away from the trap strongly.

So, there isn't an obvious way to train a purely deontological AI. In fact, we don't even know how to check how if an AI cares about consequences or deontological rules. Inner alignment with current machine learning approaches is hard. We have no physics style models for understanding the eventual intentions of super intelligent AI produced via such a process. *(TODO: ecological models?)* We don't have solid approaches for inspecting the cognition of deep agents. Or a decent understanding of what agent cognition will result from a specific training process. We don't know why or how deep learning generalizes. And it's unclear if techniques will generalize to higher intelligence and capability regimes. These issues are a theme for this entire post. My current view is that this is a difficult crux of alignment and we'll present only one (dangerous) way to proceed without

resolving these issues. *(TODO: clean up this word of difficult crux and maybe move earlier)*

That said, there are obvious ways to train deep neural networks which ensure that they will be purely deontological. For instance, consider training a (randomly initialized) model to output the value 1. Clearly such an model isn't going to be a consequentialist or even intelligent (unless you think the inductive biases of SGD are *actually* Magic). But, if the task in question might involve modeling consequences, the question of how to use current machine learning approaches to produce intelligent, non-consequentialist agents is considerably tricker.

In the superintelligent, highly capable regime, what sorts of training and objectives might produce purely deontological agents (as opposed to agents which are at least partially consequentialists)? Well, we're clearly deep into speculation land, because we don't even know how to produce a superintelligent, highly capable AI (and I wouldn't tell you even if I knew). However, I would be Very Surprised if training agents based on the consequences of their actions (outcomes) in even modestly complex environments with something resembling modern machine learning (e.g. reinforcement learning) resulted in purely deontological AIs. This is putting aside edge cases or the application of some not currently known technique. I'd also make a similar claim about AIs trained to imitate another expected utility maximizer AI. Note that constructing plans also falls into the category of outcome based training (assuming you care about whether or not those plans work!). Also be careful not to over generalize my statement: I'm just saying that you wouldn't get *purely* deontological agents, not that you couldn't get *partially* deontological agents which we will discuss later. So, this leaves the tasks which are classically associated with tool AIs such as prediction. We'll refer to these tasks as *process based* as opposed to *outcome based*. *(TODO: better name than process based, maybe a standard name which exists somewhere?)* So would process based tasks actually result in purely deontological AIs? I will hold off on speculating here, though I think the answer to this question would be useful. My understanding is that in this conversation Eliezer Yudkowsky says that he thinks that current machine learning techniques couldn't even produce an intelligent and purely deontological model. There's also some speculation in this post on safety in predictive learning

(TODO: add predication widget thing here for GPT-3 being purely deontological and GPT-n always being purely deontological.)

Now let's suppose that process based tasks do in fact result in purely deontological agents and consider if such agents can be *sufficiently useful*.

I'm not currently aware of any pivotal act which can be achieved using a process based task AI. A process based task AI could possibly help speed up alignment research, but probably not astronomically, so that isn't sufficient for a pivotal act.

(TODO: add predication widget thing here for existence of such an act.)

If purely deontological AI via process based tasks is the main approach to alignment enforced by governance, the benefits of defection seem large as tools want to be agents.

So overall, my belief is that trying to solve alignment for current ML via using purely deontological AIs is very unlikely to succeed.

4.1.4 Partially Deontological AI, what is it good for

Given the capability weakness of the purely deontological AIs we'd be able to create, perhaps we can tolerate some level of consequentialism, but instill some deontological properties. For instance, perhaps we'd like to instill deontological properties like honesty, obedience, corrigibility, or conservativeness which would override consequentialism in some cases or prohibit certain courses of action.

These deontological properties could result in incoherent decisions, but there's a problem we run into even before that: how the hell do we instill deontological properties? We're back to the difficult crux of alignment *(TODO: ok, maybe better term/language for this point)*. For any outcome based environment which rewards deontological properties, there exists an agent which simply models that deontological property as part of its utility function and achieves full marks. So maybe if you setup your deontological property encouraged environment and train to convergence you get a super intelligent, consequentialist, agent which also has the desired deontological property. But maybe not.

(TODO: talk about how difficulty in instilling deontological prefs could depend on intelligence which creates risky situation (compare ant to mouse to chimp to human or something))

There's an additional problem with partially deontological AIs which didn't exist with purely deontological AIs. If the deontological principles of an purely deontological AI splinter, the AI remains very unlikely to kill us all. It merely will have some other deontological properties potentially making the AI less useful.[3] However, if the deontological properties of an partially deontological AI splintered, but the consequentialist capabilities were retained, then it's possible that consequentialism wouldn't be overridden in important cases and the AI would kill us all.

Beyond all of these issues, we also now have to worry about the utility function of agent. While sufficient deontological properties could ensure that an AI with the wrong utility function didn't kill us all, it might not be very useful. Assuming the utility function of the AI was 'close enough' to desired, partially deontological AI could certainly be *sufficiently useful*. There can potentially be just as capable as pure expected utility maximizers. However, there are likely trade-offs between the strength of deontological properties and the capabilities of the agent. Sufficiently strong conservatism results in doing nothing at all.

[3] Unless for some reason deontological properties are likely to splinter into consequentialism?

(TODO: anything more here?)

4.2 Utility based approaches

4.2.1 Myopic Agents

Given that we assume that long-term expected utility maximizers would kill us all, what utility based approaches are left? Well, expected utility maximizers which don't care about the long run of course! These are typically described as myopic agents. Unfortunately, we currently don't know how to construct myopic agents: simply training agents with myopic reward is insufficient. This is the difficult crux of alignment discussed earlier once again. Even if we know how to construct myopic agents, they have potential safety issues mostly related to decision theory.

What about sufficient usefulness? *(TODO: speculation warning?)* The alignment tax will depend on how limited the time horizon must be to ensure safety. We'll assume that this time horizon maps onto real world time *(TODO: I'm uncertain if this is a reasonable modeling of myopic agents in general)*. The limitation on the time horizon depend on... First considering pivotal acts, it seems reasonably likely to me that killing us all takes around the same amount of time as a pivotal act. Thus, if a pivotal act must be ...

4.2.2 Getting utility closer to alignment

Even though long-term expected utility maximizers would kill us all, there's still value in the ability to produce agents with utility functions reasonably close to the desired one, particularly for complex or hard to specify alignment functions. This greatly effects the applicability of deontological and restriction based approaches.

4.3 Restrained AIs

The approaches to safety we've seen so far require solving this difficult crux of alignment (or getting lucky). Is there another way? Well, we could allow for fully blown, somewhat misaligned expected utility maximizing AI and restrain it in various ways.

For one example of a restrained AI, consider an AI which operates as a long-term goal achieving consequentialist in some sort of constrained environment (e.g. theorem proving[4]) without even knowing about the real world. Or consider AIs (hopefully) made safer by somehow removing their ability to model humans or another part of the world. Both of these AIs have had their capabilities handicapped to (hopefully) reduce danger. Additionally, the classic 'boxed' AI is an example of this approach and the typically proposed issues with this approach generalize to restrained AIs.

[4]I'm not claim that theorem proving requires long-term consequentialism, just that a long-term consequentialist could prove theorems.

A restrained AI could theoretically not be intent aligned while also being a highly capable consequentialist agent, somewhat impact aligned, and safe, However, any situation where the AI is aware of the restrictions placed on it will set up a conflict. The classic 'physically boxed' AI which is aware of the external world will manipulate and fight for freedom. Similarly, a 'knowledge boxed' AI will fight for missing knowledge if it's aware of the absence. This also applies to naively crafting AIs which are restricted by not knowing their utility function.

My current view is that it would be very difficult to prevent an AI which is aware of considerable parts of the real world from becoming aware of restrictions placed on it. Further, situating a self-modeling AI in an environment or task where it could encounter or reinvent the idea of simulation could result in the AI hypothesizing simulation and proceeding accordingly.

One possible route around these issues is making an AI which is incapable of understanding or using certain types of knowledge. Or similarly, preventing an AI from modeling itself. Unfortunately, both of these approaches seem very difficult with deep learning. We don't know how to prevent certain types of cognition or to control the cognition which results from training. Again, this is a very similar problem to the 'difficult crux' of alignment. *(TODO: clarify/revise this)*

In summary, we could either accept that the AI will become aware of restrictions and prepare for conflict or we could avoid conflict by solving a problem similar to the *difficult crux* or by severely limiting the possible tasks and environments to prevent the idea of simulation from being considered. Either way, it seems like purely restriction based approaches require a very large alignment tax or require embracing quite a bit of danger. My current (very uncertain) view is that the 'prepare for conflict' approach is a Very Bad Idea in all cases, so we'll rule it out when speculating about alignment tax. *(TODO: maybe something about how restriction could still be useful in combined setting?)*

From the pivotal act perspective, it seems very difficult to perform such an act with an AI which has sufficiently restricted knowledge or cognition about the world. However, the space of possible restrictions is large, so perhaps there are achievable ways to limit the cognition of AIs in order to ensure (or at least increase the probability of) safety while still being able to accomplish a pivotal act. I'm not currently aware of any promising directions.

The enforcement side looks similarly poor. Actors would likely perceive very strong incentives from increasing the capabilities of AIs.

4. Combining approaches

5. Strategic recommendations

A. Authors Note

Over the course of writing this post, I have noticed my views on the difficulty of alignment have shifted to be closer to that of Eliezer Yudkowsky. Perhaps this is an example of a general phenomenon: first principles contemplation of consequentialism, agency and intelligence leads to Yudkowskization (similar to carcinization).