Contents

1. Background

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. I claim that alignment techniques can be useful understood as deontological, consequentialist, 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: should 'consequentialist' category have a different name?)* I'll also be discussing and motivating a set of issues which I'll refer to as 'the hard problem of AI cognition': we don't have tools for understanding the cognition or intention of AIs produced by current machine learning methods or understanding of how cognition depends on training (c.f.

inner alignment).[1] After going through deontological, consequentialist, and capability restriction approaches at a high level and explaining where I think the hard problem of AI cognition is relevant, 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)

[1] This isn't a novel set of issues, but I haven't seen a thorough discussion of how these issues interact with various approaches to alignment. *(TODO: maybe this shouldn't be footnote?)*

Also note that I'm not necessarily claiming that it's difficult to craft an AI with specific intentions or cognition, just that we have no idea how to do so. *(TODO: is this needed? Maybe this doesn't clearly get across the idea I am intending?)*

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: anything to link with better/more focused discussion on alignment tax? Maybe also link/discuss theory practice gap?)* 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 the capabilities lead.

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, outcome expected utility maximizers (aka consequentialists) would cause an existential catastrophe if created with approaches reasonably similar to current ML. When I say 'outcome expected utility maximizer' I mean that the expected utility maximizer cares intrinsically about consequences (consequentialism) instead of about actions (deontology). We'll use the term consequentialism for the remainder of these post. We'll also go through this distinction in more detail below. This assumption is due to an inability to construct a human values utility function, an inability to perfectly inner align an agent's utility function, Goodhart's law, and instrumental convergence. *(TODO: could this be made clearer? Possibly remove some adjectives?)*
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 them). *(TODO: maybe make arguments in appendix? maybe link something?)* If you strongly disagree with these statements, please post your objections in the comments.

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

Deontological principles are rules for taking actions which aren't based on the consequences of those actions. In other words, deontological principles 'care' about actions instead of their consequences. Note that some deontological properties can be encoded or modeled using utility functions, but for others this might not be possible due to incoherent decisions. Confused or wondering about the implications of AIs having these principles? Hopefully the next sections will clarify this, so read on. *(TODO, I'm not a huge fan of this paragraph. Lots of room for improvement...)*

4.1.1 Tool AIs are purely deontological AIs

Long-run consequentialism 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 something like: We'll use tool AIs to motivate introducing a few of the core claims of this post.)

(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.[2] This 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 were a consequentialist. For a competent imitator, this the *same* as being a consequentialist. So wait! Why have we bothered with defining this class of AIs if it practically includes consequentialists 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.[3] *(TODO: maybe better justify that intentions determine generalization or link something? Possibly I should just state this less confidently: I'm not that confident this is true in a deep sense.)*

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 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.

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

[3] 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.

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 consequentialist. It seems quite plausible that this imitator would itself just become a consequentialist!

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. This is the 'the hard problem of AI cognition' which we'll be referencing throughout the 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.

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 consequentialist 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[4] 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: justify this claim)*

(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 would likely seem large to actors as tools want to be agents. *(TODO: Maybe more detail here? Or just lean entirely on Gwern? Maybe add statement like 'limiting the modeling of consequences greatly reduces usefulness')*

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

[4]Note that this depends on how intelligence is defined.

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 hard problem of AI cognition. For any outcome based environment which rewards deontological properties, there exists an agent which simply models that deontological property as consequences with some utility and achieves full marks. For instance, rewarding honesty could be modeled as 'be honest' or as 'appear honest to the overseer'. 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.

A further concern is that it might be easier to instill deontological properties in less intelligent agents. This could result from full blown consequentialism requiring relatively advanced capabilities like self-modeling, predication, and reasoning about counterfactuals. For instance, note that among life on earth intelligence seems to correlate with consequentialism. While ants (and other organisms) can look quite consequentialist from afar, this is really an emergent and less generalizable phenomenon than human consequentialism. *(TODO: maybe better better justify these claims. Note sure if ant sentences good/needed)* This creates the potential for quite a dangerous situation in which there is a smooth transition between dumb deontologist AIs and more intelligent purely consequentialist AIs which deceptively pretend to have deontological properties. Even if the transition isn't smooth, there is still potential for danger. When dialing up the intelligence knob (params, training time, etc), noticing a transition region between having deontological properties you want, some alignment failures, and then seemingly getting those properties back again should be a cause for alarm.

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.[5] However, if the deontological properties of an partially deontological AI splintered or were merely somewhat off, 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. We don't just need to ensure that we get deontological properties: we need to ensure we get the right deontological properties and those properties actually prevent existential catastrophe.

Beyond all of these issues, we also now have to worry about the utility function of agent with respect to consequences. 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

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

desired, partially deontological AI could certainly be *sufficiently useful*. There can potentially be just as capable as pure consequentialists. 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.

(TODO: anything more here?)

4.2 Consequentialist approaches

4.2.1 Myopic Agents

Given that we assume that long-term consequentialists would kill us all, what consequentialist approaches are left? Well, consequentialists which don't care about the long run of course! These are typically described as myopic agents[6]. Unfortunately, we currently don't know how to construct myopic agents: simply training agents with myopic reward is insufficient. This is the hard problem of AI cognition once again. However, I'd note that it would likely be easier *(TODO: maybe clarify what I mean by easier: higher success probability given equal resources (or maybe this is obvious?))* to create myopic agents then to instill complex deontological values. However, if we know how to construct myopic agents, they have potential safety issues mostly related to decision theory. Resolving these safety issues also likely requires grappling with the hard problem of AI cognition. Specifically, many issues could be resolved via producing an agent which follows dumb decision theory (DDT). My understanding is that getting an AI to use a specific decision theory is likely to be considerably harder than instilling deontological properties. Of course this would depend on the specific deontological property. In summary, just creating myopic agents might be easer than instilling deontological properties, but we might need to control the decision theory of myopic agents which could make the overall difficulty harder.

What about sufficient usefulness? 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 and that myopic agents follow dumb decision theory. Following dumb decision theory implies that the myopic agent will always pick the choice which maximizes utility: superrationality isn't a concern. So, the time horizon sufficient for safety depends on how fast the agent could seize power and the additional utility associated with possessing power. For instance, suppose that the time horizon was several years and that the AI could seize power in a few days and that the AI would be able to achieve much more utility over the remaining period than if it didn't seize power. This time horizon clearly results in existential catastrophe. Unfortunately, it's likely to be difficult to predict how much higher utility would be for a given agent if it possessed power (presumably utility would increase up over time with power). So, the time horizon would

[6] I've also seen purely deontological agents referred to as myopic, but here I'll be using myopic just to refer to consequentialists with little to no weight on long-run consequences.

likely need to be shorter or around the same time as it would take the AI to fully seize power. This is also difficult to predict, but it does seem to imply that enacting a pivotal act with a myopic AI would be difficult. This is because enacting a pivotal act would probably take around the same time as seizing power. However, perhaps the pivotal act could be 'split up' such that a myopic AI achieving high utility on the shorter term consequences of its actions would be sufficient.

Another potential power seeking concern is that the AI might not have time to seek all power but it would have time to seek some power (possibly causing a disaster) and this would result in higher utility. While it might be possible to stop an AI which has *only* acquired some power, if the AI wasn't stopped, it might keep acquiring some power continuously eventually resulting in existential catastrophe. The trade-off between utilisation and acquisition of power would depend on the discounting of the agent's utility function within the time horizon. This could push the time horizon down even further.

Given how short the time horizon would have to be, the usefulness analysis of myopic agents proceeds very similarly to that of purely deontological agents executing process based tasks. It would be very difficult to enact a pivotal act and the benefits of defection would likely seem large when trying to enforce alignment via governance. In the context of governance, there is also a slippery slope concern if agents become discontinuously dangerous at some time horizon. Raising the time horizon would initially appear safe. Until it isn't.

In summary, creating safe myopic agents is likely to be harder than instilling deontological properties (due to issues with decision theory) and the alignment tax is also likely to be higher. Overall, I think myopic agency isn't as promising as deontological approaches.

4.2.2 Getting utility closer to alignment

Even though long-term consequentialists would kill us all, there's still value in the ability to produce agents with consequentialist utility functions reasonably close to the desired one, particularly for complex or hard to specify utility functions. This greatly effects the applicability of deontological and restriction based approaches. Additionally, reward models or classifiers for complex or hard to specify behavior could be very useful building blocks for instilling deontological properties. This also implies that knowing how to construct these components would be useful for researching how to instill complex deontological properties.

I think that (unaligned) corporate and academic research might end up meaningfully contributing to this topic which reduces the marginal benefit of aligned work in this space. This reasoning is less applicable in scenarios where timelines and/or takeoff are much faster than corporate leadership expects.

4.3 Restrained AIs

The approaches to safety we've seen so far encounter the hard problem of AI cognition (or require getting lucky). Is there another way? Well, we could allow for fully blown, somewhat misaligned long-term consequentialist 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[7]) 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.

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. But now we've bumped back into the hard problem of AI cognition.

In summary, we could either accept that the AI will become aware of restrictions and prepare for conflict or we could avoid conflict by tangling with the hard problem of AI cognition 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 embracing quite a bit of danger. My current (very uncertain) view is that the 'prepare for conflict' approach is a Very Bad Idea, so we'll rule it out when speculating about alignment tax. We'll reconsider this when contemplating combining approaches to alignment.

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

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

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

(TODO: unfinished section, just outlined)

First of all, ... testing with restrictions and hopefully begin prepared for malicious ness. Likely to be fooled by deception. Advantage of things which reduce deceptiveness (myopic/deontological honest). Even non-deontological honest could be useful if AI doesn't think it could take over and decides not to engage in acausal trade.

Myopia + purely deontological not a thing myopia + deontological properties don't obviously synergy or anti-synergy in general.

5. Strategic recommendations

I was broadly in favor of prosaic alignment work before starting to write this post and I remain so. However, I have updated very slightly against prosaic alignment due to a considerable increase in my estimation of alignment difficulty, (more information in the author's note below). These recommendations will focus on prosaic alignment, though I won't make the case for prosaic alignment here.

My recommendations here are hopefully reasonably robust to differing views in takeoff speeds and modest changes in timelines. However, conditioning on very short (<5 years) or very long (>40 years) would probably change the analysis considerably.

While I'm in favor of working on aligning prosaic AI, I think we should actively try to discover new facts about intelligence. Specifically, I think that the alignment community is working too little on the hard problem of AI cognition. I'll propose an idea for a project and describe how I think the research focuses of the alignment community should change including examples of over and under rated topics, but first I think it's illustrative to go through an example of a project which I think is worthwhile and specific examples of additional sub-experiments I think are particularly valuable to conduct.

5.1 Redwood Research's Project

An example of work on instilling deontological properties is Redwood Research's project on getting a language model to never describe someone getting injured. It's probably a good idea to be familiar with this project before reading the rest of this section.

(Disclaimer: I'm TAing for the bootcamp Redwood Research is running and also interested in working there)

As of the time when this post was written, this project has just involved training a classifier for the predicate 'the completion describes injury'. But, the described plan is to use this classifier to train a language model which never violates this predicate. If totally successful, this would be a deontological property of the model. More generally, training classifiers to *always* identify bad behavior is useful preliminary work for instilling deontological properties.

I think a particularly useful place to extend this project is in trying to understand the changes in cognition and resulting generalization properties of various training strategies to produce models which don't violate the predicate. Here are some example experiments which I think get at this notion:

1. Use this classifier to train a language model from scratch which never violates this predicate, but only train on a dataset which isn't very diverse (for instance, only one genre of fanfiction). Does the property generalize to out-of-distribution inputs? What about adversarially crafted inputs? How does this vary as the training set is made more diverse?
2. Instead of training from scratch, instead use a pretrained language model (ideally pretrained or finetuned on fanfiction) and run the same tests described above.
3. Try to interpret and understand the difference in model parameters after training the pretrained model never to violate the predicate. It might be useful to penalize the distance in model weights from the original model.
4. Repeat above experiments but for different model sizes or training strategies and analyze which factors are relevant in how the model generalizes or changes its cognition.

Of course, there are likely cheaper and more effective ways to run experiments which get at the same ideas.

5.2 Ideas for some experiments on RL agent decision making

One domain of experimentation I'm particularly interested in is determining how decision making is modeled inside of RL agents. I'll go through an outline of an example project, but it should be straight forward to think of variants or other experiments which get at the same domain.

Construct a complex and randomly generatable environment which requires trading off various different objectives and probabilities of failure rewarding

consequentialism and ideally requiring higher levels of 'intelligence'. Add an obstacle/trap to this environment which happens to never be positive expected utility to pass through due to a large negative reward and never having a sufficiently positive reward accessible only through this obstacle. Train an agent to succeed in this environment and then see if it will pass though this obstacle/trap under any circumstances. Specifically, test if it will pass through if an out-of-distribution environment is constructed for which a positive reward larger than the obstacle penalty is visible behind the obstacle. Additionally, test how much positive reward is sufficient (if any). Hopefully, this would test if the agent is modeling the obstacle deontologically or as a negative reward in its utility function.

Here are some extensions/variants of this project which I think could worthwhile (depending on results from the base experiment):

1. Vary the complexity of the environment, training time, or the model size and architecture and determine if results change.
2. Does avoiding value function parameter sharing change anything?
3. Pre-train the agent in the same environment, but with positive reward possible for passing through the obstacle in some cases. Then, train the agent on the actual environment in which it's never a good idea to pass through the obstacle. Does this change results? How do results change as pre-training and final training times are varied?
4. Have multiple types of positive reward in the environment. Occasionally, make it appear as though net positive reward can be obtained by passing through the obstacle, but the reward is actually a lie in just this case and is still net negative. Do this using only one of the types of reward. Then, run the out-of-distribution test for passing through the obstacle using the other type of reward.
5. Try to understand where and how the utility functions is stored in the model perhaps using techniques like in the Understanding RL vision paper. Ideally it should be possible to edit the utility function by changing model weights 'by hand'. Can these techniques be used to adjust the behavior of the model with respect to the obstacle?

5.3 How focus should change

It's worth noting that my analysis here is highly speculative. It's somewhat difficult for me to model how aligned individuals working in technical AI safety are distributed. So maybe claiming there should be less focus on a specific topic is basically irrelevant because individuals working on this topic care about something other than existential risk or have very different empirical views. Similarly, I don't have a good model of how much effort is going into various topics or the difficultly of various topics. Perhaps loads of people are spending lots of time working on interpretability work, but there doesn't appear to be much work here merely because the topic is so challenging. I also don't have a good model of the skills of aligned individuals. It's possible that shifting focuses

in the way I recommend would result in people working in domains for which they have less skill or aptitude which could make shifting net harmful. As such, individuals, teams, and organizations should take their comparative advantage into account: increased comparative advantage in the current topic should push against changing focuses. Despite all of these issues, I still think this analysis is worthwhile.

I think the important work lies on instilling deontological properties, understanding the cognition likely to result from various training methods/environments and building techniques for transparency and interpretability, particularly of cognition. Its worth noting that developing architectures...

Overall: change from make model do X to gain insight into what sorts of changes in cognition occur when making model do X.

I think effort should be reallocated away from on crafting reward models for more complex or harder to write utility functions. For examples of this type of research, consider Learning from human preferences. This includes work on active learning, improving sample efficiency, and other related topics. This is based on the understanding that a large amount of research ...

Overrated:

-
-
-

Underrated:

-
-

Higher uncertainty but could be good:

-
-

Overall,

6 Alignment difficulty

A. Authors Note

(TODO: maybe move these views elsewhere?)

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).