Contents

1 Background

The late 2021 MIRI conversations 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 usefully 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. 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 AI decision making depends

on training (c.f. inner alignment).[1] I'll define this problem in more detail below, including where this problem does and doesn't show up in alignment. 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 briefly discuss 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)

If you'd like to read the late 2021 MIRI conversations (and haven't read them yet), my weakly suggested reading order is:

[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 and the strategic implications.

Also note that I'm not claiming that it's necessarily difficult to craft an AI with specific intentions or cognition, just that we have no idea how to do so.

- This post
- Zvi's gears analysis of AGI intervention
- The conversations themselves

(Of course, this reflects my biases.)

3 Assumptions and focuses

3.1 Sufficient usefulness

I wrote above that we want to make a *sufficiently useful* and safe AI. What is this *sufficiently useful* criteria? 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 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 these two models as well as different views on takeoff speed, referencing different perspectives as necessary.

3.2 Assumptions

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 similar to current ML. *(TODO: could this be made clearer? Possibly remove some adjectives?)* 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 this 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.
2. Societal and government competence and coordination aren't very high (this informs how hard it is to enforce alignment through governance).
3. AIs capable of directly causing existential catastrophe (with at least small probability) have to be quite intelligent and capable on at least some dimensions. There are other threat models worth considering, but we won't talk about them much here.

I won't make a case for why these are good assumptions here, but would be happy to chat about them 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.

3.3 Capabilities for catastrophe

Note that we'll be focusing on techniques for aligning AIs in the regime where capabilities are sufficient for unaligned AIs to directly cause existential catastrophe (of course, techniques for less capable AIs could generalize). One potential objection to this approach is that in slower takeoff scenarios, the crux of alignment could come down to determining how to get AIs which aren't existentially dangerous to meaningfully speed up alignment research. This would require 'weakly' aligning these 'less' dangerous AIs[2]. Ideally, this would be done without this work on weak alignment contributing to the very likely ongoing research on how to get AIs to do capabilities research. Perhaps this asymmetry could be enacted via the use of secrecy or asymmetrically useful weak alignment strategies. Regardless, we won't be considering this type of scenario here.[3]

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 (it's not important for this post to have intuition about why incoherent decisions are sometimes required). Confused or wondering about the implications of AIs having these principles? Hopefully the next sections will clarify this, so read on.

[2]These 'less' dangerous AIs could still plausibly cause disaster (depending on views surrounding takeoff).

[3]I haven't seen this actively discussed despite a decent number of people putting quite high probability on slow takeoff. It also doesn't seem like current alignment research is asymmetric in this fashion. Perhaps trying to asymmetrically improve alignment research speed during this period via weak alignment work should receive more attention?

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.[4] Agency is mostly just a set of capabilities coupled with (long-term) consequentialism. If we 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 associated with constructing them.

4.1.2 Purely Deontological AI, what is it good for

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.[5] They don't make explicit plans aimed at causing certain outcomes. The fact that purely deontological AIs can be capable of modeling consequences means these AIs can appear agentic. For instance, chess engines like AlphaZero or StockFish are almost certainly purely deontological AIs, but they do appear somewhat agentic. For a more absurd example, consider an AI which only cares about imitating what its actions would be if it were a consequentialist (e.g. a human or some consequentialist AI). For a competent imitator, this is 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 comes down to why the intentions of AIs matter at all.

Intentions determine behavior when out-of-distribution for intelligent and robustly capable agents. By definition, robustly capable agents robustly pursue their goals, so we'll consider the case where an agent will still pursue its goals effectively on an out-of-distribution input.[6] Nearly identical behavior on a specific distribution of tasks from the training environment could be the result of many possible goals (intentions). However, these different goals can be distinguished by different behavior on out-of-distribution inputs. For example, consider some empirical observations of objective robustness failures in which agents 'care' about a feature correlated with the reward and then pursue that feature when

[4]There are some edge cases in which my definition of a purely deontological AI doesn't match typical intuition for tool AIs. These cases will come up below.

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

[6]If an agent is unable to pursue goals effectively, it's unlikely to be able to cause an existential catastrophe, so we won't consider the potential safety issue of capability robustness.

out-of-distribution instead of the reward from the original environment. The behavior out-of-distribution reveals more information about the true goals of the agent (in so far as these weak agents can be seen as having goals). There are also more arcane considerations like deceptive alignment. A deceptively aligned agent would have the instrumental goal of doing well at training and not being detected. Its terminal goals could be anything. So clearly this agent (if capable) would perform well in training, but its true goals would result in different behaviour once it determined that it's no longer in training and there isn't need for deception. How would the agent be able to determine this? By detecting the out-of-distribution property that it's free to defect and seize power. In practice this could work via mechanisms such as detecting out of distribution inputs, modeling the world in general, checking if time has passed relative to when it was training, or simply defecting with small probability at any point[7].

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 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 can do useful things.

Did you catch it?

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

(TODO: better trap example would be nice. This one is too easy and feels silly)

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 another AI which is a competent consequentialist. It seems quite plausible that this imitator would itself just become a consequentialist!

[7] Defecting probabilistically doesn't depend on out-of-distribution inputs, but it does represent a guess from the agent that it is out-of-distribution in a deeper sense.

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 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 superintelligent AI produced via such a process. *(TODO: ecological models?)* We don't have solid approaches for inspecting the decision making of deep agents. Or a decent understanding of what decision making 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.

Is this just the entire alignment problem? Well no, it doesn't include outer alignment and it's possible that we could solve alignment without solving this issue either via the rather dangerous approach discussed in the upcoming section on restrained AI or if it simply happens to be easy to get certain cognitive properties regardless of a lack of understanding. Things could happen to work without us understanding why they work. I claim that relying on this is a very dangerous approach due to difficulties evaluating alignment, for instance, consider deceptiveness. So, my view is that the ways around the hard problem of AI cognition are dangerous (though perhaps I am missing some approaches) and that it is a difficult crux of alignment. I also think that a decent amount of alignment research isn't sufficiently focused on this problem and that we should be more actively working on it. I'll explain my reasoning and what this could look like in my strategic recommendations below.

Now let's transition back to the subject of purely deontological AI. Despite these issues, 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 a 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 there isn't even currently public knowledge of how to produce a superintelligent, highly capable AI (and I wouldn't publish it 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) could produce capable, superintelligent, and 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 more 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[8] and purely deontological model. There's also some speculation in this post on safety in predictive learning.

Elicit Prediction (forecast.elicit.org/binary/questions/8SV58Eq2d) *note, will be formatted in final post*

Elicit Prediction (forecast.elicit.org/binary/questions/Wgff0HgNf) *note, will be formatted in final post*

Now let's suppose that all 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. Pivotal acts likely require careful and superintelligent planning which requires reasoning about consequences. One possible route would be to use process based task AI to radically speed up alignment research. However, research requires a large amount of consequentialist agency which process based task AIs can't do without assistance. So, the use of AI would have to radically speed up alignment research while simultaneously still having humans do the consequentialist component. My best guess is that this bottleneck would result in insufficient research speed improvements particularly given that AI would also likely be used for capability research (depending on the capabilities lead). The analysis of other pivotal acts proceeds similarly.

Elicit Prediction (https://forecast.elicit.org/binary/questions/MhpdWy-A-) *note, will be formatted in final post*

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.

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

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 also instill 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.

The next relevant question is 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 reward for deontological properties as a desirable consequence in its utility function and achieves full marks. For instance, rewarding honesty could be modeled as 'be honest' or as 'appear honest to the overseer'. Note that in some sense this 'consequence model' is the 'correct' model: by definition, outcome based environments reward consequences. So maybe if you set up your deontological property and train to convergence you get a superintelligent, 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 for life on earth intelligence seems to correlate with consequentialism. While even insects can look quite consequentialist from afar, this is really a much less generalizable phenomenon than human consequentialism. This difference in difficulty could also result from the fact that consequentialism is the 'correct' model as discussed earlier. Overall, this issue 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 the deontological properties you want, some alignment failures, and then seemingly getting those properties back again should be 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 cause existential catastrophe. It merely will have some other deontological properties potentially making the AI less useful.[9] However, if the deontological properties of a partially deontological AI splintered or were merely somewhat off, but the consequentialist

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

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 the 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 desired, partially deontological AI could certainly be *sufficiently useful*. They 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.

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[10]. 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 to create myopic agents than to instill complex deontological values. However, even 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. 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 easier 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.[11] 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,

[10] 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.

[11] For details, please see the previously mentioned post on safety issues.

suppose that the time horizon was several years, the AI could seize power in a few days, and 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 increases over time when possessing power due to higher and higher utilisation). So, the time horizon would 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 amount of 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 required 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 affects 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 in general.

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 commercial 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[12]) 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

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

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 ability to perform cognition involving the real 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.

5 Combining approaches

First of all, it's worth noting that restrictions could (slightly) reduce the possibility of catastrophe when testing other approaches in high intelligence regimes. For this to be useful at all, it would require that these restrictions are at least slightly robust to maliciousness. Additionally, this approach is easily fooled by deceptiveness. So, approaches which reduce deceptiveness such as myopia or deontological honesty synergize well with using restrictions for testing. Even AIs which weakly value honesty might reveal malicious intentions if the AI estimates a low probability of successful power seeking and decides not to engage in acausal trade.

It would be possible to combine myopic agency with some deontological properties as an approach for safety. I'm not currently aware of any synergies or anti-synergies; it seems like both approaches would impose some alignment tax and potentially offer different safety properties.

At this time, I don't believe I have any other non-trivial insight into combining the 3 approaches discussed in this post.

6 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. See the section on alignment difficulty for more details. My 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 somewhat long (>40 years) timelines 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 on this topic 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 an existing project which I think is worthwhile and specific examples of additional sub-experiments I think are particularly valuable to conduct.

6.1 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 ways of training models to avoid violating 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 this fanfiction dataset) 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 the 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.

6.2 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 straightforward 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. The environment should reward consequentialism and ideally require 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 be worthwhile (depending on results from the base experiment):

1. Vary the complexity of the environment, training time, or the model size/architecture and determine if results change.
2. Does avoiding sharing parameters between the agent and the value function (for actor-critic methods) 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 deceptively lower than it appears in only this case and passing through the obstacle 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 function is stored in the model as well as build up a human comprehensible understanding of it. Part of this work could be done using techniques like the ones used in the Understanding RL vision paper. Ideally it should be possible to edit the utility function by changing model weights 'by hand'. Can these adjustments be used to change the behavior of the model with respect to the obstacle?

The details of this exact project could probably be considerably refined, but regardless, I think experiments exploring this general idea would be useful.

6.3 How focus should change

It's worth noting that my analysis here is highly speculative. Specifically, imagine everything in this section is prefixed with 'I speculatively think'.

There are a large number of issues or factors which are difficult to model when reasoning strategically about the focuses of the community (at least without spending much more time on analysis). 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 difficulty 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 focus 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 has a chance of being worthwhile. If it actually does prove to be decently helpful, I might spend more time doing a more detailed analysis later.

As stated earlier, the community is spending too little time working on the hard problem of AI cognition. This includes instilling deontological properties, understanding the decision making likely to result from various training methods/environments, and building techniques for transparency and interpretability, particularly of decision making itself. Trying to improve some notion of 'worst case performance' could also be important. It's worth noting that developing architectures which make understanding cognition easier could also be very useful (for instance, architectures which use factored cognition[13]). Overall, focus should shift (at the margin) from making models have behavior X to gaining insight into what sorts of changes in cognition occur when making models do X.

In addition to the Redwood Research project mentioned earlier, here are some examples of endorsed research projects/directions which work on the hard problem of AI cognition (perhaps indirectly):

- TruthfulQA and other work on truthfulness (this would depend on the exact work of course, merely improving on benchmarks doesn't imply

[13]That said, I'm not very hopeful for factored cognition itself; I don't think the approach makes the types of cognition we most care about considerably easier to understand and it likely makes implementing useful things considerably harder.

progress in understanding)
- Visible Thoughts
- Understanding RL Vision

(TODO: more examples?)

I think effort should be reallocated away from 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 enhancing human feedback, active learning, improving sample efficiency, and other related topics. This is based on the understanding that a large amount of research is being conducted in this area and based on the analysis in this earlier section of the post.

I don't think it's worth spending the time going over a bunch more topics in detail (before I know more about the reaction to this post), so I'll just do a quick and dirty more/less at the margin in the style of how Tyler Cowen does overrated/underrated. This list probably doesn't include many important topics. If you want me to add a topic or discuss a topic in more detail, say so in the comments.

Less at the margin:

- Reward modeling/preference learning (discussed above).
- Multi-agent interaction and agent-human interaction. This seems like a capability which will be developed by default in slow takeoff. In fast takeoff scenarios, alignment could require agent-human interaction, but I think it's more effective to figure out this cooperative alignment strategy and then determine exactly what agent-human interaction is needed. This is as opposed to working on agent-human interaction in general. I would also guess that capable consequentialism generalizes to successful interaction with humans (assuming intentions are aligned).
- Agent foundations. My not very confident nor well supported intuition is that proving useful things or gaining valuable understanding with this approach is sufficiently unlikely that intellectual resources should be directed elsewhere.

More at the margin:

- Trying to understand why and how deep learning learns and generalizes. I'm sceptical of mathematically rigorous approaches working, but I'd love to be proven wrong. I'm also somewhat concerned about enhancing capabilities, but I (weakly) think that many types of understanding asymmetrically benefit alignment.
- Interpretability, particularly focused on decision making. This is probably only slightly underrated, but I don't see much work in this space.

About the right amount:

- Trying to predict potential issues with various approaches to alignment.

- Forecasting more generally.
- Factored cognition.
- Truthfulness/honesty.

(TODO: more ratings?)

7 Alignment difficulty

After conditioning on timelines, takeoff, and AI emerging from deep learning, I think that many of the key cruxes of alignment difficulty are related to the hard problem of AI cognition. Specifically, will we actually develop a good understanding of AI cognition? Even if understanding is poor, how difficult is instilling desired deontological properties and inner alignment more generally? How will this difficulty depend on the intelligence of agents?

Over the course of writing this post, I have noticed my views on the difficulty of alignment have shifted to be closer to my model of the views of Eliezer Yudkowsky. Specifically, my views have shifted considerably toward thinking that alignment is more difficult due to high levels of difficulty in instilling deontological properties, particularly in more intelligent agents. Perhaps this is an example of a general phenomenon: first principles contemplation of consequentialism, agency, and intelligence leads to Yudkowskization (similar to carcinization).