Thanks to Wil Perkins, Grant Fleming, Thomas Larsen, Declan Nishiyama, and Frank McBride for feedback on this post. Thanks also to Paul Christiano, Daniel Kokotajlo, and Aaron Scher for comments on the original post that helped clarify the argument. Any mistakes are my own.

In order to submit this to the Open Philanthropy AI Worldview Contest, I'm combining this with the previous post in the sequence and making significant updates. I'm leaving the previous post, because there is important discussion in the comments, and a few things that I ended up leaving out of the final version that may be valuable.

Introduction

In this post, I argue that deceptive alignment is less than 1% likely to emerge for transformative AI (TAI) by default. Deceptive alignment is the concept of a proxy-aligned model becoming situationally aware and acting cooperatively in training so it can escape oversight later and defect to pursue its proxy goals. There are other ways an AI agent could become manipulative, possibly due to biases in oversight and training data. Such models could become dangerous by optimizing directly for reward and exploiting hacks for increasing reward that are not in line with human values, or something similar. To avoid confusion, I will refer to these alternative manipulative models as **direct reward optimizers**. Direct reward optimizers are outside of the scope of this post.

Summary

In this post, I discuss four precursors of deceptive alignment, which I will refer to in this post as **foundational properties**. I first argue that two of these are unlikely to appear during pre-training. I then argue that the order in which these foundational properties develop is crucial for estimating the likelihood that deceptive alignment will emerge for prosaic transformative AI (TAI) in fine-tuning, and that the dangerous orderings are unlikely. In particular:

1. Long-term goals and situational awareness are very unlikely in pre-training.

2. Deceptive alignment is very unlikely if the model understands the base goal before it becomes significantly goal directed.
3. Deceptive alignment is very unlikely if the model understands the base goal significantly before it develops long-term, cross-episode goals.

Pre-training and prompt engineering should enable an understanding of the base goal without developing long-term goals or situational awareness. On the other hand, long-term goals and will be much harder to train.

Definition

In this post, I use the term **“differential adversarial examples”** to refer to adversarial examples in which a non-deceptive model will perform differently depending on whether it is aligned or proxy aligned. The deceptive alignment story assumes that differential adversarial examples exist. The model knows it’s being trained to do something out of line with its goals during training and plays along temporarily so it can defect later. That implies that differential adversarial examples exist in training.

Implications of this argument

Many existential AI catastrophe scenarios rely heavily on deceptive alignment. So, if deceptive alignment is highly unlikely (<1%) to emerge for TAI, we should update our expectations for AI existential risk accordingly. This conclusion also warrants a re-evaluation of priorities for alignment research areas and between cause areas. Other possible alignment research topics include governance, direct reward optimizers, and multipolar scenarios that don’t rely on deceptive alignment.

Assumptions about the TAI training paradigm

I’m assuming prosaic development of TAI, using a training process like human feedback on diverse tasks (HFDT). The goal of the training process would be a model that follows directions subject to non-consequentialist ethical considerations. This high-level training setup is already the default training process for text models such as GPT-4, and this will likely continue because of the flexibility and strong performance it provides. I also expect unsupervised pre-training to be an important part of TAI development. You can find justifications for these assumptions here.

Foundational properties for deceptive alignment

There are four foundational properties that are key to the deceptive alignment story. These foundational properties are:

1. Goal-directed behavior - the model needs to have some goal or set of goals about the external world.
2. Optimizing across episodes/long-term goal horizons - model behavior needs to be influenced by long-term goals.
3. Conceptualization of the base goal - the model needs to understand what the base goal is in order to successfully play along in training.
4. Situational awareness - the model needs to understand that it is a model in training. Furthermore, it needs to understand how its actions now could affect its parameter updates, and therefore future goal attainment.

The order in which these develop, and the amount of training that goes on in between the development of foundational properties, will have a large effect on the likelihood of deceptive alignment.

Pre-trained models are unlikely to develop the necessary properties

Pre-trained models are unlikely to have long-term goals

In a typical pre-training process, a model's predicted output for token n is not used in the calculation or scoring of token n+1, and thus loss is calculated independently. There's not an incentive for gradient descent to learn sacrifice performance on the current word for performance on a future word in pre-training.

Furthermore, if a model gained marginally longer-term goals, but not long-term enough to enable deceptive alignment, the longer-term goals would be a competing priority and harm immediate reward in expectation. Gradient descent should therefore push against this.

The deceptive alignment literature argues that gradient descent could modify a model's goals to be long-term, *because* that would make it more deceptive and therefore increase training reward. There are therefore some concerns about long-term goals and deceptive alignment in pre-training. However, this is unlikely. For the sake of argument, suppose we have a model in pre-training that has a misaligned proxy goal and relevant situational awareness. But so far, it does not have a long-term goal.

In order to be deceptively aligned, the model has to have a long enough goal horizon to value its total goal achievement after escaping oversight more than its total goal achievement before escaping oversight. But pre-training processes are incredibly long (~the whole internet for text models). A situationally aware model should anticipate a lot of subsequent pre-training, and that the oversight will likely remain strong for a significant portion of fine-tuning. That requires a very long-term goal for any amount of deceptive alignment to be worthwhile.

Gradient descent updates parameters in the direction that most improves performance hyper-locally. Under the conditions I've described, making a myopic model's goals infinitesimally longer-term would not make deceptive alignment worthwhile and would harm training reward. Therefore, I don't expect the gradients to point toward longer-term goals.

Pre-trained models unlikely to develop situational awareness

It's hard to imagine how a model would develop situational awareness in pre-training when:

1. Situational awareness only produces deceptive alignment if the model already has long-term goals, and vice versa. Gradient descent is based on partial derivatives, which are independent of each other, so:
    a. If the model doesn't already have long enough goal horizons for deceptive alignment, then marginally more situational awareness doesn't increase deceptive alignment.
    b. If the model doesn't already have the kind of situational awareness necessary for deceptive alignment, then a marginally longer-term goal doesn't increase deceptive alignment.
    c. Therefore, the partial derivatives shouldn't point toward either property unless the model already has the other.
2. Unlike in fine-tuning, the vast majority of internet text prompts do not contain information about situational context for the model to figure out that it is an ML model. Understanding that it is a model in training is not relevant for predicting next tokens on the internet, so there's no benefit to learning this information.

Dangerous foundational property development order is unlikely in fine-tuning

How complete does the understanding of the base goal need to be?

My central argument for fine-tuning is about when an understanding of the base goal emerges relative to other properties. But how accurate and specific does that understanding have to be to avoid deceptive alignment?

The relevant factor is whether the model’s internal goal is its own understanding of the base goal, even if that understanding is flawed. Suppose for the sake of argument that we have a model with an exceptionally long-term goal and situational awareness. Its internal goal is its best, but still flawed, representation (X) of the base goal (X`). This model can't tell the difference between the base goal and the internal proxy goal, so it can't be deceptively aligned yet. In short, it thinks it is aligned. If it performs worse than it could have been on a training example because of this, the gradients should point toward a stronger concept of the base goal, and therefore stronger alignment.

For example, let's say we are training this model to follow directions from the prompt, unless they violate ethical norms (X`). The model understands that the base goal is to follow directions subject to some ethical norms but has an understanding of these norms (X) that is incorrect in some situations. Because it doesn't understand the difference, it can't act aligned for instrumental purposes, and it will get suboptimal reward when flaws in its understanding of the base goal affect its behavior. Because it is already pointing its goal at this flawed concept, the gradients should point toward improving the internal goal (X) to better match the base goal (X`). In order to become deceptively aligned, it has to have separate concepts for its internal goal and the base goal.

Now consider an alternate version of the above example, where the model has the same flawed concept for the base goal, but its internal goal is maximizing paperclips (or anything else that's represented by a separate concept from its base goal representation). This model can immediately become deceptively aligned, because optimizing for its flawed understanding of the base goal will get better training performance than optimizing for paperclips in the short run. The model can then continue learning about the base goal without systematically changing its internal goals and remain deceptively aligned.

Will understanding of the base goal come before goal-directed behavior?

Why does this matter?

A key idea of deceptive alignment is early in training, a model does not understand the base goal, so it starts by learning a proxy that becomes its internal goal. But if the model understands the base goal before it develops a mesa-objective, then assuming differential adversarial training, there will be an incentive to optimize for the base goal, rather than a proxy. If the understanding of the base goal has already been calculated, there will be no reason to create and maintain a separate proxy goal when the model could just point at the base goal.

If conceptualization of the base goal and goal-directed behavior emerge at essentially the same time, deceptive alignment also seems unlikely. If a model converged on a solid representation of the base goal, and became goal-directed in parallel, then differential adversarial examples would provide an incentive for gradient descent to optimize for the base goal.

The most dangerous case is when the model develops a proxy objective before it understands the base goal.

I’ve seen some arguments that rely on the idea that deceptive alignment will not occur in the parameters, but the activations. This argument doesn't change the conclusion from this section. If the model figured out what you wanted it to do in the activations before it became significantly goal-directed, gradients could point towards parameters that accomplish those objectives after they have been inferred.

How likely is the dangerous case?

It is highly likely that unsupervised pre-training will be the first step in the TAI training process. Models that are only pre-trained almost certainly don’t have dangerous goals. They might have extremely myopic goals like next-token prediction, but those goals will necessarily start to shift when the model moves to fine-tuning. However, because a pre-trained model will already have high-level representations of key base goal concepts such as direction following and non-consequentialist ethical considerations, all it will have to do to become aligned is to point them. These models also understand language, and we can tell them what we want them to do from the start of fine-tuning. This kind of prompt engineering is already standard practice, and it's hard to imagine why we wouldn't use it going forward.

Differential adversarial examples should provide a strong incentive for gradient descent to point at the base goal, rather than a proxy. The deceptive alignment story assumes that differential adversarial examples exist. The model is explicitly changing its behavior to avoid correction, meaning that its internal goals produce different results from the base goal on training examples. If they did not exist, that would imply that a perfect or near-perfect proxy exists on a diverse training set. It's hard to imagine what that proxy could be if the goal is to follow directions subject to some constraints in diverse situations.

The argument presented in this section only applies to domains where large amounts of pre-training data exist. For example, AI operating robotics may not work this way. However, if researchers were able to produce a robustly aligned model for domains where pre-training is available, such as text and video generation, that would likely make it a lot easier to align robotics. The aligned, non-embodied AI could help with alignment research. Also, general purpose systems that include robotics, such as a scaled-up version of Gato, will be easier to align if we can robustly align key component parts such as text modeling. Alignment of the language portion of a general agent would also enable prompt engineering to help with overall alignment.

Will understanding of the base goal come before long-term goal horizons?

Why does this matter?

If a model has long-term goal horizons, then it must be goal-directed. However, it’s possible for the model to develop foundational properties in the following order:

1. Goal-directed behavior
2. Conceptualization of the base goal
3. Optimizing across episodes/long-term goal horizons

If this happens, there will be a period of time when the model wants to optimize for a flawed proxy and understands the base goal but does not yet have the capabilities necessary for deceptive alignment. Because the model is not yet deceptively aligned, differential adversarial examples in this part of training should directly modify the model’s internal objective. If this period lasts for a significant number of differential adversarial examples, the model should be able to update its goals to point at its conception of the base objective.

The most dangerous case, which underpins the arguments described in the deceptive alignment literature, is if long-term, cross-episode goals develop before or around the same time as an understanding of the base goal.

The argument presented in this section still holds if the model is gaining these properties from the activations. The moment at which the model can figure out what you're trying to get it to do from the activations and the moment it can be prompted to have long-term goals may not happen at the same time.

How likely is the dangerous case?

A model can't become deceptively aligned unless it has long-term goals that extend beyond the current training episode. It's hard to see how gradient descent could train this cross-episode goals, especially if cross-episode goals are not a specific training objective. Gradient descent does not have a long-term plan for how it will achieve high reward. All it does is update parameters in the direction that would most improve its performance on the current training batch, **hyper-locally**. It can't systematically optimize for anything else.

Therefore, it's hard to imagine why gradient descent would incentivize sacrificing performance in the current training batch to achieve future gains in episodes that are not included in the training calculations. If you rewarded a model for things that happened in a different episode, then that would actually just expand the current episode, so it likely wouldn't generalize further than that. There may be ways to train cross-episode optimization, but it's hard to imagine it happening early in training by accident. Similarly, it is probably difficult to train long goal horizons within an episode without explicitly rewarding a model for activities that occurred long before the reward.

Therefore, to deliberately train a model to have very long-term goals, you would likely need to train it over at least the duration of the intended goal. Deceptive alignment would likely require very long-term goals if it were early enough in training that the model did not have a concept of the base goal. If the duration of the goal were years or decades, then training would likely require years or decades of training data, even with very sophisticated training techniques. Even then, predicting the effects of your actions over long time frames in complex environments is hard, even for highly intelligent entities. Credit assignment would be quite difficult. This should provide plenty of time for the model to point at the right goal before it can become deceptively aligned.

I expect ML researchers to deliberately train models to have long-term goals. However, I expect researchers to start by reinforcing a pre-trained model on shorter-term goals. This approach would almost certainly be a lot easier and create a good foundation to build on. In contrast, understanding the base goal will be an important training priority immediately.

Some have proposed that ML models could learn to have long-term goals because that would enable deceptive alignment and therefore high training performance, but this argument falls flat in fine-tuning for the same reasons as in pre-training.

Skepticism of empirical examples of deceptive alignment

There are a lot of people in the alignment community who have studied and built a wide range of deep learning models in detail. But the only cases of internally misaligned machine learning models I've seen are toy cases that deliberately create this effect by creating a perfect proxy for the objective in the training set. In contrast, the deceptive alignment story assumes that the model's proxies are not perfect proxies in training. Otherwise, what would the instrumental incentive for alignment be? I don't think these toy examples say anything about whether large, complex models will become deceptively aligned by accident. It's not surprising that you can intentionally create internal misalignment.

In contrast, I suspect that this community has put a lot less effort into simulating evolution. But the only non-contrived example of internal misalignment happening by accident that I'm aware of is from a simulation of evolution!

Gene propagation was the “reward” in this example. The goal of filtering high-reproduction organisms in training was to control for reproduction rate and isolate for other factors for gene survival. The analogy for a misaligned goal is reproducing at a high rate anyway. However, in this case, the simulation moves back and forth between a training process that discourages high reproduction rates and a deployment environment that encourages high reproduction rates. There is optimization in both settings, but **in opposite directions**. Therefore, it’s not surprising that evolution would find a way to act deceptively. Avoiding optimization in different phases of training versus deployment is a key strategy for making sure deceptive alignment doesn’t happen. Let’s look at some possible high-level training/testing regimes for TAI:

1. One long training process, followed by deployment in which there are no parameter updates - here, there is no optimization power from the testing process, so I would not expect deceptive alignment.
2. Iteratively move between training, in which updates are made, and deployment, in which updates are not made - here, you have the property where you move back and forth between training and deployment, but you do not have any optimization power in the deployment phase. Crucially, you could start with a fresh prompt in training and the model would have no way to recall information from deployment.
3. A training process in which updates are made, and deployment, in which parameter updates are made based on things that happen in the wild. In this case, you do have optimization power in both training and deployment, but the optimization power would be in the same direction (towards what you want). Therefore, I would not expect the type of dangerous optimization seen in this evolution experiment. Even if you continued iterating between training and this mixed version of deployment, you would be training it in roughly the same way on both. In practice, the training/deployment distinction might get blurry, but that should be fine.

None of these produce incentives that point in different directions for training versus deployment.

Implications for overall AI x-risk

Deceptive alignment is central to many descriptions of how transformative AI could cause an existential risk. If it is unlikely, then we should update our estimates of risk accordingly. Other AI failure modes include direct reward optimizers and some multipolar scenarios that don't rely on deceptive alignment. If deceptive alignment is very unlikely for TAI, then research on alternative governance and misalignment scenarios should take precedence over deceptive alignment. It's also worth re-evaluating how high of a priority AI risk should be. This would represent a serious shift from the status quo. As a side benefit, deceptive alignment is also the main line of argument that sounds like science fiction to people outside of the alignment community. Shifting away from it should make it easier to communicate with people outside of the community.

Conclusion

The standard deceptive alignment argument relies on foundational properties developing in a very specific order. However, this ordering is unlikely for prosaic AI. Long-term goals and situational awareness are not realistic outcomes in pre-training. In fine-tuning, it's very unlikely that a model would develop a misaligned long-term goal before its goal became aligned with the training goal. Based on this analysis, deceptive alignment is less than 1% likely for prosaic TAI. This renders many of the doom scenarios that are discussed in the alignment community unlikely. If the arguments in this post hold up to scrutiny, we should redirect effort to governance, multipolar risk, direct reward optimizers, and other cause areas.

Appendix

Justification for key assumptions

I have made 3 key assumptions in this post:

1. TAI will come from prosaic AI training.
2. TAI will involve substantial unsupervised pre-training.
3. TAI will come at least in part from human feedback on diverse tasks.

I justify them in this section.

TAI will come from prosaic AI

It's possible that there will be a massive paradigm shift away from machine learning, and that would negate most of the arguments from this post. However, I think that this shift is very unlikely. Historically, attempts to create powerful AI without machine learning have been very disappointing. Given the success of ML and the amount of complexity that seems necessary even for narrow intelligence, it would be quite surprising for TAI to emerge without machine learning. Even if it did, the order of foundational properties development would still matter, as described in my previous post.

The arguments in this post don't rely on any particular machine learning architecture, so the conclusions should be robust to different architectures. It's possible that gradient descent will be replaced by something that doesn't rely on gradients and local optimization, which would undermine some of these arguments. This possibility also doesn't seem likely to me, given the difficulty of optimizing trillions of parameters without taking small, local steps. As far as I can tell, the alignment community largely shares this belief.

TAI will involve substantial unsupervised pre-training

Pre-training already enables our AI to model human language effectively. It leverages massive amounts of data and works very well. It would be surprising for someone to try to develop TAI without using this resource. General-purpose systems could easily incorporate this, and it would take something extreme to make that obsolete. Human language is complicated, and it's hard to imagine modeling that from scratch without a large amount of data.

TAI will come at least in part from human feedback on diverse tasks

This post assumes that the goal of training is a general, direction following agent using human feedback on diverse tasks. However, the most likely alternative training regimes don't change the conclusions. For example, if TAI instead came from training a model to automate scientific research, the model would presumably still include a significant pre-trained language component. Furthermore, scientific research involves a lot of thorny ethical questions. There also needs to be a way to tell it what to do, and direction following is a straightforward solution for that. Therefore, there is a strong incentive to train non-consequentialist ethical considerations and direction following as the core functions of the model, even though its main purpose is scientific research. This approach provides a lot of flexibility and will likely be used by default.

There are also some possible augmentations to the human feedback process. For example, Constitutional AI uses reinforcement learning from human feedback (RLHF) to train a helpful model, then uses AI feedback and a set of principles to train harmlessness. This kind of implementation detail shouldn't significantly affect foundational property development order, and therefore would not change my conclusion.