
NE
E
SE
S
SW
W
NW
N

An Inside View of AI Alignment

An Inside View of AI Alignment

I started to take AI Alignment seriously around early 2020. I'd been interested in AI and machine learning in particular since 2014 or so, taking several online ML courses in high school and implementing some simple models for various projects. I leaned into the same niche in college, taking classes in NLP, Computer Vision, and Deep Learning to learn more of the underlying theory and modern applications of AI, with a continued emphasis on ML. I was very optimistic about AI capabilities then (and still am) and if you'd asked me about AI alignment or safety as late as my sophomore year of college (2018-2019), I probably would have quoted Steven Pinker or Andrew Ng at you.

Somewhere in the process of reading The Sequences, portions of the AI Foom Debate, and texts like *Superintelligence* and *Human Compatible*, I changed my mind. Some 80,000 hours podcast episodes were no doubt influential as well, particularly the episodes with Paul Christiano. By late 2020, I probably took AI risk as seriously as I do today, believing it to be one of the world's most pressing problems (perhaps the most) and was interested in learning more about it. I binged most of the sequences on the Alignment Forum at this point, learning about proposals and concepts like IDA, Debate, Recursive Reward Modeling, Embedded Agency, Attainable Utility Preservation, CIRL etc. Throughout 2021 I continued to keep a finger on the pulse of the field: I got a large amount of value out of the Late 2021 MIRI Conversations in particular, shifting away from a substantial amount of optimism in prosaic alignment methods, slower takeoff speeds, longer timelines, and a generally "Christiano-ish" view of the field and more towards a "Yudkowsky-ish" position.

I had a vague sense that AI safety would eventually be the problem I wanted to work on in my life, but going through the EA Cambridge AGI Safety Fundamentals Course helped make it clear that I could productively contribute to AI safety work right now or in the near future. This sequence is going to be an attempt to explicate my current model or "inside view" of the field. These viewpoints have been developed over several years and are no doubt influenced by my path into and through AI safety research: for example, I tend to take aligning modern ML models extremely seriously, perhaps more seriously than is deserved, because of my greater amount of experience with ML compared to other AI paradigms.

I'm writing with the express goal of having my beliefs critiqued and scrutinized: there's a lot I don't know and no doubt a large amount that I'm misunderstanding. I plan on writing on a wide variety of topics: the views of various researchers, my understanding and confidence in specific alignment proposals, timelines, takeoff speeds, the scaling hypothesis, interpretability, etc. I also don't have a fixed timeline or planned order in which I plan to publish different pieces of the model.

Without further ado, the posts that follow comprise Ansh's (current) Inside View of AI Alignment.

RLHF

I've been thinking about Reinforcement Learning from Human Feedback (RLHF) a lot lately, mostly as a result of my AGISF capstone project attempting to use it to teach a language model to write better responses to Reddit writing prompts, a la Learning to summarize from human feedback.

RLHF has generated some impressive outputs lately, but there seems to be a significant amount of disagreement regarding its potential as a partial or complete solution to alignment: some are excited to extend the promising results we have so far, while others are more pessimistic and perhaps even opposed to further work along these lines. I find myself optimistic about the usefulness of RLHF work, but far from confident that all of the method's shortcomings can be overcome.

How it Works

At a high level, RLHF learns a reward model for a certain task based on human feedback and then trains a policy to optimize the reward received from the reward model. In practice, the reward model learned is likely overfit - the policy can thus benefit from interpolating between a policy that optimizes the reward model's reward and a policy trained through pure imitation learning.

A key advantage of RLHF is the ease of gathering feedback and the sample efficiency required to train the reward model. For many tasks, it's significantly easier to provide feedback on a model's performance rather than attempting to teach the model through imitation. We can also conceive of tasks where humans remain incapable of completing the tasks themselves, but can evaluate various completions and provide feedback on them. This feedback can be as simple as picking the better of two sample completions, but it's plausible that other forms of feedback might be more appropriate and/or more effective than this. The ultimate goal is to get a reward model that represents human preferences for how a task should be done: this is also known as Inverse Reinforcement Learning. The creators of the method, Andrew Ng and Stuart Russell, believe that "the reward function, rather than the policy, is the most succinct, robust, and transferable definition of the task,". Think about training an AI to drive a car: we might not want it to learn to imitate human drivers, but rather learn what humans value in driving behavior in the abstract and then optimize against those preferences.

Outer Alignment Concerns

If a reward model trained through human feedback properly encoded human preferences, we might expect RLHF to be a plausible path to Outer Alignment. But this seems like a tall order, considering that humans can be assigned any values whatsoever, the easy goal inference problem is still hard, and that it's easy to misspecify any model that attempts to correct for human biases or irrationality. Ambitious value learning is hard, and I'm not particularly confident that RLHF makes it significantly more tractable.

It's also plausible that this approach of inferring a reward function for a task is just fundamentally misguided and that the way to get an outer aligned system is through the assistance-game or CIRL framework instead. There are definite advantages of this paradigm over the more standard reward learning setup that RLHF leverages. By treating humans as pieces of the environment and the reward function

as a latent variable in the environment, an AI system can merge the reward learning and policy training functions that RLHF separates and thereby “take into account the reward learning process when selecting actions,”. This makes it easier to make plans conditional on future feedback, only gather feedback as and when it becomes necessary, and more fluidly learn from different forms of feedback.

Scalable oversight is hard

RLHF also relies upon humans being able to evaluate the outputs of models. This will likely be impossible for the kinds of tasks we want to scale AI to perform - it’s just going to be too hard for a human to understand why one output should be preferred over another. We’d simply have to hope that reward model generalization we’d seen previously, when oversight was still possible, continued to hold. Even if we thought we’d figured out how to evaluate our models’ outputs, there’s always the chance of an inner alignment failure or other deceptive behavior evading our oversight - we’d want to be absolutely certain that our reward and policy models were actually doing what we wanted them to do.

The solutions to the scalable oversight problem seem to primarily rely on AI-assistance and/or breakthroughs in interpretability techniques. I think it’s clear how the latter might be useful: if we could just look at any model and be certain of its optimization objective, we’d probably feel pretty comfortable understanding the reward models and policy models we trained. AI-assistance might look something like recursive reward modeling: break the task that’s too hard to oversee into more manageable chunks that a human can oversee and train a model to optimize those tasks. Using the models trained on the narrower subtasks might make the original task possible to oversee: this is an idea that has been used for the task of summarizing books. It’s plausible that there are many tasks that resist this kind of decomposition, but the factored cognition approach might get us very far indeed.

Why I think RLHF is valuable

I’ll quote Paul Christiano here:

> We are moving rapidly from a world where people deploy manifestly unaligned models (where even talking about alignment barely makes sense) to people deploying models which are misaligned because (i) humans make mistakes in evaluation, (ii) there are high-stakes decisions so we can't rely on average-case performance.
>
> This seems like a good thing to do if you want to move on to research addressing the problems in RLHF: (i) improving the quality of the evaluations (e.g. by using AI assistance), and (ii) handling high-stakes objective misgeneralization (e.g. by adversarial training).
>
> In addition to "doing the basic thing before the more complicated thing intended to address its failures," it's also the case that RLHF is a building block in the more complicated things.
>
> I think that (a) there is a good chance that these boring approaches will work well enough to buy (a significant amount) time for humans or superhuman AIs to make progress on alignment research or coordination, (b) when they fail, there is a good chance that their failures can be productively studied and addressed.

I generally agree with this. Solving problems that crop up in RLHF seems likely to transfer to other alignment methods, or at least be productive mistakes. The interpretability techniques we develop, outer or inner alignment failures we find, and latent knowledge we elicit from our reward and policy models all seem broadly applicable to future AI paradigms. In other words, I think the textbook from the future on AI Alignment is likely to speak positively of RLHF, at the very least as an early alignment approach.

Promising RLHF Research Directions (according to me)

I'd like to see different kinds of feedback be used in addition to preference orderings over model outputs. This paper specifies a formalism for the reward learning in general and considers several different kinds of feedback that might be appropriate for different tasks, e.g. demonstration, correction, natural language feedback, etc. A reward model that can gracefully learn from a wide array of feedback types seems like a desirable goal. This kind of exploration might also help us figure out better and worse forms of feedback and what kinds of generalization arise from each type.

Relatedly, I think it might be interesting to see how the assistance game paradigm performs in settings where the RLHF paradigm has been applied, like text summarization. On a theoretical level it seems clear that the assistance game setup offers some unique benefits and it would be cool to see those realized.

As we continue to scale RLHF work up, I want to see how we begin to decompose tasks so that we can apply methods like Recursive Reward Modeling. For book summarization, OpenAI used a fixed chunking algorithm to break the text down into manageable pieces, but it seems likely that other kinds of decomposition won't be as trivial. We might need AI assistance to decompose tasks that we can't oversee into tasks that we can. Training decomposition models that can look at a task and identify overseeable subtasks seems like a shovel-ready problem, perhaps one that we might even apply RLHF to.

The Bio Anchors Forecast

Ajeya Cotra's Forecasting Transformative AI with Biological Anchors, to my knowledge, represents the most serious effort to predict the arrival of transformative AI - even if it's not attempting to pinpoint the exact instant that we'll get transformative AI, it posits an upper bound on the amount of time until then (Holden Karnofsky's post clarifies the difference).

The report's methodology, summarized by Scott Alexander, is:

1. Figure out how much inferential computation the human brain does.

2. Try to figure out how much training computation it would take, right now, to get a neural net that does the same amount of inferential computation. Get some mind-bogglingly large number.

3. Adjust for "algorithmic progress", ie maybe in the future neural nets will be better at using computational resources efficiently. Get some number which, realistically, is still mind-bogglingly large.

4. Probably if you wanted that mind-bogglingly large amount of computation, it would take some mind-bogglingly large amount of money. But computation is getting cheaper every year. Also, the economy is growing every year. Also, the share of the economy that goes to investments in AI companies is growing every year. So at some point, some AI company will actually be able to afford that mind-boggingly-large amount of money, deploy the mind-bogglingly large amount of computation, and train the AI that has the same inferential computation as the human brain.

5. Figure out what year that is.

There are some other biological anchors that the report also considers, besides the human brain compute estimate, such as drawing a comparison between the parameter count of a transformative AI model and the parameter count of the human genome, or the amount of compute used by all animal brains over the course of evolution - I don't place a ton of weight on these anchors myself, since I find them uninformative, and I tend to think the strength of this report lies in assuming that the current deep learning paradigm will lead to transformative AI and trying to work backward from there.

Is it actually reasonable to assume that deep learning might produce transformative AI? I think so. Mark Xu's model here is compelling regarding the differences between different AI algorithms/paradigms in terms of their reliance on computing power: each algorithm seems to have a certain "effective compute" regime within which scaling up computing power leads to predictable increases in capabilities and beyond which capability gains stall out or become vanishingly small. It seems increasingly likely to me that we're far from exiting the effective computing regime of modern neural networks - I think the results from e.g. the Chinchilla paper support the fact that we have more juice to squeeze from models very similar to the ones we've already trained (PaLM, Gato, DALL-E, etc.). Furthermore, these models already seem to be on the cusp of having real societal impact. Copilot doesn't seem to be terribly far from something that would become a part of every professional software engineer's toolkit,

and it feels like DALL-E 2 could be scaled up into something that produces bespoke illustrations for various needs on demand.

Given that, I think the most compelling objection to the utility of the Biological Anchors forecast is the claim that the development of human intelligence by evolution gives us no information about the development of AI. I take this to be the thrust of Eliezer's critique, aptly summarized by Adam Shimi:

> My interpretation is that he is saying that Evolution (as the generator of most biological anchors) explores the solution space in a fundamentally different path than human research. So what you have is two paths through a space. The burden of proof for biological anchors thus lies in arguing that there are enough connections/correlations between the two paths to use one in order to predict the other.
>
> …
>
> In his piece, Yudkowsky is giving arguments that the human research path should lead to more efficient AGIs than evolution, in part due to the ability of humans to have and leverage insights, which the naive optimization process of evolution can't do. He also points to the inefficiency of biology in implementing new (in geological-time) complex solutions. On the other hand, he doesn't seem to see a way of linking the amount of resources needed by evolution to the amount of resources needed by human research, because they are so different.
>
> If the two paths are very different and don't even aim at the same parts of the search space, there's nothing telling you that computing the optimization power of the first path helps in understanding the second one.

I think Yudkowsky would agree that if you could estimate the amount of resources needed to simulate all evolution until humans at the level of details that you know is enough to capture all relevant aspects, that amount of resources would be an upper bound on the time taken by human research because that's a way to get AGI if you have the resources. But the number is so vastly large (and actually unknown due to the "level of details" problem) that it's not really relevant for timelines calculations.

I find this argument partially convincing, but not entirely so. I don't agree that "the two paths…don't even aim at the same part of the search space," since it seems to me like we'll be optimizing AI for criteria developed by our understanding of human intelligence. If we're aiming for human-level, and eventually superhuman, AI capabilities, it seems likely to me that we're trying to optimize for desiderata not completely uncorrelated with those of evolution. A simple example is with language models - if the goal of GPT-N is to respond to any text prompt like (super)human would, we're obviously judging it by its ability to meet (super)human language benchmarks.

That said, I think it's reasonable to not update strongly off of this report, especially if recent (at the time) AI progress had already made you update to shorter or longer AI timelines. Personally, since I find a Deep Learning Based Development Model (and consequently a Deep Learning Based Threat Model) to be my modal prediction for future AI progress, this report helps provide grounding for my personally short-ish timelines (how short they are depends on who you're talking to!). I plan to address my personal timelines in an upcoming post of this sequence.