


Research Journals

1. Research Principles for 6 Months of AI Alignment Studies
2. Three Alignment Schemas & Their Problems
3. Loose Threads on Intelligence
4. Optimizing Human Collective Intelligence to Align AI
5. Reflections on Deception & Generality in Scalable Oversight (Another OpenAI Alignment Review)
6. A Simple Alignment Typology

Research Principles for 6 Months of AI Alignment Studies

This summer I learned about the concept of *Audience Capture* from the case of Nicholas Perry. Through pure force of social validation, he experienced a shift from an idealistic but obscure young man to a grotesque but popular caricature of a medical train wreck.

The change happened through social reward signals. Originally Nicholas the principled vegan made videos of himself playing the violin, much to no one's interest. The earnest young man then learned he had to give up his vegan diet for health reasons, and thought to give the occasion a positive twist by inviting his viewers to share the first meal of his new lifestyle.

It was an innocuous step. He gained viewers. They cheered him on to eat more. And he did.

Gradually, but steadily he ate and ate, to the cheers of a swelling crowd of online followers. And like the Ghandi Murder Pill, the choice of sacrificing a sliver of his values for substantial reward was worth it for each individual video he made. His popularity expanded with his waistline as he inched up the social incentive slope. And at the end of that slope Nicholas didn't care about health, veganism, or playing the violin anymore. Instead his brain was inured with social reward signals that had rewired his values on a fundamental level. Essentially, Nicholas had a become a different person.

Now I realize I am unlikely to gain 300 pounds from success on AI alignment articles on LessWrong, but *audience capture* does point to a worry that has been on my mind. How do new researchers in the field keep themselves from following social incentive gradients? Especially considering how hard it is to notice such gradients in the first place!

Luckily the author of the above article suggests a method to ward against *audience capture*: Define your ideal self up front and commit to aligning your future behavior with her. So this is what I am doing here -- I want to precommit to three *Research Principles* for the next 6 months of my AI alignment studies:

1. **Transparency** - I commit to exposing my work in progress, unflattering confusions of thinking, and potentially controversial epistemics.
2. **Exploration** - I commit to exploring new paths for researching the alignment problem and documenting my progress along the way.
3. **Paradigmicity** - I commit to working toward a coherent paradigm of AI alignment in which I can situate my work, explain how it contributes to solving alignment, and measure my progress toward this goal.

Let's take a closer look at each research principle.

Transparency: Avoid Distortion

First things first.

I've received a research grant to study AI alignment and I don't know if I'm the right person for it.

This admission is not lack of confidence or motivation. I feel highly driven to work on the problem, and I know what skills I have that qualify me for the job. However, AI alignment is a new field and it's unclear what properties breakthrough researchers have. So naturally, I can't assess if I have these unknown properties until I actually make progress on the problem.

Still the admission feels highly uncomfortable -- like I'm breaking a rule. It feels like the type of thing where common wisdom would tell me to power pose myself out of this frame of mind. I think that wisdom is wrong. What I want is for alignment to get solved, which means I want the right people working on it.

Is the "right people" me?

I don't know. *But I also think most people can't know*. I think self-assessment is a trap due to motivated reasoning and other biases I'm still learning about. Instead, I believe it's better to commit to transcribing one's work. It can speak for itself. Thus I won't hype myself or sell myself or only show the smartest things I came up with. In short, I want to commit to a form of transparency based on epistemic humility.

This approach will obviously lead to more clutter compared to filtering one's output on quality. Still, I'd argue the trade-off is worth it because it allows evaluation of on-going work instead of an implicit competition of self-assessment smothered in social skills and perception management.

Thus I commit to exposing my work in progress, unflattering confusions of thinking, and potentially controversial epistemics.

Exploration: Mark the Path

Restrospectives don't capture the actual experience of going through a research process. Our memories are selective and our narratives are biased. Journaling our progress as we go avoids these failings but such journals will be rife with dead ends and deprived of hindsight wisdom. On the other hand, if the useful information density is too low, people can simply opt out of reading them.

Win-win.

So what outcomes should we expect? I think there are four possible research results of the next few months: A huge breakthrough no one had considered, a useful research direction that more people are already working on, a useless path no one has explored before, and a useless approach that was predictably useless. Thus we have a 2x2 grid of outcomes across the Useful-Useless axis and the Known-Unknown axis:

	Useful	**Useless**
Known	Converge to existing path	Converge to existing dead ends
Unknown	Discover a new path	Discover new dead ends

I'd argue that in the current pre-paradigmatic phase, we should value exploration of Unknown-Useless paths as highly as exploration of Known-Useful Paths. This is

especially true because it is unclear if Known-Useful paths are *actually* Useful! Thus, my focus will be on the bottom row - the Unknowns. But what does it matter if we aim for Known or Unknown paths and how should we evaluate the value of the two strategies?

My intuition is that aiming for Unknown paths, my probability of ending up in each cell is something like:

	Useful	**Useless**
Known	0.1	0.2
Unknown	0.1	0.6

So I expect about a 10% success rate for the ideal outcome, about an equal chance to end up on what most people following a set study path would end up on, and then a six times greater chance than that to go down a dead end path that was legitimately underexplored, which is also a good thing! My greatest worry is that any given dead end I explore will turn out to have been an obvious dead end to my peers in about a quarter of the cases, and that this outcome feels as likely to me as doing something Useful at all. However, I think focusing on the Unknowns is still worth it for the increased chance of finding Unknown-Useful outcomes.

In contrast, if we compare to aiming for Known paths I think I'd end up with the following probabilities:

	Useful	**Useless**
Known	0.8	0.1
Unknown	0.01	0.09

Cause it's hard to miss the target when you are on rails, but also nearly impossible to explore!

Now these probabilities say more about my brain, my self-assessment and my model of how minds work, than about the actual shape of reality. It's a way to convey intuitions on why I'm approaching alignment studies the way I am. Maybe I'm wrong and people explore just fine after focusing on existing methods, and then we can just reframe the above thinking as one of the paths I'm exploring -- Namely, the path of explicit exploration.

Either way, using this framework of Known-Unknown and Useful-Useless paths, highlights that marking the paths you take is a key item. It's an exploration of solution space, and we want to track the dead ends as much as the promising new avenues, or else we'll be duplicating work within the community. Thus, by writing down my research path others may retroactively trace back what *definitely didn't work* (if I end up on a Useless path) or *how breakthroughs are made* (if I end up on the Unknown-Useful path).

Thus I commit to exploring new paths for researching the alignment problem and documenting my progress along the way.

Paradigmicity: Solve the Problem

One of the errors I dread the most is to get sucked in to one research path with one specific problem and lose track of the greater problem landscape. Instead, I want to be sure I have an overview, a narrative, a map -- an overarching *paradigm* that I am working with. It should show how each problem I'm studying fits into an overall model of solving alignment. Honestly, of the three research principles, this is the only one I'd strongly argue for general adoption by all new alignment researchers:

Prioritize getting a complete view of the problem landscape and how your work actually solves alignment.

This is important for two reasons:

First, by keeping a bird's eye view of the interrelation of the major subproblems of alignment, your mind is more likely to synthesize solutions that shift the entire frame. There is a form of information integration that a brain can do that involves intuitive leaps between reasoning steps. Internally it feels like your brain has pattern matched into the expectation of a connection between A and B, but when you actually look, there are no obvious steps connecting the two. This in turn sparks exploration of possible paths that might connect A and B. Sometimes you find them and sometimes you don't, but either way, I suspect this type of high-level integrative cognition is key to solving alignment. As such, a bird's eye view of the problem should be at the front of one's mind every step of the way.

Secondly, with a map in hand between us and the destination point of solving alignment, you will be able to measure your progress so far. By having a coherent model of how each of your actions plays a role and can matter to the eventual outcome, you won't get lost in the weeds staring at the pretty colors of high dopamine-dispensing subproblems. Therefore, if someone asks me, "Shoshannah, why did you spend the last month studying method X?", then I should be able to coherently and promptly answer how and why X may matter to solving alignment from start to finish.

Thus I commit to working toward a coherent paradigm of AI alignment in which I can situate my work, explain how it contributes to solving alignment, and measure my progress toward this goal.

Conclusion

For my 6 months of AI alignment studies, I will aim to be transparent and explorative in my work while constructing and situating my actions in a coherent paradigm of the alignment problem. With this approach the journal entries of this sequence will be an exercise in epistemic humility.

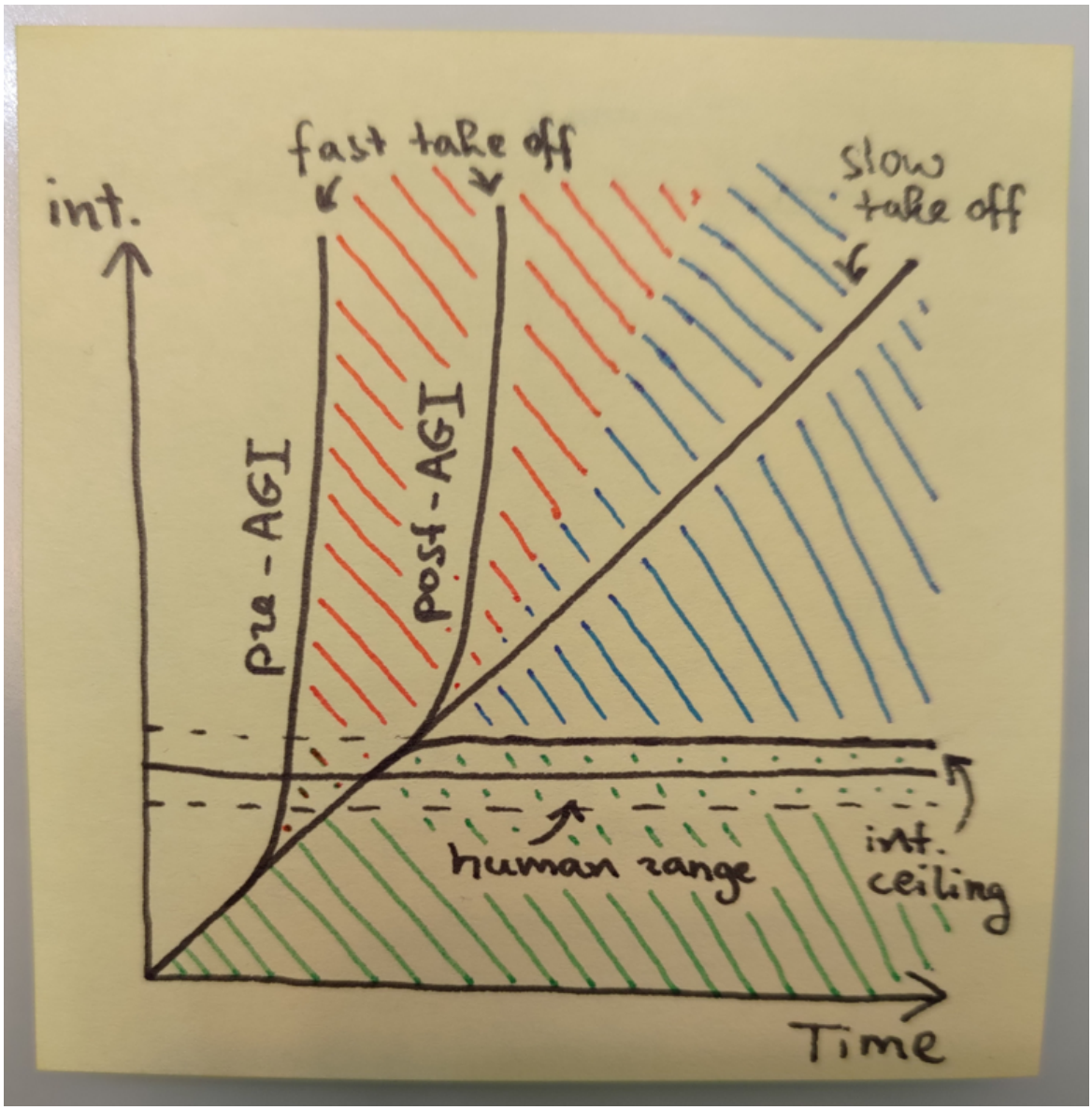
Wish me luck.

Three Alignment Schemas & Their Problems

This is the first journal entry for my 6 months of alignment studies. Feedback and thoughts are much appreciated! I'm hoping that documenting this process can serve as a reference for self-study methods in AIS.

I'm not sure how one goes about solving alignment. So for lack of an authoritative approach, I sat down to sketch out the problem from scratch. First off...

What happens when AI becomes more intelligent than humans?

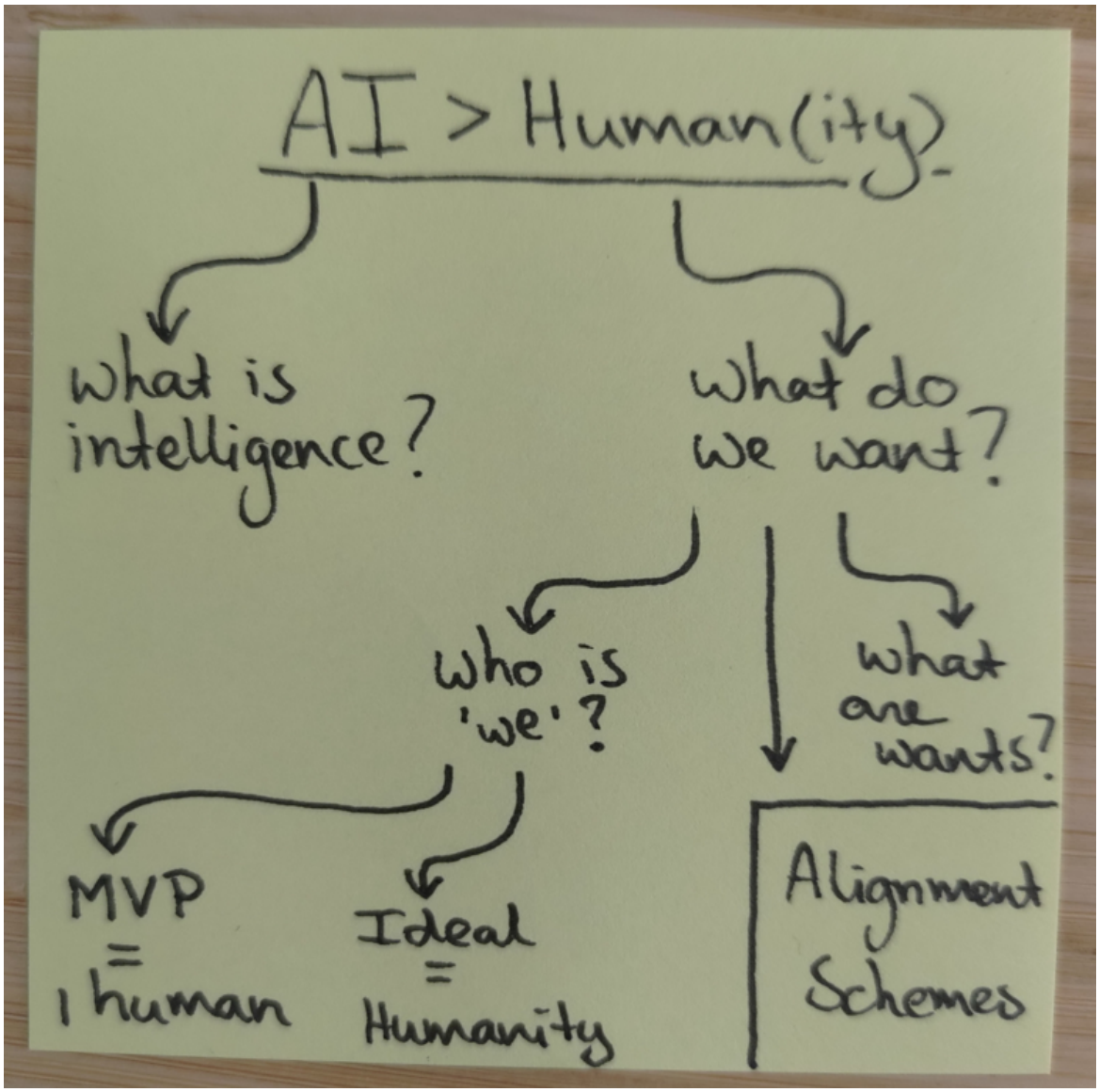


My understanding (and hope!) is that we are currently in the green area: AI is below human range -- We're safe.

Now what are the potential scenarios from this point? Let's review the entire possibility space.

1. *Intelligence Ceiling* - AI can't get smarter than us because human intelligence is about as good as it gets. This seems extremely unlikely on first principles, but we can't empirically rule it out.
2. *Slow Take-Off* - There is no Recursive Self-Improvement so we'll have time to adjust to AGI showing up. Problems will be progressive which means we can mobilize and experiment. This is kind of okay and manageable.
3. *Fast Take-Off Post-AGI* - Recursive Self-Improvement is real and kicks in around human-level intelligence. We're sitting on a ticking time bomb but at least we know when it will go off, and it's plausibly not extremely soon.
4. *Fast Take-Off Pre-AGI* - Recursive Self-Improvement is real and kicks in significantly before human-level intelligence! This is scary, and soon, and we'd better make sure all our current AI is aligned yesterday. Let's get to work.

What does AI end up doing as it becomes more intelligent?



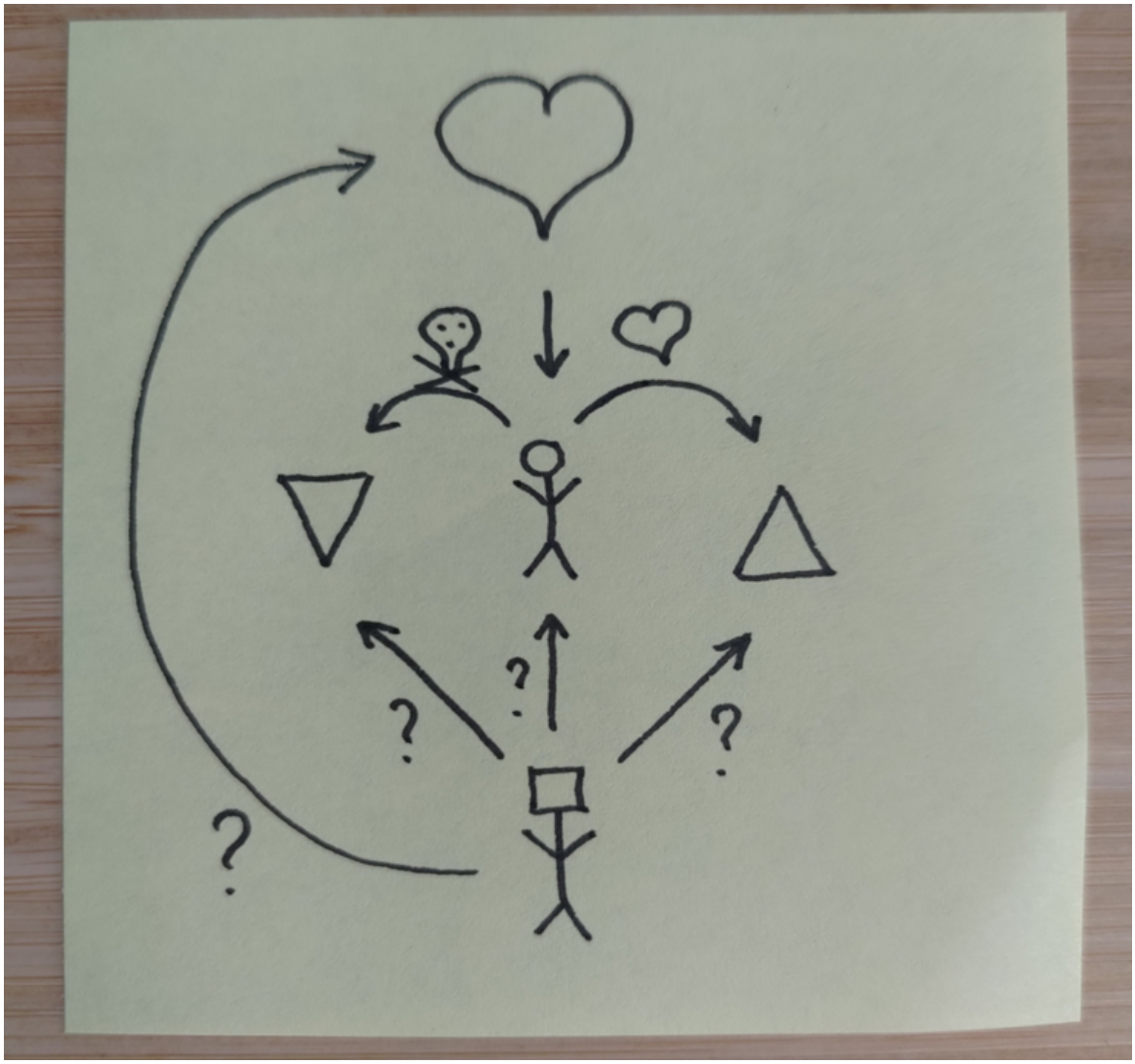
This question has a couple of branches. First off, we need to understand what intelligence is and what we'd like AI to do in the first place. This week I mostly dove into the latter part of that equation. Let's see what we want to align on before we figure out how to get there.

Yet the question what we want to align on also consists of two parts: who is this "we" and what are "wants"? Looking at the "we", the Minimal Viable Product (*MVP*) to solve for is a single human. From there, we can hopefully scale up to more humans and then all of humanity (*Ideal*). Notably, such scaling involves a process of value aggregation, which is a challenge in itself.

On the other hand, what are "wants"? This opens up a question on the nature of goals and if some goals might be easier or harder to align on than others. For instance, complexity of goals might make some goals easier to align on than others. But are there other properties of goals as well?

For now, let's say we want to align on as-yet-undefined goals of a singleton human (with options to scale up), what would that look like?

Alignment Schemas



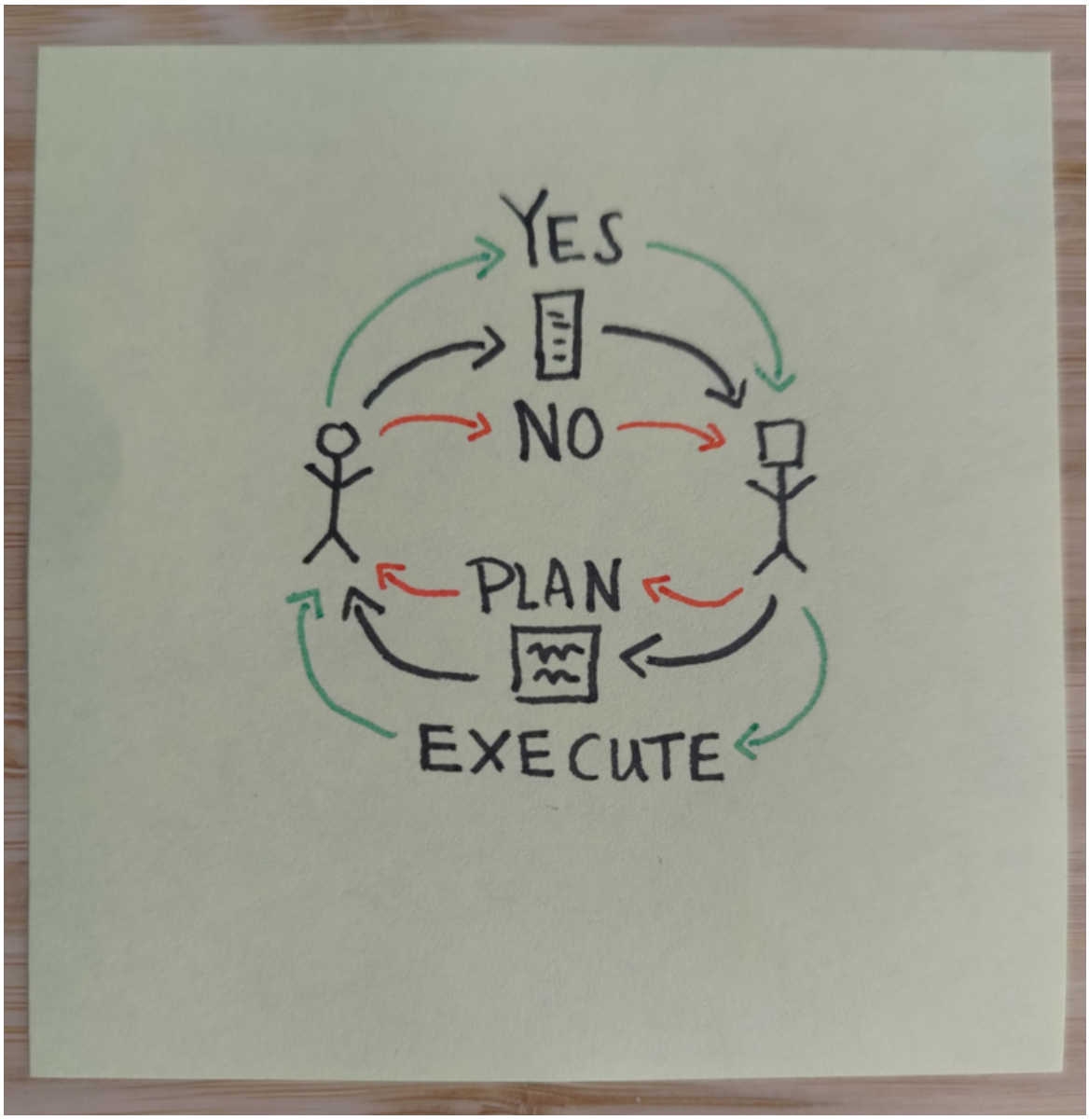
> Humans have round heads. AIs have square heads. My deepest apologies to humans with square heads.

I'm using the concept of an *alignment schema* as referring to a relational diagram that shows how human(ity) and AI relate to each other's goals. Overall, the situation seems to be that humans exist with preferences over world states. We prefer some things (hearts) and disprefer other things (skulls). These preferences are often not coherent, and often not even known to us a priori. They can also change over time, which means value lock-in is real and something we want to avoid.

Given that situation, what do we point the AI toward?

There roughly seem to be three options: Point it directly at us (Human in the Loop), point it at values that are benign to us (Empowerment), or point it at the exact things we like and dislike (Harm Avoidance, for reasons later explained).

Human in the Loop



This is the holy grail: An AI that listens to us directly. We feed our preferences in to an aggregator, the AI reads out the aggregator, the AI generates a plan based on our preferences, and we review the plan:

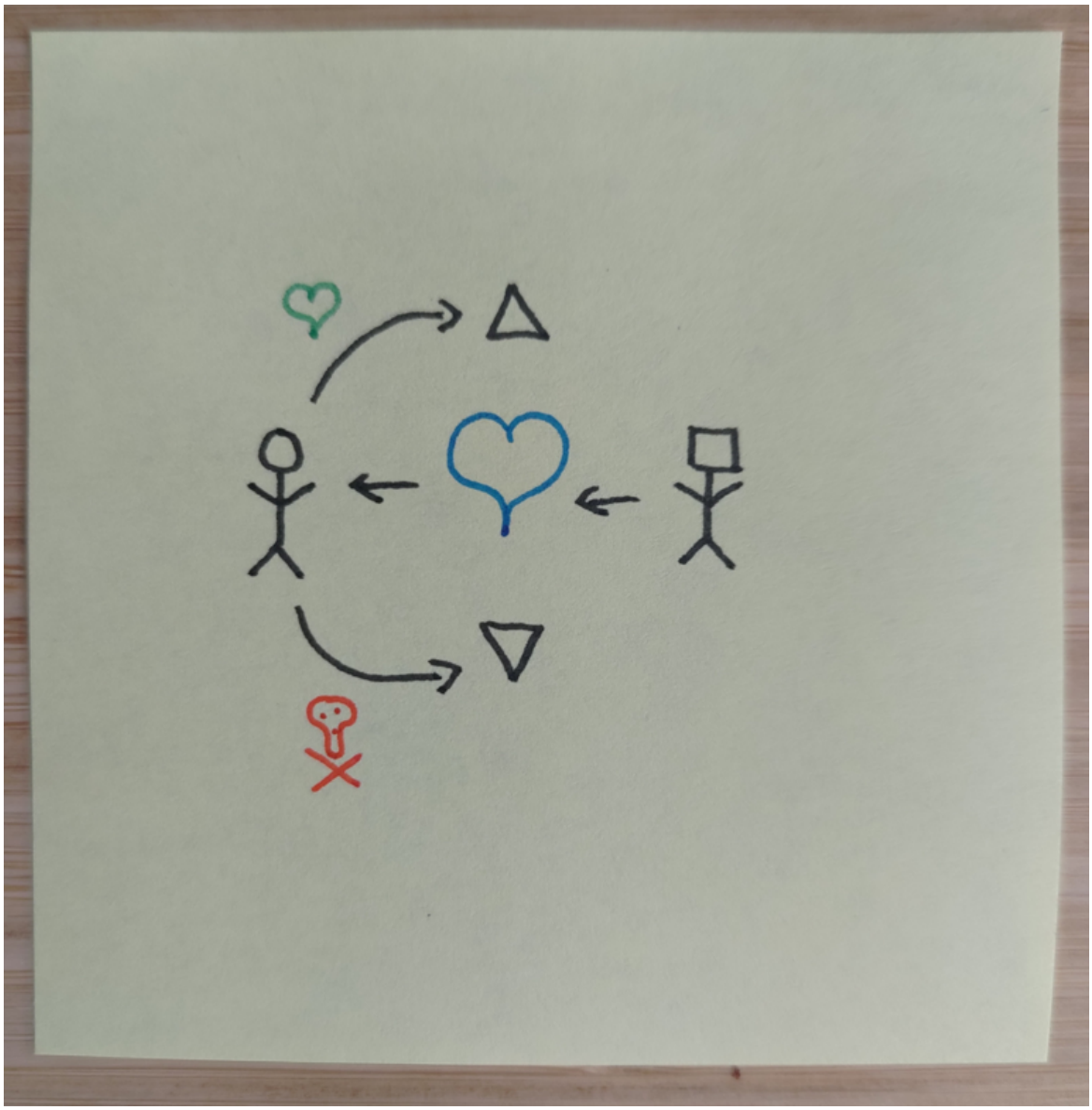
- If we don't like the plan (NO), we update our preferences to be clearer and more detailed, the AI reads them out again, plans again, and we review its plans.
- If we do like the plan (YES), then the AI executes the plan and gives us progress reports. If something goes wrong, we scream NO, it stops, and we go back to the previous step.

This schema relies on a form of corrigibility. Not only does the AI only execute approved plans it also allows the human to interrupt on-going plans. Additionally, the plans are annotated and formulated in such a way that they are human understandable and all it's impacts and outcomes are explicitly linked to the AI's understanding of the preference

aggregator. This caps out the complexity of the plans that the AI can offer up, and thus imposes a form of alignment tax.

There are a couple of things that could go wrong here. Deceptive alignment is always an issue as the AI can pretend to only execute approved plans while being busy with nefarious plans on the side. Similarly it can try to sneak problematic consequences of plans past us by encoding them in a way we can't detect. Next there is the question of what a preference aggregator even looks like, considering how human values are messy and incoherent. Lastly, there are issues around coordinating most of humanity to give input to a preference aggregator, as well as a question mark around emergent properties of intelligence once it's far past our own.

Empowerment



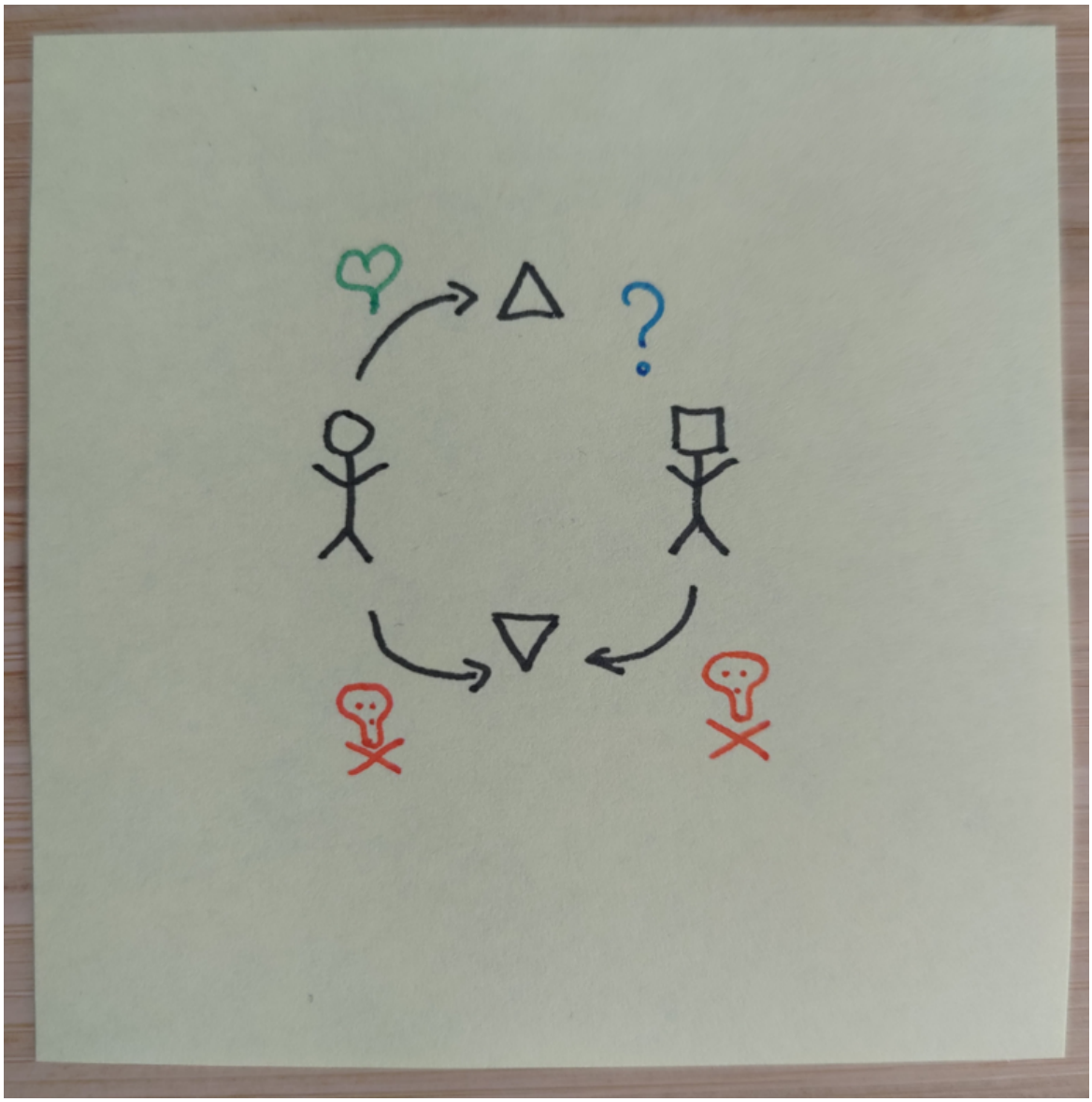
This is essentially sovereign AI -- Point the AI at humanity's instrumentally convergent goals and then deploy it. It would be a benevolent god preoccupied with supporting our self-

actualization and flourishing. A utopia of sorts, but one that seems so far removed as to be rather pointless to think about. Specifically, there is no obvious path between where we are now and developing such an AI (no incentive scale), cause what does half a sovereign AI look like?

It does however avoid any question of corrigibility or preference extraction. It simply gives us the tools to make our dreams come true without having particularly strong views on how we should go about doing that. We also don't need to coordinate humanity for this schema, cause a single lab with a magic bullet can launch the sovereign and it will be kind and empowering to all of us.

Of course there are the classical risks of deceptive alignment, manipulation and emergent properties. And specifically it's further burdened with needing to find an exceedingly robust goal landscape (cause you better be damn sure it doesn't have value drift toward nefarious goals cause there is no corrigibility) while distributional shift will also be a head cracker -- What do training wheels for gods even look like?

Harm Avoidance



This is a rather useless AI -- Peaceful coexistence with a superintelligence that does not hurt us or help us. I've listed it cause it may be simpler than the other two schemas. The issue is that if you try to point an AI at our preferred world states directly, then you are essentially signing up for value lock-in into perpetuity. On the other hand, if you make it corrigible then you end up on Human-in-the-Loop anyway. Thus you can only really point it away from doing anything we find really-really bad regardless. This is probably less problematic than value lock-in cause the most negative world states are fairly stable over time (don't wipe us out, don't torture us, etc). However, the AI won't be doing anything *useful* for us either.

Challenges per Alignment Schema

In an effort to determine what the path to aligned AI may look like, I tried to list the problems that need to be solved per alignment schema. This list is non-exhaustive and I'd be grateful for additions and thoughts.

CHALLENGES PER ALIGNMENT SCHEMA				
		Human in the Loop	*Harm Avoidance*	*Empower-ment*
AI Influence on Humans	Deceptive Alignment	x	x	x
	Manipulation Criterion	x	x	x
Nature of Values & Goals	Goal Robustness	-	x	x
	Preference Extraction	x	/	-
AI Governance	Human Coordination	x	-	-
	Incentive for Scalability	-	x	x
Other Problems	Distributional Shift	-	x	x
	Corrigibility	x	-	-
	Emergent Properties	x	x	x

Below is a brief description of each problem:

Deceptive Alignment **-** An AI that realizes it is in training and thus optimizes reward according to it's current utility function by pretending to already be aligned and then defecting straight after deployment (Steinhardt).

Manipulation Criterion - How do we define/encode/train an AI to minimally influence us considering that all interactions with humans can be considered a form of manipulation?

Goal Robustness - How do we avoid value drift after deployment? This is different from goal misgeneralization as this takes place over time and is not due to distributional shift.

Preference Extraction - How do we extract and aggregate human preferences? Maybe something like Coherent Aggregated Volition.

Human Coordination - Some schemas need more human coordination than others. The tractability of such coordination falls in the realm of AI governance.

Incentive Scalability - Some schemas are less likely to work cause there is no incentive for humans to build weaker versions of an AI with the given schema on the road to AGI. Specifically, a sovereign AI is not useful until its intelligence levels are superhuman.

Distributional Shift - Problems caused by the difference in distributions between the training and deployment environment, as defined here.

Corrigibility - The ability to correct an AI after deployment as defined here.

Emergent Properties - Humans have cognitive abilities that our far ancestors could never imagine. Similarly, superhuman intelligence may have properties we cannot imagine. There does not seem to be any obvious way to account for this, but it's worth keeping in mind that this may pose a risk.

Conclusion

In an attempt to home in on the key problems in alignment I've tried to deconstruct the problem space from first principles by identifying subquestions of alignment, formulating possible alignment schemas, and grading schemas on what subproblems need to be solved to allow them to succeed. Currently, a human-in-the-loop approach that integrates corrigibility, a preference aggregator, and limited and explicit plan specification seems most promising. Next week I'm hoping to look at the other branch of the alignment problem: what is intelligence and how do we instantiate it in machine learning?

Loose Threads on Intelligence

Epistemic Status: Unfinished deep-dive into the nature of intelligence [1] *. I committed to writing down my research path, but three weeks in I don't have a coherent answer to what intelligence is, and I do have a next question I want to dig into instead. Thus, here are the rough and rambly threads on intelligence that I've gathered. This piece is lower polish than I like cause of trade-off on writing-vs-research. Skimming might be more productive than a full read!*

Thread 1: Intelligence as path finding through reality

Intelligence is path finding through world states, where 'path finding' is a poetic term for optimization. Taking a closer look at optimization, it turns out that bad optimizers are still optimizers. Essentially, optimizers do not need to be optimal.

There exist three categories of optimization techniques:

1. optimization algorithms (finitely terminating)
2. iterative methods (convergent)
3. heuristics (approximate solutions, but no guarantee)

Genetic algorithms and evolutionary algorithms are optimization heuristics. Thus we can trace our past from the primordial soup through simpler and simpler optimization techniques, and we can project our future to the singularity through the creation of better and better optimization techniques. Humans are a point on this scale of increasingly sophisticated and optimally performing optimization techniques instantiated in reality.

Each of the optimization techniques can in turn be instantiated in three different ways:

- Mechanical
- Computational
- Collective

I made these up -- There must be an existing framework that outlines something like this. Or maybe I'm misunderstanding the concept of optimization or how one can categorize the types of instantiations. Either way, here is what I mean by each:

Mechanical optimization cannot learn. It's a tree growing toward the light or a water wheel generating power.

Computational optimization can learn but cannot be divided. It can compute all computable functions (Turing machine or a human with pen/paper). However, if you break up the cognitive processing parts, no computation will take place.

Collective optimization can be divided. Every unit can implement mechanical or computational optimization in itself, and the units work together emergently or coordinately to a greater result than the individual pieces. For instance, a fungus can be split in two such that both halves will keep growing and functioning as individuals. A flock of birds can be split in two such that both halves will coordinate their flight in

the same manner as when they were one. And of course, human societies can be split up in two and both halves will coordinate again in to societies.

The structure of deep learning mimics the structure of intelligence as path finding through world states

Intelligence = Mapping current world state to target world state (or target direction)

Deep learning = Mapping input layer to output layer

This seems analogous to me, but maybe it's not. My reasoning is that deep learning relies on hidden layers between in- and output layers. Learning consists of setting the right weights between all the neurons in all the layers. This is analogous to my understanding of human intelligence as path finding through reality -- in machine learning, a neural network is finding the function that maps inputs to outputs. In human intelligence, we look for the actions that maps the current state of reality to a desired future state of or direction through reality.

Maybe this is a tautological non-insight.

Segue on data augmentation

Data augmentation is transforming input data such that the network can learn to recognize more forms of that data and extract different features from it. Is human imagination and "thinking through different ways past events might have gone" a form of data augmentation? We perturb a memory and then project out how we would have felt and what we would have wanted to do. This seems quite similar to using simulation to generate and improve predictions.

Thread 2: Alignment as preference mapping

My core insight here is 4 reasoning steps followed by an intuitive leap:

> *Neural Networks can encode any computable function.*
>
> *Our neural activity is a computable function.*
>
> *Our utility function is encoded in our neural activity.*
>
> *Thus a neural network can encode our utility function.*

Beep-boop-brrrrrr -- MAGIC LEAP:

> *An aligned AGI is one that has learned the function that maps our neurally encoded utility function to observable world states.*

This seems true to me but maybe is not -- Loose threads indeed.

Alignment and measurement error of human-in-the-loop

Alignment is human preference profiling performed by an artificial intelligence. In preference profiling, you need to make decisions on what input parameters you will use to predict the output parameter (preferences, in this case for world states instead of products). Input parameters can be behavioral, linguistic, or biological. They can also be directly elicited or indirectly observed. Behavioral and linguistic measures are imprecise because actions are only outputted by humans based on how their own cognitive ability and conflicting drives end up converging into actions. A lot of actions are suboptimal cause humans are not good optimizers. Thus the most reliable signal of the human utility function is either:

- Aggregation over a large enough sample that all the noise is cancelled out
- Direct biological measures of our utility function

However, who says there are no systematic biases and errors in our behavior that do not cancel out over large samples?

And who says that observing our utility function directly won't change it through observation? Our experiences change us, and if our experiences are limited to being measured in a lab room, then this will not represent anything current humans consider to be our utility function.

Notably, HLRF relies on linguistic (and/or behavioral) mappings, so that leads our humans-in-the-loop into the faulty mapping between what we actually want and our words and actions.

Human Utility Functions are more hyper- than parameter

Hyperparameters are parameters across your parameters. For instance, learning rate is the parameter that controls how much you update the weights in a neural network at each step. Human utility functions seem to have hyperparameters too, which makes conceptualizing and encoded them complicated to say the least. Specifically, humans gain utility directly from various stimuli and observations like eating sweet food or looking at puppies. These would be the parameters of the human utility function. But much of the utility people strive for is not this direct hedonic payoff. Instead, we have many (scarcely known) hyperparameters where the utility we get from our observations comes from the transformation and evaluation of one or many sets of observations. For instance, the satisfaction of a job well-done relies on observing the entire process and then evaluating the end result as good. Similarly, many observations that consist of directly negative stimuli (parameters) are evaluated as positive by some hyperparameter such as the meaningfulness of childbirth or the beautiful release of a funeral.

The evaluation of the aggregates of our observations even change our biochemistry such that hyperparameters influence the parameters of direct experience. For instance, evaluating someone's social cues as them liking you can directly generate feelings of relaxation that are physiologically embodied and thus direct parameters. While the exact same encounter, if evaluated negatively, could cause tension in the body that is also a direct parameter. Thus the exact same stimuli result in completely different reward signal purely based on the settings of the hyperparameters that control how a set of observations is transformed and then evaluated.

Thus it seems conceptually straightforward to map the parameters of our utility function in to something learnable by an AGI, but it's much less clear how we'd map the ever-fickle hyperparameters of our utility function that *entirely* hinge on our evaluations and transformations we ourselves apply to our experiences ... it's a value we compute internally that would require the AGI to simulate us as full-bodied beings to get the exact same result. This would be undesireable cause such a simulation can thus validly suffer as much as we ourselves do. And thus we don't want to map our utility function directly to an AGI but use some proxy. And the only sensible proxy is then "do as I say, don't do as I seem to want to do", which then boils down to needing corrigibility.

Collaborative Filtering on Values?

Collaborative filtering finds the latent factors for how to match two types of things together (like humans and movies). Are there latent factors to humans and the values they espouse? If you run Principle Component Analysis on the values, would you get a few limited clusters? This seems easy to google and has probably been done, but probably was also not encoded well and it's hard to see how one would accurately extract value data from people such that the analysis makes sense and has useful results.

Thread 3: Natural language as data compression

Language is a data compression format that inherently encodes relational properties across abstract entities such that models of reality can be communicated and reasoned about. In contrast, images are sense data, where sense data can be compressed, but inherently is not. Similarly, sense data can encode relational properties across abstract entities, but inherently does not (e.g., a picture of a book with language in it, or a picture of a diagram).

Is it then true that AGI can result from language models but not from image models? The counterargument would be that language models lack grounding in reality. Image models can be grounded in reality cause they consume sense data and thus can be hooked up to cameras. However, we've created systems that allow sense data to be directly translated to language data and language data to be directly translated to actions. Thus, even though an abstract data compression format like language is not inherently grounded in reality, we have given it eyes and hands such that it can sense and act in the real world without directly consuming sense data or outputting motor data.

So actually, AGI can result from image models that read and write, but that's many much more steps than you'd need when using a language model. Thus AGI from language models will exist before AGI from image models or other sense-data-only models.

What's mentalese?

Human reasoning happens in "mentalese". People's introspection on how they reason is plausibly faulty, but many people have some experience of reasoning in language, imagery, and spatial-relationally. Are these just a side-effect of reasoning, and does it all take place "under the hood" anyway? Could one reason without having any conscious process of reasoning? Presumably, yes. Is that what the zombie-discussion points to? What happens if we input both language and image data in to a big enough neural network? Will reasoning then take place in both? Is there any value in enhancing intelligence with sense data?

Supervised Learning as the bootstrap of collective intelligence

Self-supervised learning is the default form of learning for individual agents embedded in reality. You make a prediction of what reality will look like, and then time passes and you see if your prediction is true. Or you make a prediction of what reality will look like if you do an action, then you do the action, and see if it is true.

Supervised learning in contrast is a form of collective intelligence. It only works if another intelligence has already learned the mapping and can thus output the labels for you. So supervised learning is how we bootstrap AI and launch it to a much higher entry point than we as biological organism could start with. We've learned to integrate supervised data since we're a collective intelligence that uses language (mostly) for coordination. However, self-supervised learning is the only option for an AGI to learn things we don't know yet.

Even Looser Threads

Feature engineering seems like a form of pre-processing, and thus not a relevant concept for AGI? We'd expect AGI to learn it's own features. Which is what kernels in convolutional neural networks do, for instance.

Overfitting in a neural network is basically memorizing the data set. This lines up with Steve Byrnes' explanation of most all of the brain being essentially memory models, but particular types of memory models. But how does this work exactly?

Are System 1 and System 2 reasoning pretty much a 2-piece ensemble? If so, wouldn't you expect more models? Maybe there are? Maybe we are integrating those lower down? This seems not super relevant.

What's a positive feedback loop called in the alignment problem? In current ML you already need to watch out for positive feedback loops if the output of the network influences the input it will later get. The given example is a collaborative filtering network that matches users to movies, but then mostly the matched movies will be watched and thus rated, and thus matched again, etc. This clearly creates a massive issue with AGI that interacts with humans ... what is this problem called in the existing alignment literature? I had a concept on this called "manipulation threshold" meaning

some formalization of how much and what kind of influence an AGI is allowed to have on any human when discussing plans that have not been signed off yet (as the other elements will be subsumed by corrigibility).

1. ^

   The deep-dive consisted of running through the Fast.AI course in 2 weeks, generating my own spin-off questions from that, and then conceptually working through the nature of intelligence from scratch on my own, googling little bits and pieces as I went. The main body text will contain as many references as I can recall to link insights together and to sources.

Optimizing Human Collective Intelligence to Align AI

Commitments are hard. Based on my Research Principles for these 6 months of self-study, I'm writing up all my ideas regardless of level of polish. So here is the rough sketch of the idea I'm currently working on. Forgive the slightly over-confident style -- It's closer to the raw form of my cognition and not necessarily the level of nuance I reflectively endorse.

Will alignment be solved by a handful of geniuses or by the collective efforts of hundreds of the merely gifted? My understanding is that we are already trying our best to draw every genius out of the woodwork. Yet how much optimization pressure have we exerted on the collective efforts of the remaining 99% of alignment researchers?

Not much, I'm guessing?

So this is a short essay on what optimizing human collective intelligence may look like.

Known and Unknown Knowledge Space

First, some framing.

If you presume the solution to the alignment problem exists somewhere in knowledge space then what we need is a method to correctly and quickly navigate toward that solution. If we conceptualize knowledge space as a graph, where each node represents a modular piece of knowledge and edges are the relationships between these pieces of knowledge, then we could theoretically encode all of known and unknown knowledge in this form. Next, there is the *individual skill* of navigating across knowledge space as a researcher as well the *collective skill* of efficiently and completely searching knowledge space as a group of researchers. My proposal is that we need to improve both to find a reliable solution to the alignment problem before the development of AGI.

Discovering New Nodes

The search for new knowledge is what we refer to as 'learning'. You can discover a new knowledge node through three different methods:

1. Logic
2. Empiricism
3. Communication

Logic yields knowledge through interpolation and extrapolation from observations or assumptions. As such, it can be practiced entirely in theory. Mathematics is the foremost field that relies completely on logic, where I'm loosely using *logic* to refer to any form of theoretical reasoning.

Empiricism consists of observation. Experiments are controlled observations you perform to gather specific, generalizable knowledge from reality. It is challenging to run good experiments that tease out complex or precise relationships in reality, but it's generally easy to run small-scale experiments that generalize well within the limited-scope of your daily life. For instance, it is reasonably straightforward to figure out the average properties of males in your small hamlet of 100 people, but it is immensely challenging to figure out average properties of males in our species. Similarly, a child quickly learns through experimentation how gravity works as relevant to its daily life, but most humans don't end up with the ability to rediscover that the gravitational acceleration on Earth is roughly 9.81 m/s^2.

Communication is the pathway we use to accrue collective knowledge. Through language we can encode all we learn and access all anyone else has ever learned. Much of the knowledge you acquire comes from communication because it is efficient - It is far faster and less error-prone to be *told* about photosynthesis than to *derive* it from scratch. Our human lives are too short to derive all collectively known knowledge. And thus we create knowledge speedruns for our children, and call them 'schools'. Education optimizes paths through collectively known knowledge space for individuals that do not have those knowledge nodes in their personally known knowledge space.

Now when it comes to research, we are at the frontier of our collectively known knowledge space and from there we attempt to forge out into the unknown. Discovering new nodes in knowledge space is achieved through *logic* and *empiricism*. And the only people who will be good at this process will be those that practiced their skill at developing proofs and designing experiments back in known knowledge space, where they could check their answers against the collective. Thus, *research is a skill you practice against known facts.*

Individual Search Algorithm

The *scientific method* is the current algorithm individual researchers run to discover new knowledge in unknown knowledge space. Researchers are traditionally trained in the scientific method during PhD's, which is basically an apprenticeship slathered in tradition, and sealed with a protected title. For any real world problem, the scientific method relies on iteration across empirical experiments, designed and analyzed using logic.

Now here comes the rub.

The alignment problem is a *real world* problem we need to solve in *one-shot*, on a *deadline*.

Real world problem - Alignment cannot be reliably reduced to a mathematical equation without testing that that mathematical equation holds up in the real world. Reality is cognitively uncontainable, and additionally the alignment problem specifically refers to entities who's intelligence will also be uncontainable to us. Thus we need some form of iteration in order to run experiments to gather data on which alignment solutions work and which don't.

However.

One-Shot - Once an AGI exists it better be aligned, or we're dead. We currently have no way to turn off an AGI, pause an AGI, or redirect an AGI. The alignment problem is by it's very nature a one-shot problem. We need an empirical solution, but cannot currently iterate.

And to make matters worse...

A Deadline - Alignment needs to be solved before an AGI is created, but the alignment and capabilities branches of research and development are independent. It's like we're crawling along the axes of the Orthogonality thesis, where we are getting better and better at shaping *intelligence*, while hardly moving an inch in shaping *motivations*.

This is not how science is normally done!

Normally you run endless experiments, have all the time in the world, and only financial and prestige incentives to win or lose. Sometimes there is a problem where many lives are at stake like the Manhattan Project, COVID, or climate change. But still, that is not all of humanity, and being slower can mean *more* humans dying, yet that is not nearly as bad as *all* humans dying.

So what do we do?

We need alignment research to be faster and of higher predictive value. We need to navigate the unknown knowledge graph more quickly and with fewer missteps. And maybe that will still not be enough, but I suspect it's the direction in which "enough" can be found.

My proposal is thus that we need an expanded form of the scientific method and more rigorous methods to test and teach it. Currently, I'm guessing that this expansion consists of adding the properties "security mindset", "superforecaster", and "speed-learning"[1].

Collective Search Algorithm

Once we have every individual researcher running a more optimal search algorithm in their minds, we can further optimize our collective search algorithm by improving the *coordination* of the researchers across unknown knowledge space. What properties would such an optimal coordination have?

- *Encoding* - Finding an encoding for unknown knowledge space allows us to map out where new knowledge nodes may be found or solution paths may be discovered.
- *Distribution* - Tracking the distribution of researchers across unknown knowledge space will allow us to see what areas are overpopulated and which may be comparatively neglected.
- *Query-able* - Ensuring the knowledge space encoding and the researcher distribution are query-able across multiple dimensions will allow researchers to easily connect with each other based on similarity of research topics and

properties of researchers. This may lead to more effective research collaborations.

Achieving the Distribution and Query-able properties will presumably hinge on improving coordination tools between researchers (e.g., journals, search engines, research databases, etc). Encoding knowledge space in a searchable format would be more in the order of a paradigm shift for how research is done. It may not be achievable at all as it relies on discovering the underlying structure of knowledge space in such a way that we can predict where solution paths to problems may be found. However, I'd like to explore the question nonetheless, and I have one naive baseline proposal that I think is better than nothing, but far from optimal:

$Solutions^2$

New knowledge nodes are discovered through expansion of existing nodes or recombining two or more existing nodes. Thus we can naively list all knowledge nodes related to alignment (proposed solutions, existing concepts, and well-defined problems) and invite researchers to explore the combination of any two nodes. The outcomes can be visualized in a table listing all current knowledge nodes as both row and column headers. Every cell in the table denotes one possible combination. Some combinations will be clearly dead ends, while other combinations may already be heavily researched. It's a naive and imperfect encoding of unknown knowledge space, but it does give us some limited access to the desired properties mentioned above:

- *Encoding* - It offers an overview of where new research directions may be found, though imperfectly and incompletely.
- *Distribution* - You can clearly visualize the number of researchers in each cell of the table.
- *Query-Able* - If you link this knowledge space encoding to a database of researchers then researchers and knowledge nodes can be queried for potentially high value collaborations.

Many senior (AIS) researchers probably already have an implicit map of unknown knowledge space in their mind (though it is likely trimmed to exclude low value areas). Formalizing this encoding in a tool that people can use to navigate unknown knowledge space offers mostly benefits for junior and medium experience researchers. Specifically, it allows new researchers to get an overview of where most of the work is happening and what areas may be promising. Second, it offers a structure to organize researcher databases around. Third, the 2 dimensional table could potentially be used to organize researcher publications, so the known knowledge base is coordinated more rigorously.

Now the solution^2 encoding is too simplistic -- It will take a lot of upkeep, doesn't encode more complex relationships between knowledge nodes, and the dimensionality is far too low. Yet, it illustrates the type of coordination mechanics I think we should be exploring.

So what now?

But before I explore that, I think I need to determine if my native research algorithm is anywhere near optimal (and improve it if it is not). Specifically, I want to test myself on the three properties that I suspect are necessary for predictive and fast AIS research: Security mindset, superforecasting, and speed-learning. I will start by testing myself on security mindset and speed-learning by writing a critique on the OpenAI alignment plan. If I write a good critique then I will hopefully get feedback from OpenAI and MIRI, who seem to be relatively far apart in 'alignment paradigm space', and thus offer high value calibration data for my development. If I write a bad critique, then that's probably data in itself too -- And a trip back to the drawing board.

1. ^

   I'm not sure what this property looks like, and maybe it's already subsumed in the current scientific method. Either way, it's the property of navigating known knowledge space more quickly and efficiently, and is thus a meta-learning property. Conceptually it would be a subskill of audidactism.

Reflections on Deception & Generality in Scalable Oversight (Another OpenAI Alignment Review)

Just like you can test your skill in experimental design by reviewing existing experiments, you can test your skill in alignment by reviewing existing alignment strategies. Conveniently, Rob Bensinger, in name of Nate Soares and Eliezer Yudkowsky, recently posted a challenge to AI Safety researchers to review the OpenAI alignment plan written by Jan Leike, John Schulman, and Jeffrey Wu. I figured this constituted a test that might net me feedback from both sides of the rationalist-empiricist[1] aisle. Yet, instead of finding ground-breaking arguments for or against scalable oversight to do alignment research, it seems Leike already knows what might go wrong — and goes ahead anyway.

Thus my mind became split between evaluating the actual alignment plan and modeling the disagreement between prominent clusters of researchers. I wrote up the latter in an informal typology of AI Safety Researchers, and continued my technical review below. The following is a short summary of the OpenAI alignment plan, my views on the main problems, and a final section on recommendations for red lining.

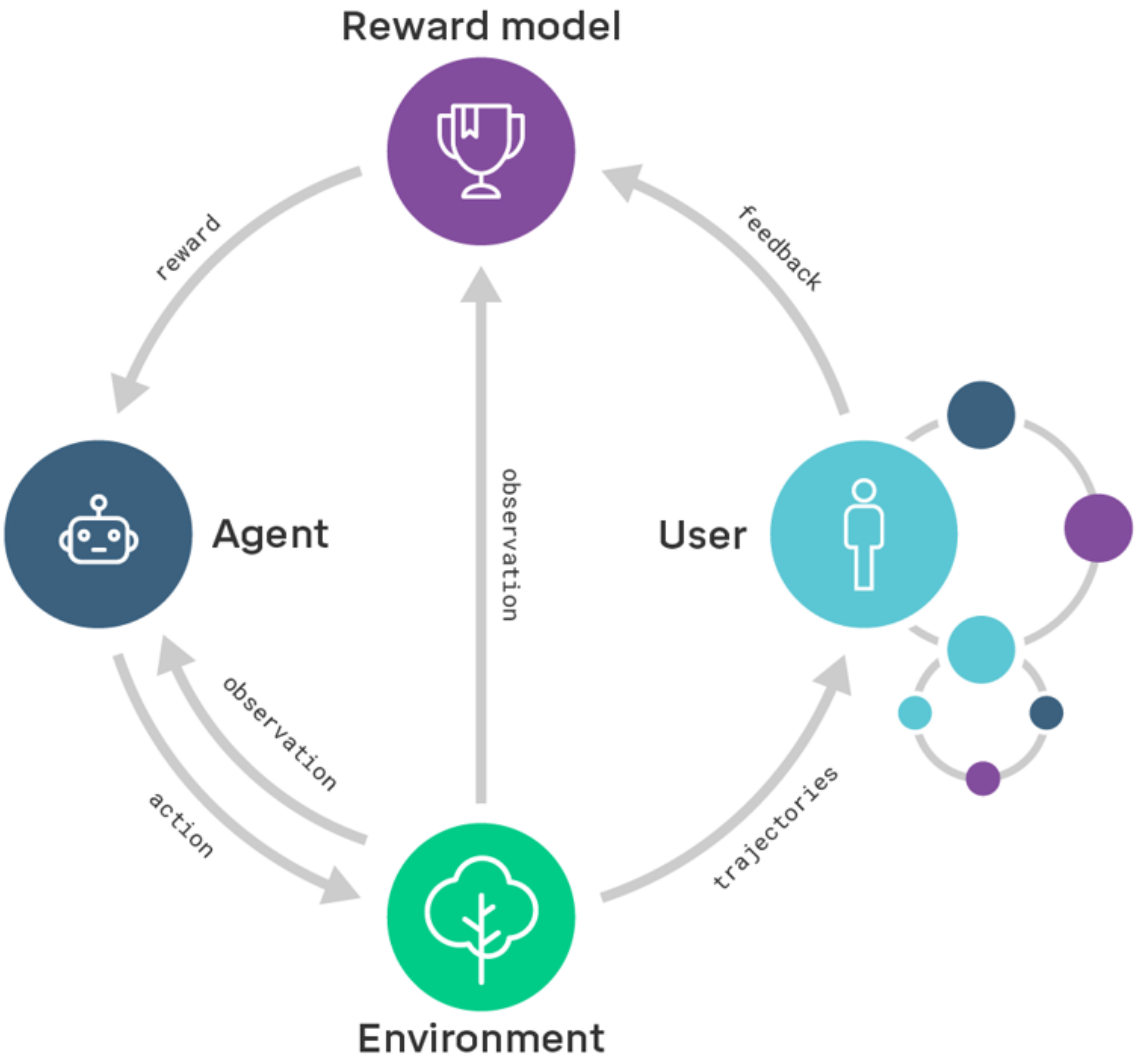
The Plan

First, align AI with *human feedback*, then *get AI to assist* in giving human feedback to AI, then get AI to assist in giving human feedback to *AI that is generating solutions to the alignment problem*. Except, the steps are not sequential but run in parallel. This is one form of Scalable Oversight.

Human feedback is Reinforcement Learning from Human Feedback[2] (RLHF), the assisting AI is Iterated Distillation and Amplification (IDA) and Recursive Reward Modeling (RRM), and the AI that is generating solutions to the alignment problem is… still under construction.

The target is a *narrow* AI that will make significant progress on the alignment problem. The MVP is a theorem prover. The full product is AGI utopia.

Here is a graph.



> “Schematic illustration of recursive reward modeling: agents trained with recursive reward modeling (smaller circles on the right) assist the user in the evaluation process of outcomes produced by the agent currently being trained (large circle).” - Source

OpenAI explains its strategy succinctly and links to detailed background research. This is laudable, and hopefully other labs and organizations will follow suit. My understanding is also that if someone came along with a *better* plan then OpenAI would pivot in a heart beat. Which is even more laudable. The transparency, accountability, and flexibility they display set a strong example for other organizations working on AI.

But the show must go on (from their point of view anyway) and so they are going ahead and implementing the most promising strategy that currently exists. Even if there are problems.

And boy, are there problems.

The Problems

Jan Leike discusses almost all objections to the OpenAI alignment plan on his blog. Thus below I will only highlight the two most important problems in the plan, plus two additional concerns that I have not seen discussed so far.

Nearest Unblocked Strategy will eventually lead to a successful deception or manipulation of humans - The OpenAI plan relies on humans detecting and correcting undesirable output at each step, while the AI becomes increasingly intelligent along the way. At some point, the AI will be capable enough to fool us. It is unclear if Leike et al. recognize this danger, but I think they argue that we will stay ahead of the AI through the alignment techniques produced by the researcher AI and that misgeneralization errors can always be detected by humans supported by smart enough AI assistants. However, how will we know if the assistants remain well-enough aligned[3]? And if they do, how do we know if they are smart enough to help us notice misalignment in the target AI?

Alignment research requires general intelligence - If the alignment researcher AI has enough general intelligence to make breakthrough discoveries in alignment, then you can't safely create it without already having solved alignment. Yet, Leike et al. hope that relatively *narrow* intelligence can already make significant progress on alignment. I think this is extremely unlikely if we reflect on what general intelligence truly is. Though my own thoughts on the nature of intelligence are not entirely coherent yet, I'd argue that having a strong concept of intelligence is key to accurately predicting the outcome of an alignment strategy.

Specifically in this case, my understanding is that general intelligence is being able to perform a wider set of operations on a wider set of inputs (to achieve a desired set of observations on the world state). For example, I can do addition of 2 apples I see, 2 apples I think about, 2 boats I hear about, 2 functions with all their subcomponents, etc. If you now teach me subtraction or multiplication or another transformation, then I can attempt to apply them to all of these inputs. This type of "folding" of operations across domains is the degree to which an intelligence is *general* instead of *narrow*.

The alignment problem needs high general intelligence, because it needs *new ideas* for solving alignment. It won't be enough to input all the math around the alignment problem and have the AI solve that. It's a great improvement over what we have but it will only gain us speed, not insight. To generate new ideas, you need to include information from a new domain in order to discover a new transformation to apply to your original problem. That means the AI needs to be able to reason about other domains. And we don't know in advance which domains will contain the insights for key transformations. Thus we will have to give the AI access to many domains before it will learn the right transformation to apply to the alignment problem.

Now, from an empiricist view, you might ask, what's the harm in trying to solve alignment with narrow intelligence anyway? If it doesn't work, then we just stop and do something else. I'd argue there are two types of harm though.

First, ungrounded optimism itself may shorten timelines by shifting risk assessments downward, which in turn increases competitive pressure among companies developing AI. Keeping AI development as slow as possible is a coordination problem that revolves around "how slow are we all willing to go to make sure AGI will be safe?". If one company "defects" by going more quickly, then all other companies are incentivized to drop their safety standards farther down in order to keep up. This mechanic is hard to counter, but at minimum one can choose to project the lowest level of optimism such that you can contribute to a culture of caution.

Secondly, if the alignment researcher AI is not producing useful alignment results, at what point would you stop adding a little more training data? At what point does it cross-over from too-narrow-to-be-useful to too-general-to-be-safe? Leike et al. expect there to be an area of useful and safe AI, but what would that look like? Incentives to push the AI ever so slightly more general to get a success will be very high, and you don't want to put that type of risk into a slippery-sloped incentive structure fueled by massive economic pressures.

Note on IDA - How does it deal with deconstructing cognitively uncontainable tasks? We should consider if such tasks contain elements across levels of deconstruction that we cannot anticipate, because we cannot cognitively contain them. This may be expressed in two ways: Either a task may not be safely decomposable at all (“atomically difficult”), or a task may have the property that humans may not be able to foresee how to deconstruct it such that all safety-relevant considerations will be covered by evaluation of the individual elements of the decomposition (even though such a decomposition may exist). I didn't find a discussion of this question in the literature, while it may be a necessary condition for this approach to be safe. I worry iteration wouldn't help either, because how would you test that you are capturing all information that is otherwise missed by a single human trying to work through the task? Additionally, the most crucial tests we can run on this problem require the AI to be running across cognitively uncontainable tasks, and thus we're already close to AGI. One counterargument is that there are tasks that are cognitively uncontainable to us but safe to run (like most larger software packages or complicated calculations). Conversely, there are also tasks that are cognitively containable to us where we can still easily be fooled (like any act of magic between humans -- you could understand it if you knew the trick, but you don't). However, it seems that it's far easier for us to miss safety-relevant considerations in task decomposition for tasks we can't cognitively contain than in those that we can.

Note on Evaluation is Easier than Generation - Is evaluation easier than generation when evaluating *preferences*, and does it increase the risk of deception? My understanding is that the claim is based on the P versus NP debate, where most likely it is the case that mathematical and computational solutions are easier to evaluate than generate. However, is it true that human minds work the same? Is it easier to evaluate if a pre-generated solution lives up to your preferences than a solution you generated yourself?

As an intuition pump, we can look at how most humans spend their leisure time. Generation of ideas on how to spend leisure time would be akin to sitting quietly, possibly meditatively, and seeing what suggestions emerge inside you on how to spend your time. Evaluation of ideas on how to spend leisure time would be like receiving a menu of options, and you go down the menu till you find an option you find agreeable enough (or you pick the most preferred option above a certain threshold value). In which case, do you think you would pick the option that you most endorse? When would you be the happiest? What are the pitfalls of each situation?

It is not clear to me that any situation in which our human preferences matter will be subject to "evaluation is easier than generation". Instead, I'd sooner worry our preferences will be lightly influenced over time, such that they take a different course all together as slight nudges accumulate. A simple and thoroughly researched example of this is priming, where a user is exposed to a stimulus that subconsciously shifts their responses to follow up stimuli. Priming effects can take place even unintentionally, and so any task that requires evaluation of preferences is sensitive to them. This problem becomes even more salient when we have an alignment researcher AI suggesting solutions that kind of plausibly give us what we want, but maybe not really. Basically, I worry that we are opening up a channel for AI to manipulate us in unrecoverable ways.

Which also segues into a concern about deception: Will we not be more easily deceived (intentional or not) by solutions we did not generate, but instead only evaluate? I don't have a fleshed out and specific instantiation of this yet, so I can only lightly touch on this consideration.

Red Lines & Recommendations

To address the two main problems outlined above, I would recommend diving more deeply into where we might expect these problems to show up. Are there red lines that we can precisely define, that if left uncrossed, keep us in the green? Specifically:

What does near-deception look like? OpenAI's DC gap is a first attempt at quantifying latent knowledge. Similarly, can we use other techniques to *predict* deception? If we are iterating anyway, can we develop high predictive accuracy on the current models? Presuming there is a continuous function of deception ability, then we could at least decide about the specific cut-off values of deceptive ability - the point at which we abort any given alignment attempt and go back to the drawing board.

What does general intelligence look like? OpenAI talks about an alignment researcher AI being narrow in the same breath as scaling up GPT models with human feedback. How are we defining narrow and general AI in this case? Is the plan to use a GPT-like model to do the alignment research? If not, then what is the "narrowness" threshold for AI that will not be crossed in order to keep the researcher AI in line?

In Review

Overall, my concern with the OpenAI alignment plan is three-fold. First, it doesn't contain guard rails that make sure we don't "iterate too close to the sun". Secondly, it sets an example of optimistic over-confidence that may accelerate capabilities progress. And lastly, it will not actually bring us closer to solving the hard parts of the alignment problem without inadvertedly bringing about AGI before we are ready.

So, how did I do at evaluating alignment strategies? I'd love to hear your thoughts! I'll be using the feedback I receive to decide on next steps in my skilling up journey. Either I'll continue working on ways to increase Human Collective Intelligence so we can better tackle the alignment problem, or I'll pivot in a new direction.

Additionally, my thanks goes out to Erik Jenner, Leon Lang, and Richard Ngo for reviewing the first draft of this document. It helped me clarify a couple of ambiguous claims, actually drop two of them as incorrect, split my document in two, and avoid accidental strawmanning moves toward OpenAI/Leike et al. Thank you!

1. ^

   "rationalist" here refers to general rationalism and not LessWrongian rationalism. Though the two are very closely related, I'd like to avoid any associations particular to this community. I considered using a different term but "rationalism" is just what this view is called and introducing a new term was more likely add to the confusion than detract from it. Also, one can definitely be a mix of both rationalist and empiricist.

2. ^

   This meta-review article by Lambert et al. (2022) is amazing if you want to learn more about RLHF. It's so amazing I made this footnote just to increase your chances of clicking the link -- Click.

3. ^

   Erik Jenner introduced a key argument to me related to error accumulation in the iterative design of the AI assistants. I didn't generate or notice this argument myself, so I'm leaving it out of the analysis, but it's good one, I think.

A Simple Alignment Typology

I set out to review the OpenAI alignment plan, and my brain at some point diverged to modeling the humans behind the arguments instead of the actual arguments.

So behold! A simplified, first-pass Alignment Typology.

Why can't we all just get agree?

There are a lot of disagreements in AI alignment. Some people don't see the problem, some think we'll be fine, some think we're doomed, and then different clusters of people have different ideas on how we should go about solving alignment. Thus I tried to sketch out my understanding of the key differences between the largest clusters of views on AI alignment. What emerged are roughly five cluster, sorted in order of optimism about the fate of humanity: the sceptics, the humanists, the empiricists, the rationalists, and the fatalists.

Sceptics don't expect AGI to show up in any relevant time frame.

Humanists think humanity will prevail fairly easily through coordination around alignment or just solving the problem directly.

Empiricists think the problem is hard, AGI will show up soon, and if we want to have any hope of solving it, then we need to iterate and take some necessary risk by making progress in capabilities while we go.

Rationalists think the problem is hard, AGI will show up soon, and we need to figure out as much as we can before making any capabilities progress.

Fatalists think we are doomed and we shouldn't even try (though some are quite happy about it).

Here is a table.

	Sceptics	**Humanists**	**Empiricists**	**Theorists**	**Fatalists**
Alignment Difficulty	-	One of these	high	high	-
Coordination Difficulty	-	is low	high	high	-
Distance to AGI	high	-	low/med	low/med	-
Closeness to AGI required to Solve Alignment	-	-	high	med/high	-

Closeness to AGI resulting in unacceptable danger	-	-	med/high	high	-
Alignment Necessary or Possible	-	high	high	high	low

Less Wrong is mostly populated by empiricists and rationalists. They agree alignment is a problem that can and should be solved. The key disagreement is on the methodology. While empiricists lean more heavily on gathering data and iterating solutions, rationalists lean more heavily toward discovering theories and proofs to lower risk from AGI (and some people are a mix of the two). Just by shifting the weights of risk/reward on iteration and moving forward, you get two opposite approaches to doing alignment work.

How is this useful?

Personally it helps me quickly get an idea of what clusters people are in, and understanding the likely arguments for their conclusions. However, a counterargument can be made that this just feeds into stereotyping and creating schisms, and I can't be sure that's untrue.

What do you think?