

Amplified Wise Simulations for Safe Training and Deployment

Chris Leong

While this proposal has weaknesses, please evaluate it comparatively, and judge it in relation to the weakness and requirements of other plausible alignment strategies.

1. Problem overview

"The saddest aspect of life right now is that science gathers knowledge faster than society gathers wisdom" - Isaac Asimov

The Importance of Wisdom: The AI industry is focused on training highly intelligent agents. Unfortunately, this may lead to bad outcomes if we accidentally produce rogue agents or if capabilities proliferate and the offense-defense balance favours the attacker. This approach increases the influence of unwise actors who can train and deploy models faster by avoiding thinking about or mitigating any potential consequences.

Concrete sub-problem: The future is much more likely to go well if we can produce AI advisors that allow us to make better decisions about the safe development and deployment of AI. This will be the initial focus of this project as even if building systems that directly help here isn't viable, it will allow us to better understand the problem.

Bigger picture: We may later pivot or expand the project to cover advisors on topics like: making high-level strategic decisions, understanding perspectives around AI, non-manipulative communication and persuasion, negotiation and conflict resolution and navigating emotions and biases. Sharing these technologies with responsible actors may be a key part of a [pivotal process](#) and these technologies could assist with almost any alignment plan.

2. Your solution

Wise AI advisors: I want to train simulations of wise folk who have been trained to use a minimally manipulative LLM by using imitation learning. Next, I want to amplify these agents using techniques like RAG, [debate](#), [iterated amplification](#).

What will we ask them? These advisors wouldn't tell you what to do, but would instead help guide you towards making up your own mind. While we could load relevant facts into context via RAG, human advisors are often able to assist people in thinking through issues without deeply understanding all of the specifics.

What do you mean by wisdom? An exact definition is unnecessary given that the idea is to train multiple sims that are wise according to different notions of wisdom, however, it is useful to contrast intelligence to wisdom. Someone who is intelligent might come up with a brilliant strategy for winning a war, but someone who is wise

might realise that the war could have been avoided in the first place. Someone who is intelligent might invent a superweapon, but someone who is wise might realise that it threatens civilization as a whole.

Why not just use a decision framework? Hopefully, these agents will be more flexible than a framework and be able to help you figure out when you're trying to apply a framework in an invalid context or how to resolve conflicting principles.

Why use amplified imitation learning? Attempting to train directly on wisdom suffers from the usual problems of the optimisation algorithm adversarially leveraging your blind spots, but worse because wisdom is an especially fuzzy concept.

Attempting to understand wisdom from a principled approach and build wise AI directly would require at least 50 years and iteration through multiple paradigms of research.

In contrast, imitation learning provides a target we can optimize hard on.

How would these agents be trained? We would fine-tune a base model to produce different advisors. We might have multiple levels of fine-tuning, ie. fine-tuning on the data from all the wise people we are using, then fine-tuning the specific advisors.

We would want to find wise folk who would be open to collaborating on this project. Ideally, they'd have a large corpus of existing work which we could fine-tune on, then we could supplement this by requesting responses to specific questions. We could use out-of-distribution detection to determine when responses are less likely to be accurate and to figure out what additional data we'd need to collect.

Minimally manipulative LLM: An LLM where the base model is limited to a certain strength with RLHF used to reduce manipulation. Normal LLM's optimise for engagement due to commercial realities. A non-manipulative LLM would be trained to avoid certain emotional arguments or loaded words.

Ambitious version of this project: While the initial scope is narrow, in the best case, humans assisted by amplified advisors would form a cybernetic system that could undergo a "wisdom explosion" where advances in wisdom are used to unlock further wisdom and train wiser advisors. Even if this capped out at peak human wisdom, this might still be sufficient to enable a successful pivotal process.

3. Pilot experiment

For this hackathon, I focused on a field study approach as opposed to a formalised experiment¹. My intuition is that in the early stages of a project, it's better to run informal tests to determine whether your assumptions are roughly in the right ballpark. Formal experiments are unlikely to be valuable until that has been pinned down.

¹ I took [Studying the Alien Mind](#) as inspiration

One leading contender to my proposal for training Wise AI would be to just use a regular LLM. I choose to test Anthropic’s Claude, since it is more flexible than ChatGPT. Fine-tuning a model to be wise would be more work than fine-tuning, so I decided to focus on the latter for this quick experiment and to see what the limitations were to better clarify my understanding of where sims could add value.

I used Claude’s artifacts feature to help me incrementally create three documents: a persona, a guide on providing wise advice and a more general advisor’s guide to cover auxiliary issues in providing AI ([see Github](#)).

From the process of generating this, Claude seemed much wiser than I expected. The proposed principles were highly quality and less handwavey. I also realised that I hadn’t previously fully taken into account the possibility of using AI to help figure out how to create a wise AI, since I was able to iterate on these artifacts extremely fast.

Engaging in this exploration made me realise that in order to be competitive these Wise AI agents would have to fit well into a workflow involving both humans and a more usual LLM. I hadn’t realised quite how capable Claude was before, since this was my first time utilising the artifacts feature to iterate on a document. This would allow the generation of a draft decision framework extremely quickly.

I said “more usual” above because I would want to train an LLM to be especially non-manipulative as this is the only way I could see this going well. I realised that training a non-manipulative LLM could be a separate start-up idea.

Unfortunately, I ran out of Claude tokens, so I had to test my documents on ChatGPT. I tried asking it to advise me on whether transformative AI should be OpenAI. ChatGPT is pretty good at listing various benefits and downsides, but some considerations are more or less decision-relevant. I tried asking it about the scenarios in which open-sourcing transformative AI would be good and when it would be bad. It really struggled to understand what I meant by a scenario, even after I tried giving it multiple hints, and this made me somewhat pessimistic about this working. At the same time, it’s too early for me to give up, without further prompt engineering.

One update I made from this experiment is that these Wise AI advisors should be able to send a specific query to a normal LLM and highlight the considerations that they believe to be most relevant as this is one area where normal LLM’s struggle.

4. Process

Timeframe	What will you do?
Next 3 months	Simultaneously test both the underlying theoretical assumptions and run cheap experiments with demo systems. There is no need to train sims until we understand the limitations of a custom-prompted LLM.
2025	Attempt to find some initial “customers”, really just qualified people who will use your product and provide feedback.

	Main workstream: Quickly test a bunch of sims of people with existing corpuses (with appropriate permissions). We can use long-context windows with prompt caching for initial tests to avoid fine-tuning for initial tests. Start collaborating with a few wise people to gain experience with the data collection process for training more accurate sims. Lower priority workstreams: Improving sim accuracy, amplifying sims, using sims to assist in running the company
2026	Produce an MVP with at least one sim that at least a few organisations are willing to pay for, even if at a significant discount, mostly as a way to ensure that we are delivering value. Start adding more sims and experimenting with amplification.
2027	Produce a tier that organisations are willing to pay a significant price for, again mostly to ensure that we are delivering value. Refine the product until it is rock-solid. The highest tier may be restricted to non-profits or customers who choose to become investors.

5. Impact on AI safety & key risks

The company would aim to be selective in who its customers are.

Any technology developed will diffuse, but the diffusion of wise AI is less risky than intelligent AI, because foolish actors are less likely to appreciate wisdom. If they somehow do start making use of such systems, there's a decent chance that they become wise enough to stop making foolish decisions that put society at risk.

6. FAQ

Who counts as a 'wise figure worth training on? We'll be using our subjective judgment. This is less limiting than it seems as different actors can use different subsets of figures. If our judgment is flawed, hopefully, these figures will help us realize this.

Why the focus on inspiring people to make wise decisions rather than making them directly? I suspect that inspiring people to be wise can be achieved at a significantly lower level of capabilities than required for producing good object-level advice, which is important as we want wisdom sooner rather than later.

Mightn't the AI pursue the goals of the people you train it on? Unclear, as it's trained to reproduce what they would say, not what would they do. Nonetheless, trusting the wrong people will doom you whether it's humans or AI.