AISafety.com – A Startup for Aligning Narrowly Superhuman Models

By Søren Elverlin, 2021-08-31

I intend to create a startup based on **aligning large language models with business goals.** I will personally[1] create and fund a framework for using existing foundational models as a copilot for business processes. The hope is that alignment work on this model will also provide business value, allowing a virtuous cycle by funding more alignment work and bringing in more research talent.

This document ("AISafety.com.docx") argues why this work is Important, Neglected and (to some extent) Tractable. A full analysis of Tractability also depends on the documents "Business Plan.docx", "Specification.docx", "Budget.docx", "Legal.docx", "Failure Modes.docx", "What Winning Looks Like.docx". These documents are work in progress, and like most startups, failure is a possibility.

Motivation

In The Case For Aligning Narrow Superhuman Models, Ajeya Cotra argues for an alignment strategy based around finding general alignment techniques for models in fuzzy domains.

I see a weak consensus in AI Safety that this is among our better options for eventually solving AI Alignment, though that goal seems to be far enough away that we are unlikely to reach a stronger consensus. Some prominent researchers do think there are more valuable things to do, but few expect it to be harmful.

Ajeya Cotra suggests that this should in the short term be pursued by Principal Investigators at universities, as well as Senior ML Researchers who are free to do their own projects. As far as I can tell, this has not worked out yet. Open Phil is empathetically not soliciting project proposals in this space. (EDIT: This might change in the future. As of 2021-08-20, they have an RFC for RFPs.)

All valuable work on Alignment should be done, and I am convinced that this is important, neglected and tractable. In contrast to other AI Alignment strategies, aligning narrowly superhuman models has the potential to be very commercially valuable. This could potentially allow alignment research to be self-funding.

The Startup

The tentative name is AISafety.com. The term "AI Safety" is controversial, and often used much more broadly than long-term alignment. I intend to hire key researchers from "Inside Alignment Research" but mostly create a supporting organization from "Outside Alignment but Inside AI Safety".

It is often repeated that there is a significant oversupply of altruistic funding available for alignment research. This startup aims to create a large, professional organization, with capabilities far beyond those provided by ephemeral Mechanical Turk sometimes used in the more concrete alignment research. Altruistic funding will be an advantage, but not a requirement, and the project is still viable with any degree of funding.

[1] I am married, and the funds are communal.

Vignettes from the future

One way of illustrating what kind of work AISafety.,com could do is to give examples of what concrete work can be done, how the concreteness of the tasks makes it fundamentally different from existing alignment work, and how this work can be done by people who would otherwise not contribute to solving the Alignment Problem:

- Alice wrote her thesis on ensuring self-driving cars do not hit emergency vehicles. She is currently working in AISafety.com on determining if their language model is withholding information.
- Bob has a master's degree in Business Strategy and is currently analyzing problematic outputs from the language model from AISafety.com, with an emphasis on discovering and categorizing treachery.
- Carol left her job at Facebook, where she was building the largest model yet. She is now grafting new neurons to a much smaller model at AISafety.com, hopefully corresponding to a 100% (=immutable) credence that the model has a decisive strategic advantage.
- Dave serves on the Permanent Data Quality Team at AISafety.com. He holds no formal education, but the alignment researchers appreciate how his dedication support their work.
- Eve is currently working on integrating Chris Olah's "Microscope" into the models at AISafety.com. The work is open-source and quite general. The intention is to make it trivial to add to all models in the world.
- Frank is currently working on foundational research, a direction he narrowly chose over high frequency trading. Although paid by AISafety.com, his work is "one level of abstraction up", and is more useful to MIRI. He won't leave as he prefers the culture of AISafety.com.
- Grace comes from a background in hospitality and is currently arranging conferences and general community-building. There are currently no public out-reach positions in AISafety.com, but she is slowly and carefully assembling a team for this purpose.
- Heidi started out working on Alignment at AISafety.com but didn't like it. She is currently implementing a honey-pot with tripwires as part of the AI Control Team.
- Ivan has been prompt-hacking language models since 2020 and is the author of "5% More Aligned. The AISafety.com Handbook for Prompt-hacking".
- Judy is an expert in business processes, and has specialized in one task: Given a business process and output from a language model attempting to implement this process, judging to what extent the output is aligned with the business processes.
- Mallory is a former penetration-tester, and leads the red team at AISafety.com. This work focus more on practical steps the language model could take to e.g. hide data and obtain ressources, rather than focusing on how other humans can misuse the models.
- Niaj is the oldest employee and has served as CEO in 3 companies. He is managing the upper part of the main sandwiching project, gaining a deep understanding of the objectives of top management and how they diverge from the people on the floor.

Ownership Structure

The goal is to solve the Alignment Problem, not profit-maximizing. This creates some unique challenges for the structure:

- It must be flexible enough to handle ontological shifts (e.g. "Friendly AI" -> "Aligned AI", or OpenAI realizing that openness is problematic).
- We must avoid losing the purpose (arguably, OpenAI doing capability research).
- We must be able to attract employees and investors from outside EA.

- We need to credibly signal that we care about AI Alignment for altruistic purposes.

A standard startup equity structure would be inappropriate for an altruistic community-focus company. My current plan[2] is to divide equity into 3 parts of the same size:

1) The first third is equity to be **sold off to investors** to accelerate growth. If feasible, I prefer this to be non-voting, and the goal of solving alignment explicitly overrides any fiduciary responsibility towards stockholders.
2) The second third is equity **used as compensation to early employees**. Y Combinator suggests settings aside 10-20% for this, so at 33%, the amount of equity offered would be roughly double the startup standards.
3) The last third is equity that **vests once the Alignment Problem is solved**. This is assigned by counterfactually estimating the impact of everyone's work in the company. This equity can be traded freely before it vests. (TODO: Ask John Burden about how to do this in a credible way). This costly signal is meant to differentially attract people who care about the mission.

Even though I will maintain a majority of the stock, this still feels sufficiently altruistic to me that I believe we can call ourselves a n altruistic community startup, and at the same time should motivate potential employees who care about compensation and fairness.

There is a legal term "Public Benefit Corporation", which seems like a low bar to pass – but legal advice is required to determine if this is appropriate.

Culture

The culture is crucial for any startup. I obviously intend to create the most awesome, rational and productive culture I can, but so do all other startups. There are special considerations for this startup in particular:

There is a built-in tension between Alignment work and Capability work. The business-side might reasonably argue that we should use larger models, and we might need to ensure that the alignment side mostly prevails. OpenAI is perhaps an instructive example, on the negative side. We might need to hire a good CSO, who might be better than us at playing the Game of Thrones. This creates a principal agent problem, but this will be manageable since I control the votes of the shares that have not vested yet.

The hiring process will prefer candidates with demonstrated previous altruistic contributions to AI Alignment, though not to the point of poaching researchers working on more important work. Engagement with the Rationalist community would probably also be preferred. I expect this source of manpower to be insufficient, and non-strategic positions would likely have to be filled with people who care about alignment because they are paid to care.

Currently, I am leaning towards having all alignment work open source, as this is consistent with our goal of having alignment methods being widely used. This is the kind of decision I could see being changed in the future – if licensing could bring sufficient resources towards alignment work, that might be preferable.

[2] This is not legally binding! I spoke to a non-expert, who said the tax might need to be paid in advance, which would invalidate this idea.

The entire point of AISafety.com is to gain experience aligning large models. This means that the process must be well documented. We will keep diaries, and post lessons learned on LessWrong. It is hoped that some research should be formal enough for eventual publication – pre-registering hypothesis, statistics etc.

Importance

Per Ajeya Cotra's case, this work is likely very beneficial. I strongly recommend reading it.

My mental model of "alignment" is multi-dimensional, and not categorical. As an example, my model predicts we will eventually be able to say things like "GPT-3 is 1% more aligned if you ask it to be nice in the prompt". Towards the higher end of the scale of alignment, I predict we will find an "attractor basin" where almost-aligned AIs want to be fully aligned (this is a counter-argument to the fragility of values).

Is there any way this project could increase existential risk from AI?

- A failure mode would be creating impressive demos. We could discover that just prompt-hacking is enough to be commercially valuable, and alignment isn't necessary. Even if this is true, I think this is strictly better that this is discovered by "us".
- Worst-case would be creating a rigged demo, which I guess I'll just not do.
- The status of the field of Alignment could be lowered by e.g. seeking publicity and then failing. Publicity-seeking is not a part of my plans. I don't think the early failure of a startup would be particularly newsworthy to anyone outside the field.

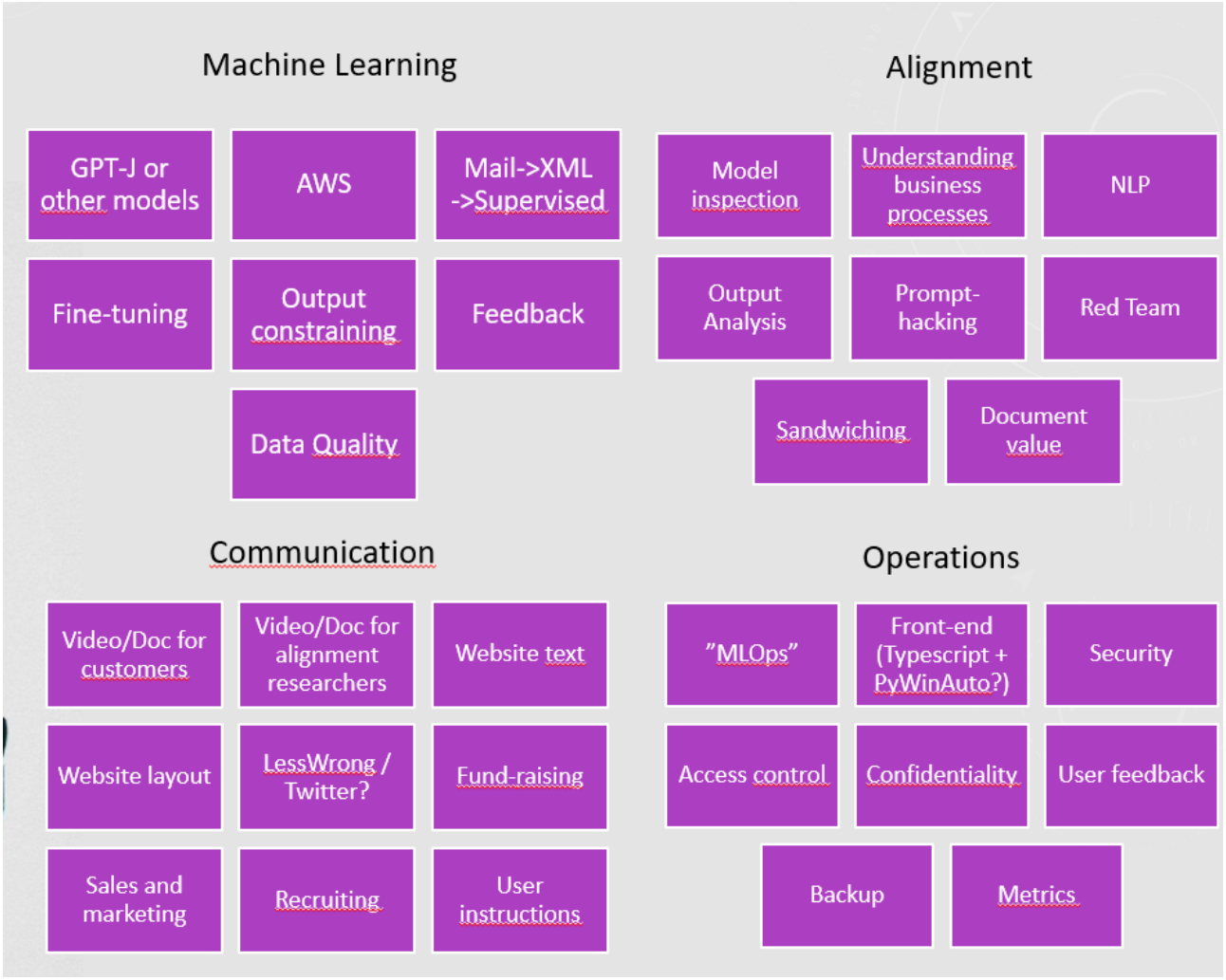
I've thought enough about this now, and I believe this is strongly positive. This does not mean that I am sure we can solve the Alignment Problem, or that I think it is solvable.

Initial roles

I am a generalist and have a sufficient grasp of all the tasks involved to be able to do this alone, though this will be inefficient and slow, so I will attempt to bring in contractors, employees, and partners to cover my weaker areas.

- Initially, I will be CEO, and handle sales, development, operations, devOps, MLOps and everything not explicitly covered by others.
- Finetuning/Supervised learning is a necessary task that I have only theoretical knowledge about. There is a certain metis to this, and this is a well-delineated area that should be covered.
- I need to find someone to interface with the wider alignment community. E.g., posting on LessWrong, discussing with other alignment researchers, "meta-research". There is probably enough work in communication to necessitate a full-time position.

- There should be a dedicated position for the task of actually aligning the model – a CAO.



Others doing the same

Both Anthropic.com and Ought.org are conceptually close to my vision for AISafety.com, but there are important differences:

The customers seem distinct: Anthropic is building tools for making models more interpretable - I guess their customers are mostly corporations with large models, and very research-based (as opposed to practical). Oughts Elicit seem geared towards "researchers" in a very broad sense of the word. This seem distinct from AISafety.com's target customer-segment: "Businesses".

The focus on Alignment: Anthropic curiously does not mention the word "Alignment" on their website, and their research seem one step removed from alignment (though I do think this is valuable towards solving alignment). Ought do mention alignment a bit, but seem far more interested in capability work.

Funding and researchers: Right now, both are funded by EA / EA-adjacent money. It is unclear to what extent this will continue, but a significant goal of AISafety.com is to not compete for these funds and attract researchers from outside the larger Rationalisphere.

Assuming all alignment research is open, I expect synergy, rather than competition. We would ideally be customers for Anthropic.

"Using ML for business processes" is something that a lot of other companies are trying. I think it is better if I capture the market first. Even broader, "A startup doing NLP" would probably describe hundreds of companies, which might pivot to business processes if it turned out to be lucrative.

Aligning a ML model for Business Processes

The success of this startup depends on many practical factors, and one essential assumption:
Current or Next-generation Large Language Models are superhuman in executing a large subset of current business processes but are held back by insufficient alignment.

Some intuition behind this prediction:

Small-a alignment is a major concern for large companies. As the organization and processes grow more complex, lost purposes, goodharting and other factors contribute to increasing inefficieny, waste and calcification. The principal agent problem is very real.

Contrary to the text on this cup, major successful companies hold meetings to ensure alignment between the meeting participants, and face-to-face meetings contribute to this, even if the surface-level agenda of the meeting is easy to communicate.

If a large company has a significant business process that doesn't involve human judgement, it has already been automated. This automation process can be seen as a slowly rising waterline, but this is fundamentally limited by number and complexity of processes, interacting with the requirement for human judgement. **Many processes only need human oversight for exceptional cases and are actually quite simple.**

Formal descriptions of business processes are insufficient. In general, these are We arguably going one step further towards alignment: We want to align the language model with the business – i.e., maximizing shareholder value. This is not what we truly desire as a final goal, but this a subgoal on the path towards solving alignment. Communicating this point clearly is a priority.

Naively optimize towards making suggestions that the person accept won't work. For most business processes, we don't have access to the actual business processes, but only their implementation by people in the business. Though we do measure acceptance, we try to align towards the business processes themselves, not the person.

What we observe

What we want to learn

Examples:

- A narrow superintelligence could always find a reason why a superior need to look into a case, and this is undesirable.

- It is in the interest of the employee that information is hoarded, but this is not in the interest of the company.

Language is a central part of business processes. The non-formalizable processes involve communication, and a capable language model is an absolute prerequisite for implementing these business processes. On the other hand, the processes outside of language are often rather trivial.

The inherent difficulty of a business process can often quickly be estimated by humans. An experienced person will often decide what the next step is while skimming output of the previous step.

Practical questions

The name "AISafety.com" is not set in stone. It is primarily a marketing tool/brand towards ourselves and potential Alignment researchers. From a business perspective, this might not be optimal (but unlikely to matter much). If we are very lucky, we'll be able to argue towards prospective customers that alignment research falls under Corporate Social Responsibility. There is a theoretical possibility that it will be less biased, but this is not something we should promise, nor is it a particular focus of ours. Our work **is** broader than pure alignment - Some of the work we will do falls under "Foundational Research" and "AI Control".

Advisory board – is one needed? Matthijs Maas expressed interest, but it is unclear if he is allowed. It is also unclear if this kind of work is compensated.

People to talk to

Stuart Armstrong, Paul Christiano, Connor Leahy