Strategic Vision for AIS Group Support

Agustín Covarrubias | 20 Dec 2023 | Public

This is an early document planning the execution of the AI Safety Group Support project at CEA. Some light updates have been made since it was written to reflect some subsequent changes in the project's hosting organization, but it does not include anything from February 2024 onwards.

Introduction

In the last few years, the AI Safety talent pipeline has grown from just a handful of university groups, university labs, and organizations in the area, to a large, decentralized community of organizations trying to rapidly build the capacity needed to substantially scale the width and quality of the talent pipeline, especially given the extreme growth of public interest in AI Safety and the fast reduction of the stigma that used to be associated with the field in the eyes of many people in government, academia, and the public.

To date, there are now dozens of active AI Safety groups and dozens of organizations working in the fieldbuilding space, covering a spectrum of functions in the pipeline, including outreach, training, career advising, funding, in-field prioritization and strategy. Many of these organizations have started performing some of these functions in an ad-hoc capacity, trying to rapidly fill gaps in the pipeline as they arise.

In the particular case of the early talent pipeline (the segment of the pipeline that captures people interested in working in AI Safety through things like university groups, workshops, and research fellowships), there is currently no single organization coordinating initiatives in the area, even though these seem to have an established track record of being among the most effective projects in the space. Due to the lack of established group support initiatives, CEA has switched to performing some of these functions from its Groups team, which until recently, was almost fully specialized in EA groups, only providing ancillary support to cause-specific groups.

Currently, CEA provides support to AI Safety university groups through programs like the Organizer Support Program (OSP), which, for the last two semesters, has piloted supporting AI Safety organizers by connecting them with experienced mentors that can guide them in organizing a group, as well as by organizing events (like the recent group organizer summit), which bring together community builders to meet each other, discuss strategic considerations, skill up and increase their motivation to do community building.

CEA has decided to take on a new specialized role on its Groups team, an AI Safety University Group Support Lead, which will essentially incubate a new project, independent of CEA, fully focused on providing non-monetary support to AI Safety groups. The plan is

that this role will help pilot new infrastructure in the area while planning to spin off this infrastructure into an independent organization over the next six months.

This proposal provides my strategic vision for the role and its functions for the period before this spin-off, as well as giving a glimpse at how its functions could change in the future.

Vision and Mission

This role is expected to take on two main hats:

- Building and improving the infrastructure supporting the early-stage talent pipeline, in particular for university groups (the primary role)
- Helping coordinate different organizations and stakeholders across the pipeline, identifying gaps and bottlenecks in it, and informing fieldbuilding strategy (a secondary role)

Building infrastructure for the early-stage talent pipeline

The biggest gap this role hopes to address is the lack of specialized support to AI Safety university groups, especially given the pace of changing strategic considerations, the need for resources that attend to these specific considerations, and the early evidence pointing to fieldbuilding in this space being particularly cost-effective compared to other interventions.

On one hand, I expect some of this support to look a lot like the existing support being provided to EA groups by CEA, and indeed, some of this is happening already through programs like OSP, or recent events like the OASIS[1] or UGOR[2] organizer retreats. However, extensive resources targeting AIS groups specifically are still lacking, and programs like OSP are still piloting the best ways of supporting AI Safety groups, while still having a public identity that's heavily focused on EA groups. For this reason, I expect this role to spend most of its time, especially over the first few months, trying to create the necessary infrastructure and resources needed to bring AI Safety group support to "feature parity" with EA group support.

On the other hand, I expect that some of this support will look very different from the one currently provided to EA groups, for example, focusing on the unique challenges posed by the need for technical training, very different communication strategies, and rapidly changing needs and opportunities in the ecosystem. A major challenge for this role will be developing, testing, and refining these new resources, like new types of programming, new communication strategies or events, working closely with established groups to figure out what works and what doesn't, instead of drawing only from the existing experience with EA groups.

---

[1] OASIS was a small retreat run in Berkeley during early February of this year, which was focused on helping top AI Safety group organizers plan out their semester activities. This was run by Constellation.

[2] CEA has run different kinds of organizer retreats over the last few years. The latest one was the creatively named University Group Organizer Retreat (UGOR), which happened a few months ago, and brought top EA and AIS organizers together in the UK.

Helping to coordinate work across the talent pipeline

While in the short term, I expect this role to be focused almost exclusively on the infrastructure and capacity needed to support university groups, I think this role could be well positioned to also work closely with organizations all across the AIS pipeline (from university groups to established labs) in order to help identify bottlenecks and gaps in the pipeline (for example, missing infrastructure for training, the need for more specialized career advising or not enough people with a background in information security) and helping to orchestrate projects to quickly fill these gaps, both internally and externally.

Currently, this function is spread among several organizations, like Constellation, FAR AI, MATS, CBAI, LISA, Center for AI Safety, BlueDot Impact, Rethink Priorities, Alignment Ecosystem Development (AED), among many others. Based on conversations with some of these relevant stakeholders (particularly MATS and Constellation), it seems it would be unwise for this role to focus too much on general coordination, considering that some organizations are already doing this pretty well. At present, the main axis of necessary coordination seems to be improving resource sharing and coordination between already established university groups, but I expect this demand may change depending on future assessments of the overall state and needs of the talent pipeline, which should follow naturally from performing the stakeholder engagement performed below. For this reason, I've mostly focused on infrastructure and emphasized less on this aspect of the role.

Key uncertainties

Main document: Key Strategic Uncertainties | AIS Group Support

Short-term projects

AIS Resource Centre

The most straightforward short-term project this role should focus on is on trying to improve the state of resources targeted at new AI Safety UG organizers, and it seems like a good model of what this could look like is the EA Groups Resource Centre. I expect developing a resource centre with anything near to resource parity to the EA resource centre would take substantial amounts of time, but it's possible this role (performed at 1-1.5 FTE if a second person joins the project, like Nikola) could develop an MVP centre relatively quick, addressing most of the low-hanging fruits for new organizers.

A major rate-limiting factor will be tweaking some advice as we get feedback, especially from some of the new group organizers (like the ones currently in OSP) about the types of resources that would be useful, but it seems there is already plenty of value in just centralizing some of the existing community resources, and particularly in following the pointers for improvement already laid out in Nikola Jurkovich's AI Safety University Group

Guide, and combining that with some adapted resources from the EA Resource Centre, as well as references to relevant projects by the Alignment Ecosystem Development (AED) project.

I expect that under that this minimal version of this project could be set up in approximately two months with 0.3 FTEs of work from one person. This project could then be expanded with new resources based on ongoing feedback from established and new organizers.

Fieldbuilder Support Program (FSP)

Starting next semester, I think the best path for continuing to provide equivalent to the Organizer Support Program (OSP) would be to take learnings and feedback from the last two rounds of running the pilot AIS program, and spinning this off into a separate project, which I've affectionately named the Fieldbuilder Support Program (FSP).

Regardless of whether this role is initially housed at CEA, this program will be run at a time when we already expect to have transitioned to another organization, so the key challenges of running this spinoff would be:

- Replicating the key workflows already in place for OSP (including outreach, applications, mentor-mentee assignments, supervision, and assessment).
- Building a new network of mentors, and trying to transfer over the ones already at OSP.
- Figuring out key changes in strategy, particularly around outreach and marketing of the program outside of EA and existing AI Safety UGs.

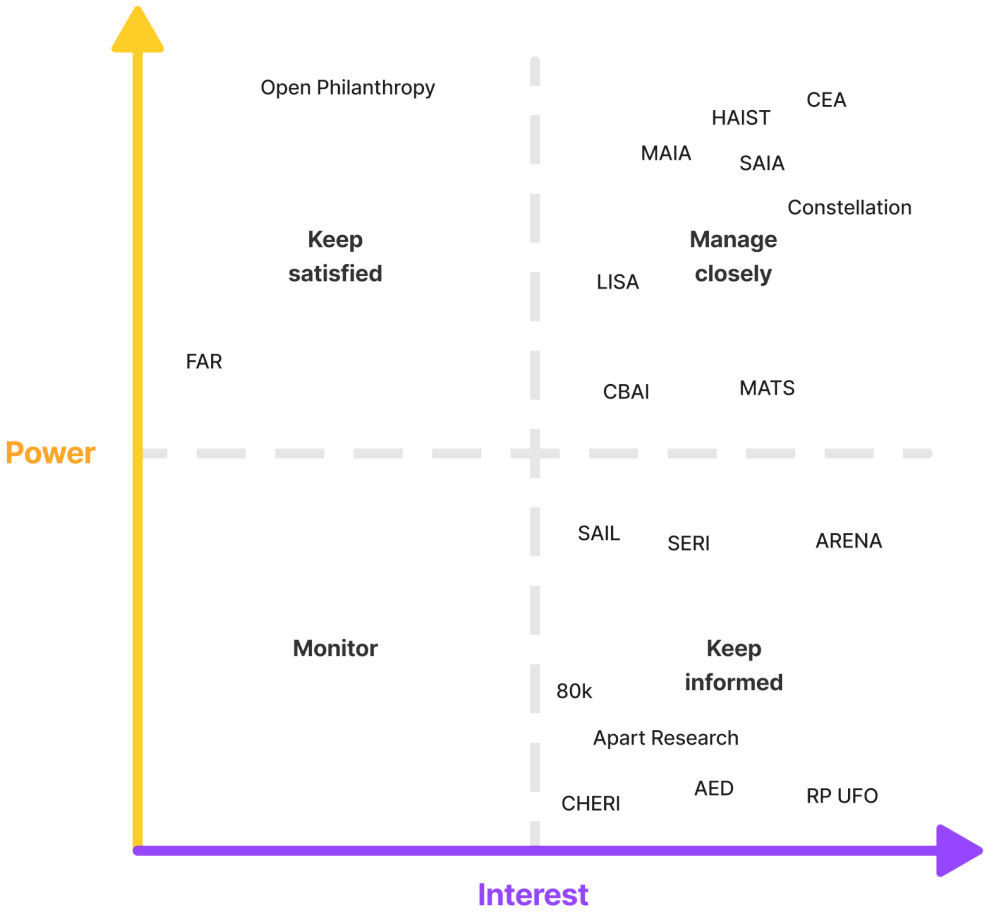
The details of this will probably depend a lot on the capacity of the organization hosting this role, but I think most of the work over the next 6 months would consist on planning around the first semester of FSP and thinking through all the major changes to the program.

One thing I would be interested in experimenting with is trying to align FSP with opportunities for technical upskilling (especially in ML). This could plausibly be implemented by either connecting the organizers with existing opportunities for upskilling, or organizing some ad-hoc upskilling program within every FSP cohort.

Information gathering and stakeholder engagement

I think a key factor in whether this role succeeds or not will be whether it manages to successfully engage with many relevant stakeholders of the AI Safety ecosystem, starting with the established university groups but also including everything from upskilling programs to general organizations working in the AI Safety ecosystem.

I think the first weeks of this role should be focused on creating a detailed stakeholder engagement plan, detailing all relevant stakeholders and a plan for dealing with each of them (if necessary) over the next few months. For example, a very incomplete but still somewhat useful power-interest matrix for this could look like:



Accordingly, a basic MVP of this plan could look something like:

- The first month should be spent on the ground working with HAIST, SAIA, and MAIA to try to figure out key strategic concerns, major learnings and trying to secure their buy-in for the role.
- Within the first three months, this role should have clear buy-in from LISA and MATS, have a good understanding with Constellation and Open Philanthropy, and open communication paths with Apart Research, ARENA, MLAB, AED, SERI and 80,000 Hours.
- Within the first 6 months, this role should at least have a somewhat complete network established with people at OpenAI, DeepMind, Anthropic, FLI, PauseAI, CHAI, EffiSciences, CHERI, GCP, SAIL, etc.

This is referential, but I expect defining a plan like this should be a high (if not the highest) priority of this role during its first few months.

Public-facing fieldbuilding network

Especially if this role is to capture value outside the EA community, I think building a separate public-facing identity for the functions and programs performed by this role is important. This essentially involves creating a separate, consistent “brand” for the programs

run by this role and making it well-known inside the AI Safety ecosystem. This is probably key for this role to function effectively as a focal point for university group organizing, as well as attract potential organizers to relevant programs (like FSP). Building this identity will be a key step in spinning out of CEA.

This could be as simple as figuring out a name, creating a logo, mounting a basic website, and publishing forum posts about it, but it could also fit within a wider outreach strategy. This depends on key considerations around how to best reach potential organizers, especially if we expect a significant amount of them to be outside the traditional EA networks, but it might turn out that this is less critical, especially depending on how high we want the standard for new organizers to be. If outside outreach is less essential, the focus should be put on stakeholder engagement (as addressed previously).

Supporting fieldbuilding prioritization

While some research has been done on trying to model the impact and cost-effectiveness of AI Safety field-building programs, the most detailed models so far (particularly those done by the Center for AI Safety, as well as some models internal to Rethink Priorities) are still pretty simplistic, and I have a strong suspicion that they might not adequately capture some key dynamics of AI Safety university groups, especially regarding spillover effects, that could not only prove significant for fieldbuilding prioritization, but also determine key aspects of university group strategy going forward, especially for new groups. For example, I think it would be very helpful to develop moderately detailed models that can help us get a grasp of the effectiveness of outreach to professors and graduate students, which could, in turn, inform some of the strategy resources that this role would develop.

Based on my previous experience doing CE modeling at Rethink Priorities, I suspect this is something that could be reasonably performed with less than 0.3 FTE over less than two months. For this reason, I think it could be a reasonable low-hanging fruit for this role to try to tackle.

Possible future directions

Most of the projects and vision I've covered here relate to roughly the next six months of this role, but it's also worth pointing out some possible future directions, especially if the initial attempts to build the relevant capacity are successful:

- **More active stewardship of the AIS Talent pipeline:** While this would imply a much broader scope, this role could be well-placed to improve coordination along the pipeline about talent needs, bottlenecks, and gaps, and try to leverage the existing capacity to either address these directly, or working with other organizations to attempt to mobilize projects addressing these areas.
- **More tightly integrated AIS programming:** Current introductory programming for AI Safety (like AISF) has been developed to target a very broad target audience. While many of its "forks" have tried to address shortcomings of these problems specifically

as it related to running them in university groups, I think there remains a lot of low-hanging fruits for high leverage programming that we haven't even started considering yet, especially around more active programs and workshops instead of reading-focused fellowships. Developing and coordinating experiments around new programming of this kind could become a key function of this role (or of its surrounding organization).