This is a template for project proposals for AI Safety Camp.

It's ok to submit a proposal that is not yet finished. If you apply in time, we'll help you improve it.

Click "Share" in the upper right corner to turn on sharing before sending us your document.
We recommend ***giving commenting access to "Anyone with the link"***, so that we can share your draft with trusted advisors. However, if you want more control, you can instead just give access to linda.linsefors@gmail.com, remmeltellenis@gmail.com and robertkralisch@gmail.com.

Please name your document "[Your name] - [The name of your project idea]"

Examples
To give *you an idea of what and how to write, here are four accepted projects from AISC9:*

- Lawyers (and coders) for restricting AI data laundering
- TinyEvals: How Language Models Speak Coherent English?
- Towards Ambitious Mechanistic Interpretability
- Modelling Trajectories of Language Models

If you are accepted to lead a project, then the final version of your project plan will be posted publicly on the AISC webpage at the start of November. But don't worry. You will have time to revise it before then.

Target audience
Write the application with a potential applicant in mind.

[The name of your project idea]

Summary

A short description of your project. Just a few paragraphs to help any reader to get an overview of what you want to do, and to decide if they want to read more.

When promoting the project, we'll sometimes post the summary together with a link to this document.

The summary should be 50-200 words.

The non-summary

A longer description, including anything you think is relevant. This should include the motivation for the project and roughly what steps are involved.

If you are unsure what to write, here's some questions to think about:

- Theory of change: If the project succeeds, how would this be useful for reducing risk from AGI/TAI?
- What are assumptions in terms of how AGI and/or human society would work under which the theory of change is tractable?
- Project plan: What are the steps you need to complete to finish this project?
  - What's the first step?
- Backup plan: What can go wrong, and what's the backup plan if that happens?
- Scope: What is and isn't part of the project?
  - What's the *most* ambitious version of this project?
  - What's the *least* ambitious version of this project?

This section is expected to be the majority of your document!

It is a good idea to divide the no-summary into subsections. Format it in whatever way makes sense for your project. If you don't know how to do this, look at these examples [links: TBD].

Output

Part of the format of AISC is that projects have a beginning and an end. At the end of the project, what will you have produced?

A blogpost? An academic paper? A github repo? A web-tool? Something else?
How will you share the outcome of your work to the world?

Risks and downsides (externalities)

Does your project have any risk or other potential downsides? I.e. what's the risk that your project turns out to be net negative for the world? E.g. infohazards, potential AI capabilities progress, etc.

It's important to be aware of any risk that comes with your project. However some projects will be much riskier than others, and some projects might not have any notable risks.

(This section is *not* about things like "step X was harder than we thought so we did not reach our goal". That would be part of planning, and goes in the non-summary. This part is about how your project might make the world worse rather than better. If the worst that can realistically happen, is that you don't do anything, then you don't have downside risks. This will be the case for some projects but not others.)

Acknowledgements

Who has contributed to this research proposal?
Is there any specific writing or person who has been a major influence?

Team

Team size
What team size are you aiming for?

Normal team size is 3-5 people including you, but you can go for bigger or smaller as long as it makes sense for your project.

Research Lead (You!)
Your name.

Your contact info *if* you want to make this information available to team member applicants.

Information about you which is relevant for the project, e.g. how does your background relate to this project.

How much time (e.g. average hours per week) do you commit to spend on this project if it happens?

Roles (optional)
Some projects come with well defined roles, others don't. If you have specific roles in mind, you can list them here.

If you want to recruit someone to do project admin or other support roles, you're wellcome to do so. Just remember that you have final responsibility to make sure your team runs well. E.g. If you delegate the job of scheduling meetings, and that person fails, you'll need to be ready to step in to pick up the slack. This goes for any task that is a bottleneck for the whole project.

Skill requirements
What skills are needed for this project?

If you are unsure what to write, here's some questions to think about:

- What minimum skills or understandings does any team member need to be able to contribute to this project?

- What diverse skills or backgrounds would you value having in your team, even if they are hard to find? Dream big: If you could get any person with any skills, what skills would they have?
- Are there any skills that are needed for this project that you don’t have yourself, and therefore need someone else to bring to the project?

If you have specific roles for your project, you might want to list skill requirements for each role.