Project planning: Iteration 2

*Planning out one’s projects explicitly can help you to be intentional about your work and what they are trying to achieve through it. It’s also a way of creating tighter feedback loops, both from others and yourself.*

*As always, the prompts here are intended to be helpful, and you should amend or ignore those that don’t feel right. If you would like to modify the template in any way, please feel free to edit or add questions as you see fit.*

### Name: Kay Kozaronek Date: 29.06.2022

# The project

### Describe the project in 2-5 sentences.

* Test “Training Language Models with Natural Language Feedback” on new OpenAI critique dataset
  + 100 training examples -> 13000 examples (OpenAI dataset)
  + If the algorithm incorporates human feedback on the new as well as it did on the original dataset, it’s a success. (e.g. shows generalizability of algorithm)
* So main goal is to advance outer alignment research.
* Publish results (as part of Jérémy’s next paper)

### What is the purpose of this project? Why do you want to do it?

* Get more familiar with state of the art LM (GPT-3)
* Advancing outer alignment methods

### What are the goals and subgoals of the project?

* Write my first paper
* Publish my first paper
* Build portfolio and network to find employment at AI safety organization

### How will you know if it’s worked?

*What would be different about the world if this project worked versus if it didn’t? Is there any way of measuring or verifying whether those changes happen?*

* Find employment afterwards / continue collaboration with mentor (Jérémy)
* Paper gets published on Arxiv by the end of September (1 month after project finishes)

# Approaching the project

### What are you going to do during this project period?

Consider setting milestones that are

* Specific,
* Measurable,
* Attainable by the end of this summer and in this environment,
* Relevant to your goals,
* Time-bound, i.e. come with a deadline.

Approach:

1. Set up weekly goals and make it public (accountability structure)
2. Set up Github repository for project
3. Read the necessary papers:
   1. Get familiar with OpenAI critique paper
   2. Get familiar with OpenAI learning to summarize from human feedback paper
   3. Jeremy paper
4. Replicate Jérémy’s paper with smaller model (e.g GPT-2)
5. Do Data Analysis of Critique dataset
6. Apply Jeremy algorithm to OpenAI dataset
7. Use OpenAI API to implement code from step 4
8. Iterate on results (e.g Hyperparameter Tuning)
9. Write up a draft of my results
10. Get Feedback
11. Finish draft
12. Publish brief summary on LW

### What are your biggest uncertainties about what you should do?

### How could you acquire information that would reduce these uncertainties?

### What failure modes are you worried about? (Imagine you failed; can you guess why?)

### How could you avoid or mitigate these failure modes?

### What can help you? Who can help you? (Who can already do this?)

### 