ML ALIGNMENT & THEORY SCHOLARS

Research Plan

MATS Program – Summer 2024 Cohort

Due: Friday 19 July 2024 by 6 pm PT

Task description

The first milestone of the in-person research phase of MATS, the Research Plan, is due on Friday, July 19 by 6 pm PT via **this submission form**. The ~2-5 page research plan consists of a description of your proposed research project during the MATS program and a justification of why you chose this project in consultation with your mentor. The research plan is intended to provide the MATS team and, potentially, an external grant manager with data on MATS research projects, as well as encourage scholars to critically evaluate their project's theory of change. We recommend that scholars exercise reasoning transparency in their research plan and aim for clear, concise communication rather than being exhaustive or unnecessarily technical.

The research plan should contain:

- A brief project description of no more than 120 characters;
- An abstract of no more than 1000 characters;
- A description of a plausible AGI threat model, or risk factor(s), that your project intends to address;
- A theory of change for how your project might plausibly contribute to reducing the risk of the chosen threat model/risk factor(s), the main failure modes you might encounter, and a justification for why you chose this project in particular;
- An outline of the research activities you plan to conduct during the MATS program, including any experiments and planned output (e.g., blog post, symposium talk, etc.);
- A brief outline of potential future research activities post-MATS given you successfully accomplish the research activities outlined in the previous section.

We estimate that the research plan should take 5-8 hours, though some scholars might choose to hone their research plan for significantly longer. In your research plan it is acceptable to, for example:

- Address unknowns and ask questions;
- Use either bullet points or prose to clearly and concisely outline your project;
- Go over two pages in length;
- Link to words, pages, and definitions that are not obvious to an outside reader;

- Use section headings to structure your document;
- Use first person prose;
- Solicit feedback from peers;
- Submit a single document as a research team.

If you are unsure about submission logistics and requirements, please err on the side of caution and ask Ryan. If you or your mentor believe that your research plan would contain infohazards, please let Ryan know ASAP, and we will discuss alternative formats. If you have any other questions, feel free to post them here, send Ryan a message, or ask anonymously.

Task purpose

This milestone should not get in the way of conducting research, but instead serve as a focus for critically evaluating your research project and its significance for AI alignment. We encourage you to think about the full scope of your research agenda, including work you plan to accomplish during and after MATS. Consequently, the research plan intentionally is designed to mimic our sense of the ideal grant application for the Long-Term Future Fund (LTFF). Regardless of whether you are applying for the extension program, we believe it is important to understand how to persuasively and tactfully explain the importance of your research in a way that appeals to funders and others in the AI safety community.

Although the LTFF application and the research plan are **two different documents**, the LTFF application will require much of the content found in the research plan. For example, the research plan primarily focuses on work you have done and will continue doing during the MATS program and the LTFF application primarily focuses on work you anticipate doing post-MATS. However, both require you to think about your project's threat model and theory of change. It is okay if the nature of your work or experiments deviates somewhat from your plan. If your project changes significantly during the extension program, you may need to contact the LTFF for advice.

The primary evaluation metric for progression in the MATS program (i.e., for the London extension) remains your mentor's assessment. The research plan will not replace mentors' evaluations; however, you should assume your mentor will review your report and account for it in their evaluation. Additionally this will serve as the basis for your grant proposal (e.g., to the Long-Term Future Fund) should you choose to apply for the Extension Phase of the MATS Program. You are welcome to publish your research plan on LessWrong or elsewhere after your submission. You should expect to receive feedback on your research plan by the end of the summer program.

If you require help in planning or refining your research plan, we recommend you consult your mentor and the Research Management team, and share drafts via the #feedback-on-posts-or-drafts Slack channel.

Resources

- Research plan, LTFF, and Extension Program FAQs

AI threat models/risk factors

- Joe Carlsmith, "Scheming AIs: Will AIs fake alignment during training in order to get power?"
- Ajeya Cotra, "Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover"
- Rohin Shah et al., "Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals"
- Joe Carlsmith, "Is Power-Seeking AI an Existential Risk?"
- Dan Hendrycks, Mantas Mazeika, and Thomas Woodside, "An Overview of Catastrophic AI Risks"
- Dan Hendrycks and Mantas Mazeika, "X-Risk Analysis for AI Research"
- DeepMind Alignment Team, "Threat Model Literature Review"
- Richard Ngo et al., "The alignment problem from a deep learning perspective"
- Many more LessWrong posts

Theory of change

- Effective Thesis, "Creating a Theory of Change"
- Michael Aird, "Building Your Theory of Change for Your Research" [Video]
- Open Philanthropy, "Reasoning Transparency"

Funders

- EA Funds: Long-Term Future Fund
- Open Philanthropy:
    - Early-career funding for individuals interested in improving the long-term future
    - Request for proposals: benchmarking LLM agents on consequential real-world tasks
- Survival and Flourishing Fund; to apply for funding, you will need a fiscal sponsor (e.g., Ashgro, BERI, Rethink Priorities Special Projects)
- Manifund
- Foresight Institute: AI Safety: Neuro/Security/Cryptography/Multipolar Approaches
- Cooperative AI Foundation: Cooperative AI Research Grants
- Center on Long-Term Risk: CLR Fund
- AE Studios: AE Grants
- National Science Foundation: Safe Learning-Enabled Systems

Example research plans

Please note the examples below are from past cohorts, and the evaluation criteria have since changed. See the rubric below for the current evaluation metrics.

- Example research plan #1 - Shashwat Goel
- Example research plan #2 - Anonymous
- Example research plan #3 - Anonymous
- Example research plan #4 - Anonymous (Group research plan)

Rubric (100 points total)

The following rubric is designed to provide a comprehensive assessment of each research plan based on the criteria and guidelines provided. Scores for each category are cumulative, leading to an overall evaluation of the proposal's quality.

1. Threat Model/Risk Factor (20 points)

- **Importance:** Does the proposal explain the potential impact of the proposed intervention on the chosen risk factor/threat model?
- **Tractability:** Does the proposal explain the feasibility of the proposed intervention on the risk factor/threat model?
- **Neglectedness:** Does the proposal explain the neglectedness of the proposed intervention on the threat model/risk factor?

2. Theory of Change (40 points)

- **Impact mechanism:** Does the proposal explain how the proposed research activities are expected to impact the threat model or risk factor?
    - Does the proposed mechanism of impact on the risk factor/threat model seem reasonable?
    - Does the proposal indicate any key assumptions or unknown variables that might affect the impact of the proposed research activities on the risk factor/threat model?
- **Risk analysis:** Does the proposal conduct an adequate risk analysis for the proposed research activities?
    - Are any potential "failure modes" of the proposed research activities indicated? Is there a plan to mitigate these?
    - Are any "dual-use research" or "infohazard" concerns indicated? Is there a plan to mitigate these?
- **Reasoning transparency:** Does the proposal display reasoning transparency?
    - Is the theory of change presented in a logical and coherent manner?
    - Are key uncertainties and assumptions indicated?

3. Planned Activities and Outputs (40 points)

- **Specific/Measurable:** Are proposed research activities and outputs specific and measurable?
    - Does the proposal include specific research activities and open questions to be answered (e.g., ML training runs, policy analysis, mathematical proofs, literature reviews, outcomes on benchmarks, stakeholder interviews, workshops)?
    - Does the proposal include specific outputs/goals (e.g., paper/report submission, LessWrong post, policy memos, talks, white papers)?

    - Will the proposed activities and outputs yield sufficient information to evaluate the success of the project?
- **Attainable/Realistic:** Are proposed research activities and outputs achievable within the scope of the program (including extension phase, if appropriate)?
    - Is there sufficient time to complete the proposed research activities?
    - Are there sufficient resources (e.g., mentorship, computing funds) to complete the proposed research activities?
- **Time-bounded:** Are research activities and outputs time-bounded?
    - Does the proposal include a tentative timeline of planned activities (e.g., LTFF grant, scholar symposium talk, paper/report submission)?
    - How might the timeline change if planned research activities are unsuccessful?