Progress report 2024-01 – 2024-06

Modeling Cooperation

Paolo Bova, Ben Harack, Jonas Emanuel Müller, Tanja Rüegg

Executive summary

Modeling Cooperation aims to improve safety-enhancing cooperation among competitors for the development of transformative AI. During the reporting period, we decided to mainly focus on the software we implement to facilitate the Intelligence Rising workshops which educate decision-makers about navigating AI competition risks.

After achieving feature parity with the previous solution in the last reporting period, we worked on preparing for the second release of the software during the next reporting period. This included gathering and addressing feedback from user tests, improving maintainability and robustness, and reducing technical debt.

Additionally, we increased our time spent on fundraising. Together with Technology Strategy Roleplay, the organization that runs the Intelligence Rising workshops, we were able to secure a $77,116 grant from the Foresight Institute.

During the reporting period, Modeling Cooperation consisted of seven team members corresponding to 1.25 FTE. Up until now, the permanent full-time team members received fixed wages of $30,000/year, while the part-time team members worked for $15.625/hour on a contract basis. Since these wages were initially determined in 2020, we offered all team members a yearly inflation adjustment starting in April 2024.

Description

Modeling Cooperation aims to improve safety-enhancing cooperation among competitors for the development of transformative AI (TAI). To achieve this goal, we focus on the following three work areas:

- Building software such as the Intelligence Rising application to leverage serious games that educate decision-makers about navigating AI competition risks.
- Creating tools for AI governance researchers to increase the impact of their AI governance research.
- Conducting research and authoring reports and papers that use modeling and simulation to evaluate ways to shift the incentives of AI race competitors toward more safety.

We consider the mitigation and prevention of dangerous competition for TAI high-impact because such dynamics could strongly incentivize the competitors to underinvest in safety, which in turn could lead to an increased risk of disaster.

Team

In 2023, we learned that our current funder's grant round schedule had changed. In previous years, we were able to secure funding from the Survival and Flourishing Fund (SFF) for the upcoming year in our funding round in August or September. Last year, the applications were due by the end of June.

Instead of bringing the application forward to June, we decided to wait until the next funding round. This decision was based on three reasons:

1. We wanted to finalize the redesign and extension of the Intelligence Rising software before submitting a new application.
2. We only learned about the new funding schedule shortly before the new deadline.
3. The funding we received the year before as well as the decision of our fiscal sponsor, Convergence Analysis, to waive the fees for the year 2021, ensured we could cover the costs of our current team members for 2024.

Using our carry-over from last year and fiscal sponsorship fees we had set aside, we were able to continue employing Jonas Emanuel Müller full-time as well as Paolo

Bova and Tanja Rüegg part-time. Ben Harack continues to finance his part-time contributions through a personal grant he received from Survival and Flourishing Projects. Jasmine Brazilek, Miles Tidmarsh, and Vasily Kuznetsov remain on hiatus.

During the reporting period, our seven team members corresponded to 1.25 FTE. Up until now, Jonas has received fixed wages of $30,000/year while the employed part-time team members have worked for the corresponding hourly wage of $15.625 on a contract basis, which means they only get paid when working and not when anything prevents them from doing so (e.g. sickness or vacations). Since these wages were initially determined in 2020, we offered all team members a yearly inflation adjustment starting in April 2024. Paolo decided to waive the increase and Tanja waived her full wages until further notice.

Strategy

At the beginning of the reporting period, we re-examined our planned work in light of our financial and human resources as well as our work to date and short-term goals.

During the previous period, we continued developing our software to facilitate the Intelligence Rising workshops—the workshops are based on a simulation game that allows participants to work together in teams to explore AI governance scenarios as they advance through an AI technology tree.

At first, the goal was to complement the existing ad-hoc combination of Google Slides and messaging tools by making many new features available that weren't possible before. Now, the goal is to solely use our software (as well as audio calling for online workshops) to facilitate the workshop. Thus, we have been working toward feature parity with the previous solution and were able to achieve it during the last reporting period. Our goal for this reporting period is to gather and address feedback from user tests and improve the maintainability as well as robustness of the software to prepare for the release of the second version of the software in the next reporting period.

We assess such collaborations to be promising projects for Modeling Cooperation because they help to enable other researchers and policymakers to gain an intuitive understanding of AI competition models or AI development scenarios—a research direction our stakeholders repeatedly expressed interest in—and because we are

exceptionally well-positioned to build user-friendly tools like web apps based on AI governance research results given our backgrounds and skills.

As mentioned in our team section, we learned in 2023 that our current funder's grant round schedule had changed. This resulted in us deciding to wait until their next funding round and to allocate more resources to fundraising in the second half of 2023. We used this time to create an initial application for Foresight's AI safety grant together with the organization Technology Strategy Roleplay (TSR) that runs the Intelligence Rising workshops. Additionally, we started exploring other options to potentially establish the Intelligence Rising software as a self-sustaining project with funding independent of Modeling Cooperation.

Therefore, we decided to focus on improving the Intelligence Rising software, finalizing Foresight's AI safety grant application, and looking into additional funding sources during this reporting period.

Accomplishments

Intelligence Rising software

During the previous reporting period, we implemented the features that were missing to achieve feature parity between our Intelligence Rising software and the previous solution.

Our software consists of a sleek interface for the participants and a more advanced interface for the facilitators. It allows the facilitator to (automatically) group the participants into teams, lead them through multiple levels of the AI technology tree, track the progress of publicly or secretly researched papers and products within the tech tree, keep track of the teams' stats and actions at any given state, record additional custom or predefined ad-hoc events during the workshop which might influence the outcome of the game, and create non-player controlled teams (NPC) teams to steer the outcome of the workshop. Additionally, the participants can create forecasts of their chances to (be the first to) discover a paper or product depending on the resources they would invest, follow the state of the AI technology tree, and submit actions.

Intending to release the second version of the software during the next reporting period, we focused on improving and user-testing the software. We accomplished the following achievements during this reporting period:

- **Gathered and addressed feedback from user tests with facilitators and participants.** Having achieved feature parity with the previous solution during the last reporting period, both TSR and Modeling Cooperation conducted user tests with facilitators and potential participants. As a result, we addressed the feedback as follows:
    - **Implemented a new desktop sidebar layout to increase the visibility of the most relevant game elements.** The user tests showed that the world state should always be visible and that the timeline was sometimes difficult to find. Due to the limited real estate of the main page of the app containing the tech tree, all other game elements were either positioned in the sidebar, which only allowed for one expanded item at a time, or on other sub-pages. Based on the user feedback, we introduced a second sidebar for the desktop view, which allows for the world state and timeline to be always visible for facilitators and participants (unless they actively take steps to hide them).
    - **Redesigned the main page to reduce the number of tabs.** Related to the feedback above, the users asked whether it is possible to reduce the number of tabs present within the Intelligence Rising software. In addition to adding the second sidebar, we took additional steps to reduce the complexity of the software. This resulted in us being able to remove a full level of tabs.
    - **Introduced an application help button.** To increase the ease of use, we added a help button that opens a summary helping the facilitators and participants to navigate the different elements of the software.
- **Achieved a sustainable database schema by implementing various data deduplication strategies.** The game state data in the software has achieved such a significant size that we ran into issues when sending it over the network. Thus, we leveraged several data deduplication strategies that reduce the amount of data we store and send over the network. This significantly increases the scalability of the software, which is especially important before broader adoption of the software by TSR which is planned for the next reporting period.
- **Reduced technical debt to improve the long-term code quality, reduce maintenance costs, and enhance the system's ability to adapt to future changes.** After a longer period of focusing on extending the scope of the Intelligence Rising software to achieve feature parity with the previous solution, we made sure to assess and reduce the technical debt. According to Wikipedia, technical debt refers to the implied cost of additional work in the future resulting from choosing an expedient solution over a more robust one. Sometimes,

prioritizing speed over robustness can be the right choice, as long as the technical debt is resolved before it significantly increases future costs. Thus, we decided to address the technical debt before the release of the second version of the software which is planned for the next reporting period.

Fundraising

- **Secured a $77,116 grant from the Foresight's AI safety grant together with TSR.** After we created an initial version of the funding application during the last reporting period, we submitted the finalized version in February 2024. We decided to create a shared application because we're working toward establishing the Intelligence Rising software as a self-sustaining project with funding independent of Modeling Cooperation. Thus, the application was restricted to the Intelligence Rising software and asked to cover the costs for maintenance and basic feature requests for a year. In May 2024, we received confirmation from the Foresight Institute that they'll fund the full requested amount of $77,116.
- **Applied to Schmidt Futures' open call of the Virtual Institute on Grand Strategy and AI.** In March 2024, we learned about this Schmidt Futures opportunity. Among other things, they were looking for applicants who use LLMs to model the behavior of foreign actors and who use AI to improve games that are in the same genre as Intelligence Rising. Thus, we applied with a proposal to extend the Intelligence Rising software using LLMs—a project we've been interested in for a while, since it would address one of TSR's main bottleneck on training new facilitators by allowing training workshops to be run even if no or only a few human participants are available. Unfortunately, they informed us in May 2024 that they aren't able to fund the project despite its solidity because they funded similar projects and wanted a strong diversity on the final list.
- **Applied to SFF**. After SFF had changed its grant round schedule in 2023, we decided to wait until the next funding round to apply again for general support of Modeling Cooperation. The applications were due in June 2024. By then, we already had confirmation that the Foresight Institute was going to fund the request covering the development of the Intelligence Rising software for a year. Thus, we applied with a more ambitious request to grow Modeling Cooperation as an organization.

We also submitted the Schmidt Futures and SFF application to the Nonlinear Network Fund. The fund manager informed us that there's no time we should expect

to hear back because the funders on the platform all have different funding schedules.

Roadblocks

- **A change in our current funder's grant round schedule resulted in us having to spend more time on fundraising.** In previous years, we were able to secure funding from the Survival and Flourishing Fund for the upcoming year in our funding round in August or September. Last year, the applications were due by the end of June which wasn't feasible for us. Luckily, the funding we received in 2024 covered our expenses for the entire year 2023. That being said, it still resulted in us deciding that we need to spend more time on fundraising which we'd have otherwise been able to spend on the Intelligence Rising software.
- **The Intelligence Rising software became so large that the game state data achieved such a significant size that we ran into issues when sending it over the network.** The goal of fully replacing and extending the previously used ad-hoc combination of Google Slides and messaging tools—except for audio calling—resulted in a feature-rich application. While this allows for an efficient facilitation of the workshops, it necessitated the implementation of various data deduplication strategies to reduce the amount of data we store and send over the network.

Future goals

These goals are subject to change and contingent on funding as Modeling Cooperation aims to use its limited resources to investigate the most impactful research questions within AI governance. Thus, we regularly re-evaluate our planned work taking new research opportunities and feedback from our stakeholders into consideration.

Short-term goals

- **Release the second version of the software we implemented to facilitate the Intelligence Rising workshops.** During this reporting period, we focused on gathering and addressing feedback from user tests, improving maintainability and robustness, and reducing technical debt. Having completed these preparations, we aim to release the second version of the software during the next reporting period.

- **Finalize the extension of our technical report *Safe Transformative AI via a Windfall Clause* and upload the new version of our report to arXiv.** In addition to evaluating a Windfall Clause, designed in a public information setting and applied to the model presented in the paper *Racing to the Precipice*, we managed to extend our findings to the private and no information settings present in previous works during the first half of 2022. Building upon this success, we were also able to examine when disclosure of information is helpful or harmful to a Windfall Clause. We plan to incorporate our new insights into our revised technical report which we then would like to publish on arXiv.

Longer-term goals

- **Further investigate the Windfall Clause policy when applied to a model with more parameters than the one presented in the paper Racing to the precipice.** For analysis similar to the examination of the Windfall Clause policy, we think it is useful to start with a simple AI competition model to gain a basic but thorough understanding of the policy's most important dynamics. Once we author a paper about our initial analysis, we then plan to advance our research by applying the Windfall Clause policy to a more complex model that is more representative of real-world AI developments.
- **Expand our ongoing and start new collaborations with AI governance academics to build tools for other researchers and policymakers to help build intuitions for AI competition.** When discussing the impact of potential research directions, our stakeholders repeatedly expressed interest in projects that enable other researchers and policymakers to gain an intuitive understanding of AI competition models or AI development scenarios. Given the background and skills of Modeling Cooperation's team members, we would argue that we are exceptionally well-positioned to build user-friendly tools like web apps based on AI governance research results. This view was strengthened during our ongoing collaboration with Robert Trager, who already expressed interest in continuing the project as well as significantly expanding its scope in the future. We would like to work toward building an AI governance modeling platform that enables other researchers and policymakers to explore our computational models and those of our collaborators.
- **Build upon the achievements Jonas accomplished during the project "Using Bayesian ML to find more interpretable solutions to AI race models":** Having been awarded a personal grant by Survival and Flourishing, Jonas worked on

making numerical optimal solutions of an AI race model interpretable by humans using MIT's new Bayesian program learning ML system DreamCoder.

    - **Submit DreamCoder improvements implemented as part of the aforementioned project:** Jonas would like to contribute to the DreamCoder project by submitting his improvements to the researchers. To do so, he plans to document the new infrastructure and polish the bug fixes and extensions. We hope that these improvements help the researchers at MIT as well as other researchers outside of MIT to benefit from the DreamCoder ML system, which provides a safer alternative to current approaches to AGI.
    - **Refine the interpretable strategies gained as part of the aforementioned project and incorporate the solutions into our AI competition simulations:** Jonas would like to refine the results in terms of improving the accuracy as well as robustness and training DreamCoder to provide solutions for games with more dimensions. Afterward, he would like to use the result code from DreamCoder to make the agents in our existing AI competition simulation behave optimally.
- **Write a series of blog posts elaborating on Modeling Cooperation's approach, methods, and models.** One goal of our relaunched website is to allow anyone to stay up to date with our work. In addition to announcing our research results and research support software tools, we also want to share the insights we gained during our work for other researchers to benefit from. One example is our comparison of multiple analytical methods to solve dynamic AI competition games which could help other researchers to decide on a suitable approach without having to try out the various approaches themselves. Another example is our review of often technical and complex literature relevant to AI competition, such as game theory and industrial organization, which we could write up tailored to other AI competition researchers.
- **Integrate statistical features into our Monte Carlo simulation tool.** We would like to integrate features such as null hypothesis significance testing, parameter sampling, and t-digest support to further increase our economists' productivity.