Progress report 2023-07 – 2023-12

Modeling Cooperation

Paolo Bova, Ben Harack, Tjankui Lamon, 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 focus on the software we implement in collaboration with Shahar Avin to facilitate his and his team's Intelligence Rising workshop which educates decision-makers about navigating AI competition risks as well as fundraising.

We continued developing the software with the goal of fully replacing and extending the previously used ad-hoc combination of Google Slides and messaging tools—except for audio calling. This included allowing the workshop facilitator to create custom events during the workshop and representing concerns that arise when participants navigate AI competition risks. We also added a UI for managing non-player-controlled teams and ensured GDPR compliance. Additionally, we created an initial version of a shared funding application together with Intelligence Rising.

Tjankui Lamon decided to move on from Modeling Cooperation in October. The other full-time team member still receives fixed wages of $30,000 annually and the employed part-time team members work for $15.625/hour on a contract basis. During the reporting period, Modeling Cooperation consisted of eight team members corresponding to 1.8 FTE.

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

As in 2022, Modeling Cooperation entered 2023 with the support of a $83,000 grant from Jaan Tallinn as part of the Survival and Flourishing Fund. This allowed us to continue employing Jonas Emanuel Müller and Tjankui Lamon 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. At the end of October, Tjankui decided to move on from Modeling Cooperation and go back to working on various smaller projects as a freelancer instead due to it being a better fit for her career plan.

During the reporting period, our eight team members corresponded to 1.8 FTE. Jonas continues to receive fixed wages of $30,000 annually and the employed part-time team members work 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). For Tjankui, we

decided to increase her wages when we extended her contract in February 2023 due to higher living costs.

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 reporting period, we significantly extended our software to facilitate Shahar Avin's and his team's Intelligence Rising workshop—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 on adding all necessary features to completely replace and improve upon anything that was previously possible with Google Slides or messaging tools and plan to continue doing so during this 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.

Additionally, we learned that our current funder's grant round schedule had changed at the end of the previous progress report. 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. This year, the applications were due by the end of June. We decided to wait until the next funding round instead because we learned about the deadline only shortly beforehand, we wanted to finalize the current redesign and extension of the Intelligence Rising software first, and the funding we received last year 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 a few months into 2024.

That being said, we decided to allocate more resources to fundraising in the second half-year to explore other options and look into establishing the Intelligence Rising software as a self-sustaining project with funding independent of Modeling Cooperation.

To reach these two goals, we decided to focus on the Intelligence Rising software as well as fundraising during this reporting period.

Accomplishments

During the previous reporting period, we extended and redesigned the first version of the Intelligence Rising software. The software consists of a sleek interface for the participant and a more advanced interface for the admin. It allows the admin 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 at any given state, and record additional ad-hoc events during the workshop which might influence the outcome of the game. 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 and follow the state of the AI technology tree.

Our goal is to fully replace and extend the ad-hoc combination of Google Slides and messaging tools Intelligence Rising used previously—except for audio calling. Therefore, the following accomplishments focus on the Intelligence Rising software.

- **Added custom instability events allowing the admins to introduce additional events influencing the gameplay.** During the workshop, global stability is represented as a value between 0 and 10. When global stability is low, the admin can trigger instability events. The effects of the events can vary greatly, from lowering the global stability further to affecting the characteristics of specific teams or increasing the global stability. Additionally, teams can choose to mitigate the events to minimize the negative effects. Until now, the admin could only choose from predefined instability events. Now, all events are custom and built on the fly using a template. This allows the admin to choose predefined instability events, edit their effects and mitigation options, or create fully custom ones.
- **Implemented the ability to edit, and manage non-player controlled teams (NPC teams) within the Intelligence Rising software.** During the last reporting period, we added the functionality to create NPC teams. The goal of integrating

NPC teams was to always enable admins to properly convey the potential dangers of TAI. For example, the admins would be able to intervene by managing additional teams that ensure the participants end up in an AI race in case the participants act unrealistically cooperative. To allow admins to edit and manage the NPC teams, we created a new interface with the following functionalities:

    - When creating new NPC teams, the software makes sure to warn the admin when a team already exists and deals gracefully with any typos related to whitespace.
    - The name of an NPC team can be edited at any given time during the workshop without creating backward-incompatible changes.
    - For each NPC team, the admin can choose for which views within the software the team should be available. The software contains four different views: 1) the world state representing the characteristics of each team, 2) the tech tree displaying which technologies can be researched, 3) the discovery table summarizing the researched technologies, and 4) the forecast table allowing the user to calculate the probability of discovering a specific technology. By default, the NPC teams show up everywhere that the participant teams show up.

- **Added concern cards informing the participants about concerns that arise when specific items on the technology tree are researched.** Concerns are a specific type of event that the admin can trigger in response to the successful research of an item on the technology tree. The participants are then given the chance to mitigate these effects by taking action. The concerns are now represented on the technology tree and the admin can also create custom concerns.
- **Ensured GDPR compliance by developing a system that allows complete deletion of personal identifiable information (PII) upon request.** While the database maintains a full history, we introduced a feature that erases both the current state and history of PII without breaking important data links. By moving sensitive information to a separate database table and using anonymous UserIDs, the system preserves necessary relationships while fully anonymizing user data.

Additionally, we created an initial version of a funding application for Foresight's AI safety grant together with Intelligence Rising. 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.

Roadblocks

- **Tjankui's decision to move on from Modeling Cooperation reduced our human resources and tied them up since she had to hand over her tasks to other team members.** Tjankui left Modeling Cooperation at the end of October to go back to working on various smaller projects as a freelancer instead due to it being a better fit for her career plan. Since our human and financial resources didn't allow for finding a replacement, her tasks had to be handed over to our other software engineers. Overall, her departure as well as offboarding reduced our human resources.
- **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. This year, the applications were due by the end of June which wasn't feasible for us. Luckily, the funding we received last year covers our expenses for this entire year. 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.

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

- **Improve and optimize the software we implemented in collaboration with Shahar Avin to facilitate his and his team's Intelligence Rising workshop.** After focusing on extending the features of the Intelligence Rising software during the last two reporting periods, we'd like to focus on improving and optimizing the current version of the software. This consists of two main areas: 1) Reducing technical debt to improve the long-term code quality, reduce maintenance costs, and enhance the system's ability to adapt to future changes. 2) Getting a sustainable database schema ready before broader adoption by the

Intelligence Rising team since the game state data in the application has achieved such a significant size that we ran into issues when sending it over the network. For the latter, we plan to implement various data deduplication strategies that reduce the amount of data we store and send over the network.

- **Finish the application for Foresight's AI safety grant together with Intelligence Rising and explore additional funding sources.** We have been looking into establishing the Intelligence Rising software as a self-sustaining project with funding independent of Modeling Cooperation. During this reporting period, we started this process by creating an initial version of a funding application for Foresight Institute together with Intelligence Rising. We are planning to finalize this application as well as explore additional potential funding sources.

Longer-term goals

- **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.
- **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.