Progress report 2023-01 – 2023-06

Modeling Cooperation

Paolo Bova, Jasmine Brazilek, Alex Dietz, Ben Harack, Vasily Kuznetsov, Tjankui Lamon, Jonas Emanuel Müller, Tanja Rüegg, Miles Tidmarsh

Executive summary

Modeling Cooperation aims to gain better insight into the dynamics of competition for the development of transformative AI to look for opportunities to promote safety-enhancing cooperation among competitors. During the reporting period, we decided to fully focus on the software we implement in collaboration with Shahar Avin to facilitate his and his team's Intelligence Rising workshop where participants explore AI governance scenarios. Having tested the first version of the software during the previous reporting period, the Intelligence Rising team expressed interest in the software not just complementing the ad-hoc combination of Google Slides and chat messages used previously but fully replacing and extending it instead.

With that goal in mind, we extended the features of the software, redesigned the interface, improved the user experience, and further improved our infrastructure to ensure a cost-effective and robust setup. After welcoming two new team members in November with a fixed-term contract, we were able to extend one of the two contracts—with increased wages due to changed circumstances. While the other full-time team member still receives fixed wages of $30,000 annually, the employed part-time team members work for $15.625/hour on a contract basis. Currently, Modeling Cooperation consists of nine team members corresponding to 2.4 FTE.

Description

Modeling Cooperation aims to gain better insight into the dynamics of competition for the development of transformative AI (TAI) by building research software tools and conducting research to look for opportunities to promote safety-enhancing cooperation among competitors. 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

Like 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 one additional person full-time as well as Paolo Bova and Tanja Rüegg part-time.

Our two newest team members, Tjankui Lamon and Alex Dietz, joined Modeling Cooperation as full-time software engineers at the beginning of November 2022 to replace Paolo's reduced work hours since he switched to working part-time after starting his Ph.D. at the beginning of October 2022. Both of them initially signed a fixed-term contract until the end of January 2023 because the extension of the contracts was contingent on fit and funding. Unfortunately, the above-mentioned grant only allowed us to continue funding one of the two new team members. In January 2023, we decided to extend Tjankui's contract while parting ways with Alex. Our decision was based on our current needs and with Tjankui being a software engineer focusing on UI design, she impressed us by finding and implementing creative and feasible solutions to complex challenges.

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

Until now, our permanent full-time team member receives fixed wages of $30,000 annually and the temporary full-time team members as well as 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, vacations, COVID-19). When discussing

the desired extension of Tjankui's contract, she informed us that her circumstances had changed insofar as she was dependent on higher wages to cover her living expenses. Together, we decided to increase her wages until the end of the year by offering her the money we didn't spend last year due to Jonas waiving some of his wages, Ben covering his wages through another grant, and reduced costs for legal consultants and freelancers.

During this reporting period, Jonas decided to waive an additional 4,800 Swiss francs he earned as a freelancer outside of Modeling Cooperation. This allows us to start building up a small runway again and, thus, increase the job security for our team members. At the moment, our nine team members correspond to 2.4 FTE.

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 released the first version of 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. To let the Intelligence Rising team test the software in their actual working environment, we organized a workshop with Shahar Avin as the game administrator.

The feedback was so positive that the Intelligence Rising team expressed interest in increasing the project scope: At first, the goal was to complement the existing ad-hoc combination of Google Slides and chat messages by making many new features available that weren't possible before. Now, the goal is to solely use our software (as well as a voice chat application for online workshops) to facilitate the workshop. Thus, it is important to add enough features that it completely replaces and improves upon anything previously possible with Google Slides or chat messages. To reach this goal, we decided to fully focus on this project 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.

Accomplishments

As mentioned in the previous section, we decided to fully focus on our Intelligence Rising software during this reporting period. The goal of the software is to simplify, automate, and improve the facilitation of Shahar Avin and his team's Intelligence Rising workshop. The workshop is a round-based simulation game that allows participants to work together in teams to explore AI governance scenarios as they advance through an AI technology tree.

During the last progress report, we released the first version of the software which consists of a sleek interface for the participants 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.

Together with the Intelligence Rising team, we decided to increase the project scope so that the software not just complements but fully replaces and extends the ad-hoc combination of tools to facilitate the workshops used until now. Therefore, the following accomplishments focus on the Intelligence Rising software.

- **Extended the features of the Intelligence Rising software to completely replace the ad-hoc combination of Google Slides and chat messages used previously.** Until now, teams submitted up to three actions per round of the game to the admin via chat messages. The goal was to add this functionality to the software, inform the admin about the submission in real-time via an indicator, and introduce a more standardized process to communicate the decisions. To do so, we had to fulfill the following requirements:
    - Talent assignment:
        - Each team starts with a pre-defined number of talent points and can collect additional ones by researching items on a tech tree. The tech tree works as follows: Different research items on the

technology tree result in different rewards and unlock new technologies on the path toward transformative AI. The last technologies that can be researched in the technology tree are forms of transformative AI.

- Every team can assign its available talent to work on researching papers, products, or AI capabilities that haven't been publicly researched yet but are already unlocked on the tech tree. The more talent points a team assigns to an item, the higher the likelihood of them discovering it. To support the decision-making for the teams, the software already offers a forecasting tool to determine the chances of each team discovering the item (first) depending on the number of talent points one expects each team to assign to it.
- Each talent assignment can either be done in secret or public.
- The admin should be able to overwrite the decisions of the teams. For example, a team might assign all of their four available talent points to one product. If the admin then reduces this to three talent points, the team must be able to re-assign the remaining talent.
- We implemented this functionality by adding two talent counters to each tech tree item: one for secret and one for public assignments.

- Two free actions:
  - In each round, every team can submit two actions they want to take. This can be any action that the actual state or AI company could do in the real world. For example, team China could launch a cyber attack against Alphabet to steal its latest research.
  - While the first of these actions is always public, the team can choose whether the second action should be public or secret.
  - Once the user submits an action, an indicator notifies the admin about the new submission in real time allowing the admin to incorporate the action into the workshop.
  - The admin formalizes the actions into a quantitative effect on the game. For example, the admin would usually determine the success probability of China launching a cyber attack against Alphabet by comparing the two teams' cyber power (which is being tracked as part of the world state). If the cyber attack ends up being successful, the admin could mark one of Alphabet's

latest technologies in the technology tree as also having been discovered by China. If the action was public, it would be visible to all teams. If the action was secret, it would be hidden from other teams.
    - We implemented this functionality as two free-text fields with the second one containing a button allowing the user to choose whether the action should be public or secret.
- **Redesigned the Intelligence Rising software and improved the user experience.** While adding the functionalities mentioned above better supports the facilitation of the workshop, it also highlighted the need for a redesign of the software to allow for easier and faster navigation within the interface. First, we identified the areas of improvement concerning the user experience. Second, we created different mockups of the overall structure and individual pages—always optimizing for mobile devices too. Third, we discussed our suggestions with Shahar and collected additional feedback. Fourth, we decided on our target vision together before breaking it down into tasks and prioritizing them accordingly. The largest changes included adjusting the website structure, adding a sidebar containing functionalities such as the actions as well as talent assignments, and standardizing various design elements across all functionalities.
- **Implemented the ability to create non-player controlled teams within the Intelligence Rising software.** Until now, the software only supported teams consisting of participants.
- **Added pre-defined instability events allowing the admins to introduce additional events influencing the gameplay.**
- **Implemented optimistic updating to significantly enhance the application's responsiveness.**
- **Further improved our infrastructure for the Intelligence Rising software to ensure a cost-effective and robust setup.**

Roadblocks

- **Having very limited funding forced us to part ways with one of our software engineers.** When Paolo reduced his work hours due to starting a Ph.D. at the beginning of October 2022, we were excited to welcome not one but two new full-time software engineers at the beginning of November 2022—Tjankui Lamon and Alex Dietz. Both of them initially signed a fixed-term contract until the end of January 2023 because the extension of the contracts was contingent

on fit and funding. Unfortunately, the available funds only allowed us to continue funding one of the two new team members. In January 2023, we decided to extend Tjankui's contract while parting ways with Alex. Additionally, we had to agree to increase the new permanent full-time team member's wages after the trial period since the wage wouldn't cover their living expenses at their usual place of living.

- **Not being able to outsource creating new versions of our contracts blocked some of our researchers' resources.** The limited funding not only affected our staffing but also the resources of our team members. During this reporting period, we had to create new versions of the contracts between our fiscal sponsor and all paid team members of Modeling Cooperation. Due to offering our new permanent full-time team member the money we didn't spend last year, our team members had to take this work on themselves instead of being able to outsource this task.

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 in collaboration with Shahar Avin to facilitate his and his team's Intelligence Rising workshop.**
- **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 which 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 which 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 own 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 researcher 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.