Progress report 2022-07 – 2022-12

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 launched our research software tool SPT Model which we built in collaboration with Professor Robert Trager. The tool implements the Safety-Performance Tradeoff model created by Robert Trager, Paolo Bova, Nicholas Emery-Xu, Eoghan Stafford, and Allan Dafoe. It allows other researchers and decision-makers to explore how improvements in technical AI safety could affect the safety choices of competing AI developers.

We also released the first version of the software we implemented in collaboration with Shahar Avin to facilitate his and his team's Intelligence Rising workshop where participants explore AI governance scenarios. Furthermore, we welcomed two new team members in November to replace the reduced work hours of one of our full-time team members who switched to working part-time after starting their Ph.D.—resulting in slightly expanding our human resources. Currently, Modeling Cooperation consists of nine team members corresponding to 2.4 FTE. While the permanent full-time team members receive fixed wages of $30,000 annually, the

temporary full-time team members as well as employed part-time team members work for $15.625/hour on a contract basis.

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

Modeling Cooperation entered 2022 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 full-time and Tanja Rüegg part-time. Paolo Bova has been one of our full-time contributors since 2019 and switched to working part-time after starting his Ph.D. at the beginning of October 2022. To fulfill the need for software engineering resources in our ongoing collaborations, we decided to replace Paolo's reduced work hours—and slightly expand our human resources—by adding two full-time software engineers, Tjankui Lamon and Alex Dietz, to our team. Both of them will initially be working for us from November 2022 to February 2023 with the extension of the contracts being contingent on fit and funding.

Ben Harack started to finance his part-time contributions through a personal grant he received from Survival and Flourishing Projects. In combination with Jonas waiving 3,350 Swiss francs he earned as a freelancer outside of Modeling Cooperation during this reporting period, we can continue building up a small runway and, thus, increase the job security for our researchers. Jasmine Brazilek has been continuing to support Modeling Cooperation part-time on a voluntary basis while Miles Tidmarsh and Vasily Kuznetsov have remained on hiatus. At the moment, our nine team members correspond to 2.4 FTE. Our permanent full-time team members receive 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).

We would like to quickly introduce our newest team members, Tjankui and Alex:

- Tjankui started as a software engineer in 2008, building web applications for small companies and startups where she developed her passion for software and UI/UX. She also worked as a consultant for one of the biggest banking and telecom companies in Belgium.
- Alex is a software engineer in the Boston area, with a focus on frontend web development. He also holds a Ph.D. in philosophy, and has published articles on well-being and collective obligations.

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. Being close to the launch of our research software tool built in collaboration with Robert Trager and having created a test release of our software to facilitate Shahar Avin's and his team's Intelligence Rising workshop during the previous reporting period, we decided to fully focus on these two projects.

- Together with Robert Trager—a professor of international relations at UCLA and strategic modeling team lead at GovAI—we decided to apply some finishing touches and launch the research software tool during this reporting period. The tool will allow other researchers and decision-makers to explore how improvements in technical AI safety could affect the safety choices of competing AI developers.
- In our project with Shahar Avin, who is a senior research associate at CSER, we decided to continue implementing software for his and his team's Intelligence Rising workshop—a simulation game that allows participants to explore AI governance scenarios with other teams as they advance through an AI technology tree. The goal for this reporting period is to release the first version of the software and to conduct field user testing, i.e. letting the team of Intelligence Rising test the software built for facilitating the workshop in their actual working environment. Later on, the software will enable us to explore how access to different tools can influence the decision-making of participants, who may in the future hold key positions relevant to AI governance.

We assess these 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

- **Launched our research software tool implementing the Safety-Performance Tradeoff model we built in collaboration with Robert Trager.** During the first half of 2022, we revised the Safety-Performance Tradeoff model created by Robert Trager, Paolo Bova, Nicholas Emery-Xu, Eoghan Stafford, and Allan Dafoe which we display in our research software tool. In preparation for its launch, we also polished the tool with regard to its ease of use—increasing accessibility, clarity, and design—and created a blog post elaborating on the project which we shared with a number of researchers from GovAI when we presented the new research software tool to them. Building upon our progress, we finalized and launched our tool during this reporting period by completing the following four steps: First, we incorporated the feedback we received regarding our planned blog post after presenting the tool to the other AI governance researchers. Second, we applied some finishing touches which included the integration of our Modeling Cooperation branding as well as website and the replacement of the previously implemented web analytics with a solution that is not just GDPR-compliant but doesn't even use cookies or collect personal data—thereby adhering to the recent decision of ruling Google Analytics illegal in different European countries. Third, we polished the explanatory texts and newly introduced website tour to improve the process of getting to know our research software tool. Fourth, together with Robert Trager, we announced the launch of our software research tool on the Effective Altruism forum to allow other researchers and decision-makers to explore how safety insights could affect the safety choices of competing AI developers.
- **Released the first version of the software we implemented in collaboration with Shahar Avin to facilitate his and his team's Intelligence Rising workshop and conducted field user testing.** This workshop allows participants to explore AI governance, and in particular AI competition, with other teams as they advance through an AI technology tree. We created a test release and held a presentation showing and explaining the functionality of the software to the

game administrators during the previous reporting period. To advance to releasing the first version of the software, we finalized two large work packages during this reporting period: We finished the implementation of the tech tree through which participants advance to explore AI governance scenarios with other teams—one of the core components to facilitate the workshop—and we migrated our database and our framework due to renaming of the database we currently use. Afterward, we were ready to release the first version of the software and, as mentioned in the short-term goals of our previous progress report, eager to conduct field user testing, i.e. letting the team of Intelligence Rising test the software in their actual working environment when conducting a workshop. Thus, we organized a workshop with Shahar Avin as the game administrator and multiple participants who worked in teams as they advanced through an AI technology tree. Shahar Avin used the software 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. The participants used the software to create forecasts of their chances to (be the first to) discover a paper or product depending on the resources they would invest and to follow the state of the AI technology tree.

Roadblocks

- **Having very limited funding increased the difficulty of finding suitable team members to replace the reduced work hours after one of our full-time team members switched to working part-time after starting their Ph.D.** Our team members are highly dedicated and mission-aligned individuals who either contribute on a voluntary basis or for a wage that just covers their living expenses. This is especially difficult for full-time team members since we can only offer short-term job security due to the current funding process covering a year at a time. Accepting such a low wage is only possible because our team members have a financial safety net provided by another job or family and friends. These circumstances increased the difficulty of replacing the reduced work hours after one of our full-time team members switched to working part-time after starting their Ph.D. We were lucky to find two people whose current career goal and/or place of living allowed them to accept our current wage. Nevertheless, this meant we already knew that we could only permanently hire one of the two new team members. 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.

- **After the initial user field test, our collaborator desired additional features to allow him and his team to conduct the Intelligence Rising workshops using mainly our software and completely replacing the application that is currently used for the game. While we were excited about this positive feedback, it resulted in a delay regarding letting the broader team of Intelligence Rising test the software built for facilitating their workshop.** The goal for this reporting period was to release the first version of the software and to conduct field user testing, i.e. letting the team of Intelligence Rising test the software built for facilitating the workshop in their actual working environment. After successfully releasing the first version of the software and organizing a workshop with Shahar Avin as the game administrator, he expressed interest in us implementing additional functionality allowing him and his team to conduct the workshops using solely our software. We agreed to the desired changes since it increases the usefulness of the software knowing that this will result in delaying letting the broader team test the 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

- **Release the second version of the software we implemented in collaboration with Shahar Avin to facilitate his and his team's Intelligence Rising workshop.** During this reporting period, we released the first version of the software and organized a workshop with Shahar Avin as the game administrator. Afterward, Shahar desired additional features to allow him and his team to conduct the Intelligence Rising workshops using solely our software. We are excited about the positive feedback and plan to extend the functionality of the software during the upcoming reporting period. Then we can continue to conduct field user testing, i.e. letting the team of Intelligence Rising test the software built for facilitating the workshop in their actual working environment

- **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.