Progress report 2021-07 – 2021-12

Modeling Cooperation

Paolo Bova, Jasmine Brazilek, Ben Harack, Vasily Kuznetsov,
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 completed the first release of our research software tool SPT Model which we built in collaboration with Associate 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 started our collaboration with Shahar Avin to implement software for his and his team's Intelligence Rising workshop where participants explore AI governance scenarios.

Finally, we revised and extended our technical report evaluating the Windfall Clause policy when applied to the model presented in the paper *Racing to the Precipice* and furthered our collaborations with Robert Trager's and The Anh Han's teams. Currently, Modeling Cooperation consists of seven team members corresponding to 2.2 FTE. While the full-time team members receive fixed wages of $30,000 annually, the employed part-time team member works for $26/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 2021 with the support of a $74,000 grant from Jaan Tallinn as part of the Survival and Flourishing Fund as well as a $9,200 personal grant from Survival and Flourishing for a project proposed by Jonas Emanuel Müller. This allowed us to continue employing Paolo Bova and Jonas full-time as well as Ben Harack part-time. During the two months of full-time work, Jonas set aside for the project he had proposed, his wages were covered by his personal grant. Even though he was granted more than his current hourly wage at Modeling Cooperation, he decided to waive the additional amount of $4,000 as well as a total of 2,800 Swiss francs he earned as a freelancer outside of Modeling Cooperation in 2021 to start building up a small runway and, thus, increase the job security for our researchers.

Tanja Rüegg and Jasmine Brazilek have been continuing to support Modeling Cooperation part-time on a voluntary basis while Miles Tidmarsh and Vasily Kuznetsov have remained on hiatus. Overall, Modeling Cooperation consists of seven team members corresponding to 2.2 FTE. Both full-time team members receive fixed wages of $30,000 annually and the employed part-time team member works for $26/hour on a contract basis, which means he only gets paid when working and not when anything prevents him from doing so (e.g. sickness, vacations, COVID-19).

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. Since Jonas completed the project “Using Bayesian ML to find more

interpretable solutions to AI race models" for which he had been awarded a personal grant during the first half of 2021, we were able to allocate our available 2.2 FTE freely.

After starting our collaboration with Robert Trager and laying the foundation for our collaboration with Shahar Avin during the first half of 2021, we decided to dedicate almost all of our resources to these two projects.

- Together with Robert Trager—an associate professor of international relations at UCLA and strategic modeling team lead at GovAI—we decided to continue working on our research software tool implementing an AI competition model with the goal of completing the first release 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 implement 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. Our software will facilitate the workshop and 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.

Building upon the work we did in the previous reporting period, we decided to revise and substantially extend our technical report in which we evaluated the Windfall Clause policy—an *ex ante* commitment by AI firms to donate a significant amount of any eventual extremely large profits—when applied to the model presented in the paper *Racing to the Precipice* by Armstrong, Bostrom, and Shulman. This revision and extension consist of incorporating feedback that we had received from other researchers at the end of the previous reporting period and evaluating the Windfall Clause under a scenario with private information.

Additionally, we decided that Jonas would spend any remaining time contributing to the DreamCoder software project which he used during his project "Using Bayesian ML to find more interpretable solutions to AI race models". Building on the

achievements during the previous reporting period, the goal is to document and polish the improvements he implemented and to submit them to the MIT researchers who created DreamCoder. We hope that the new infrastructure, bug fixes, and extensions help other researchers within and outside of MIT to benefit from the DreamCoder ML system, which provides a safer alternative to current approaches to AGI.

Accomplishments

- **Completed the first release of our research software tool implementing the Safety-Performance Tradeoff model we built in collaboration with Robert Trager:** Together with Nicholas Emery-Xu, Eoghan Stafford, and Allan Dafoe, Robert Trager created an AI competition model called the Safety-Performance Tradeoff model. To allow other researchers and decision-makers to explore how improvements in technical AI safety could affect the safety choices of competing AI developers, we built a research software tool that showcases both analytical results and numerical results. After having built a first internal version during the first half of 2021, we extended and refined the research software tool for our first release in August 2021 as was desired by Robert Trager. At the beginning of the reporting period, the research software tool only contained three charts showcasing the analytical results of the model and functionality to allow users to enter custom parameters. In addition to showcasing the analytical results, we reproduced the numerical results and implemented functionality for automatically downloading and then showing them in the research software tool. Furthermore, we more than doubled the amount of currently available charts and we introduced sliders to ensure users always enter valid input values. To help the users understand and explore the Safety-Performance Tradeoff model, we added an explanation of the model, a definition of its parameters, three preset scenarios, and a summary of its key insights to the research software tool. To complete the first release, we also adapted the current layout and design to accommodate the additional functionalities and information. During the course of this project, team member Paolo Bova contributed substantially to the further development and analysis of the model, and so was invited to become the second author of the corresponding working paper.
- **Started our collaboration with Shahar Avin to implement software to facilitate his and his team's Intelligence Rising workshop:** This workshop allows participants to explore AI governance scenarios with other teams as they advance through an AI technology tree. As a first step, we wrote a Monte Carlo

simulation in a notebook to represent the rules of the workshop game as code. This allowed us to analyze the game and find ways of increasing its robustness together with Shahar Avin. Afterward, we built an internal version of a software that will be used to conduct the workshops in the future. During this reporting period, we implemented functionality for the game administrators including creating, switching between, and deleting new games. Furthermore, we started implementing the technology tree allowing the administrators and users to keep track of the world state within the game representing which papers and products have already been researched and whether this information is public or private.

- **Revised and substantially extended our technical report *Safe Transformative AI via a Windfall Clause* about the evaluation of the Windfall Clause policy when applied to the model presented in the paper *Racing to the Precipice*:** During the previous reporting period, we found that in our model the Windfall Clause doesn't just encourage a safer competition for TAI but that it is also often in the firms' best interests to pledge a significant share of windfall profits to socially good causes. While our report examined the public information setting for the design of a Windfall Clause, we now succeeded in extending our findings to the private and no information settings present in previous works. Despite the challenges that private information presents for negotiating a Windfall Clause, we were able to find a simplification that allows us to outline results for when a Windfall Clause is viable. Thanks to these new results, we were also able to examine when disclosure of information is helpful or harmful to a Windfall Clause. Additionally, we incorporated some feedback that we had received from other researchers at the end of the previous reporting period.
- **Worked on and laid the foundations for future collaborations with The Anh Han as well as Robert Trager and their teams:** Together with The Anh Han and his team, we discussed multiple model ideas before experimenting with and implementing several evolutionary game theory models. Furthermore, we helped Nicholas Emery-Xu, a Ph.D. student of Robert Trager, with several proofs for the private information setting of his model and provided him with our code for solving the public information setting. Nicholas has used these results in an initial draft which he aims to submit in 2022.

Roadblocks

- **Having too few human resources available resulted in postponing project milestones in coordination with, but against the desire of, our collaborators**

and delaying the submission of improvements to the DreamCoder software implemented during the previous reporting period. Given the limited amount of human resources and the inefficiency of context switching, we decided to focus on one collaboration at a time. Accordingly, we had to postpone some of the project milestones in coordination with our collaborators even though they would have benefitted from an earlier delivery. Similarly, we needed to delay preparing and submitting the improvements to the DreamCoder software despite already having them implemented a few months prior. Given that multiple AI governance researchers expressed interest in collaborating with us, we would love to resolve this roadblock by hiring an additional team member and/or outsourcing specific tasks once our financial situation allows us to do so.

- **Not having a stable source of funding forced us to spend additional time minimizing recurring infrastructure costs to ensure the research software tools for our collaborators are less dependent on our financial situation.** We are in the privileged position to build research software tools with our collaborators that currently aim at facilitating a workshop about AI governance and displaying the results of established AI governance researchers. While this allows us to advance their core work, it also increases the importance of providing reliable software and infrastructure. However, uncertainty about future funding means that there is a constant risk that we would have to shut down our services in case of a lack of funding in the future. This is made more difficult by the current funding process which only covers a year at a time. Therefore, we had to spend additional time on cost-saving to reduce the monthly costs of keeping the research software tools running.

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

- **Polish and launch our research software tool implementing the Safety-Performance Tradeoff model we built in collaboration with Robert Trager:** Having completed the first release of our research software tool

visualizing the AI competition model created by Robert Trager, Paolo Bova, Nicholas Emery-Xu, Eoghan Stafford, and Allan Dafoe, it's time to get the word out. Together with Robert Trager, we plan to polish the research software tool and announce it, among other channels, on the Effective Altruism forum to allow other researchers and decision-makers to explore how improvements in technical AI safety could affect the safety choices of competing AI developers. A successful launch includes polishing the research software tool with regard to its ease of use—increasing accessibility, clarity, and design—and publishing blog posts elaborating on the project.

- **Release the first version of the software we implemented in collaboration with Shahar Avin to facilitate his and his team's Intelligence Rising workshop:** While we built an internal version of the software during this reporting period, we would like to release the first version of the software during the upcoming reporting period with the goal of user field testing, i.e. letting the team of Intelligence Rising test the software in their actual working environment when conducting a workshop. To ensure go-live readiness, we first need to finish the implementation of the login system, the game administration section, and the tech tree through which participants advance to explore AI governance scenarios with other teams. Then, we have to set up the infrastructure to run the application in a browser and synchronize the state of the game in real time between the game administrators and game participants. Additionally, we plan to give a presentation showing and explaining the functionality of the software to the game administrators.
- **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 applying it 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 this reporting period. 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 finalize the incorporation of our new insights into a revised version of our technical report 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 collaborations with AI governance academics to build tools for other researchers and policymakers to help build intuitions for competition toward the development of TAI:** 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 resulting 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 governance 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.