Progress report 2021-01 – 2021-06

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 AI race dynamics to look for opportunities to promote safety-enhancing cooperation among race participants. During the reporting period, we evaluated the Windfall Clause policy when applied to the model presented in the paper *Racing to the Precipice*. Our technical report shows that joining the Windfall Clause is often in the firms' best interests and encourages a safer race for transformative AI. Thus, we hope that our findings motivate other researchers and policymakers to prepare a Windfall Clause to increase the probability of safe development of TAI.

We also started a collaboration with Robert Trager to build a web app implementing the Safety-Performance Tradeoff model created by him, Eoghan Stafford, Nicholas Emery, and Allan Dafoe, and relaunched our website to properly reflect our work and community. Furthermore, 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 for which he was awarded a grant from SAF. Modeling Cooperation consists of seven team members corresponding to 2.2 FTE. While both 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 AI race dynamics using game theory and computational modeling to look for opportunities to promote safety-enhancing cooperation among race participants. We consider the mitigation and prevention of AI races high-impact because a race toward transformative AI (TAI) could strongly incentivize its participants 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 as well as Jonas full-time and 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 1,875 Swiss francs he earned as a freelancer outside of Modeling Cooperation during this reporting period to start building up a small runway and, thus, increase the job security for our researchers.

Tanja Rüegg has been continuing to support Modeling Cooperation part-time on a voluntary basis. Jasmine Brazilek (see biography below) joined our team at the beginning of the reporting period, after already supporting us in the early stages, and has been contributing part-time on a voluntary basis as well. Additionally, Vasily Kuznetsov has joined Miles Tidmarsh in being on hiatus. Overall, Modeling Cooperation consists of seven team members corresponding to 2.2 FTE. Due to limited financial resources, we adjusted our wages as follows: Until the end of 2020, all employed team members worked for $26/hour 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). Moving forward, this only applies to the employed part-time team member while both full-time team members receive fixed wages of $30,000 annually.

We would like to quickly introduce our newest team member, Jasmine. She has Honors degrees in chemistry and microbiology from Monash University where she

conducted research on Malaria diagnostic techniques. Today, she has experience as a technology consultant in areas such as DevOps and security. She is passionate about existential risk reduction and animal welfare.

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. Having been awarded a personal grant from Survival and Flourishing at the end of 2020, Jonas set aside the equivalent of two months of full-time work during the reporting period to focus on his proposed project "Using Bayesian ML to find more interpretable solutions to AI race models".

With this in mind, we decided to dedicate almost all of our remaining resources to further pursuing the evaluation of the Windfall Clause policy—an *ex ante* commitment by AI firms to donate a significant amount of any eventual extremely large profits—proposed by the Future of Humanity Institute. After spending some time investigating this line of research during the second half of 2020, we assessed it to be a promising research opportunity. Thus, we decided to apply the proposed policy to the game presented in the paper *Racing to the Precipice* by Armstrong, Bostrom, and Shulman with the aim of authoring a write-up discussing our results.

We decided to spend the rest of our time working toward transitioning from our outdated website consisting only of a brief introduction to our initial research team as well as links to our first model implementations to an up-to-date website that properly reflects our work and community.

As a result, at the beginning of the reporting period, we had already assigned all of our available resources to the three projects mentioned above which all paid into our short-term goals. But in May 2021, we were offered the unique opportunity to start collaborating with Robert Trager—an associate professor of international relations at UCLA and strategic modeling team lead at GovAI—to build a web app implementing an AI race model. Despite the resource scarcity, we decided to accept the project because of its perfect fit with our longer-term goal of collaborating with AI governance academics to implement tools to help build intuitions for AI races.

Accomplishments

- **Authored the technical report *Safe Transformative AI via a Windfall Clause* about our evaluation of the Windfall Clause policy when applied to the model presented in the paper *Racing to the Precipice*:** In our model, we found that the Windfall Clause doesn't just encourage a safer race 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. In fact, the scope for agreeing to a mutually beneficial Windfall Clause is surprisingly large since firms have a stronger incentive to join as the race becomes more dangerous. It can even be in the best interests of players to join a Windfall Clause when they dislike donating via the Windfall Clause more than letting another firm win the race (where dislike is captured as a parameter in our model). Finally, players are also more willing to commit to a Windfall Clause as they learn about their relative positions in the race. We hope that these encouraging findings motivate other researchers and policymakers to prepare a Windfall Clause as part of the policy toolkits to increase the probability of safe development of future AI systems.
- **Worked on making numerical optimal solutions of an AI race model interpretable by humans using MIT's new Bayesian program learning ML system DreamCoder:** In 2020, we extended the game presented in the paper *Racing to the Precipice* with a logit discrete choice model from the discrete choice literature in economics. While this allowed us to distribute the win probabilities among race participants based on how much progress toward TAI they have made instead of always having the race leader be the final winner, it was no longer possible to analytically derive optimal strategies for the race participants. Thus, we solved the model numerically under the public information scenario leading to a result consisting of a large number of data points that represent the optimal actions for samples from the whole parameter space. While one can plot the resulting many-dimensional space, it is difficult for humans to extract understandable concepts from such large amounts of data. Instead, humans need code (or math formulas) to be able to interpret the behavior of functions. In this project, Jonas worked on increasing the interpretability of these results by representing them as code using DreamCoder—a system that emulates how humans generalize from a handful of examples by coming up with short programs, or algorithms, and uses Bayesian program learning to build concepts compositionally from those learned earlier. While doing so, he was able to report the following five achievements:

- Built a new DreamCoder cloud infrastructure which was necessary to overcome the lack of support for running DreamCoder in the cloud for users outside of MIT.
- Created bug fixes for DreamCoder because some ways of running the software didn't work as expected for the use case of investigating the numerical optimal solutions of AI race models.
- Extended DreamCoder to output its resulting code not just as a PDF but also in the programming language Clojure allowing for investigating the solutions, testing their accuracy, and incorporating the resulting strategies into our AI race simulations in the future.
- Trained DreamCoder to represent the optimal strategies for seven different variations of an AI race model as symbolic functions that build on each other. Furthermore, DreamCoder extracted its own functions that represent shared concepts between the different strategies. The symbolic functions make it possible to read and interpret the optimal strategy data as mathematical functions.
- Became more familiar with ML and especially the Bayesian program learning approach used by DreamCoder which is especially relevant for Jonas' future research because of its grounding in cognitive science, focus on AGI, as well as its interpretability (which could also be applied to AI safety).

- **Started our collaboration with Robert Trager to build a web app implementing the Safety-Performance Tradeoff model:** Together with Eoghan Stafford, Nicholas Emery, and Allan Dafoe, Robert Trager created an AI race model called the Safety-Performance Tradeoff model. In our collaboration, we aim to build a web app that allows other researchers and decision-makers to explore the aforementioned model and investigate a wider range of parameters than would be possible analytically. So far, we have built the first internal version of the web app with three charts showing the analytical results of the model and allowing the users to enter custom parameters.
- **Relaunched our website www.modelingcooperation.com:** Up until now, our online presence was limited to the website we launched in 2019 consisting of only a brief introduction to our initial research team as well as links to our first model implementations. Since then, Modeling Cooperation has grown significantly and we wanted to build a website that properly reflects our work and community. While implementing our new website, we focused on enabling the users to achieve the following goals: getting a first impression of Modeling Cooperation, looking into our research as well as research support software in more detail, staying up to date with the work we are doing, and getting in touch

with us. Additionally, we ensured that the website is GDPR compliant because our target audience includes EU residents.

Roadblocks

- **Having too few resources available increased the risk of us not meeting the project deadlines, decreased our efficiency due to context switching, and forced us to postpone an additional collaboration as well as important administrative tasks.** While we couldn't pass up the unique opportunity of collaborating with the established AI governance academic Robert Trager, it amplified the scarcity of our available resources. Meeting those unexpectedly tight project deadlines was only possible due to the tireless commitment of our team members and even then it was only a matter of days. In addition, our researchers had to focus on multiple projects at once and didn't have enough time to fulfill all of their other responsibilities. This roadblock could be resolved by additional financial resources which would allow us to hire an additional team member and/or outsource specific tasks.
- **The limitations of MIT's new ML system DreamCoder made for a difficult start to the project "Using Bayesian ML to find more interpretable solutions to AI race models".** Even though DreamCoder has good documentation on how to get started and works well for the use cases provided by the researchers, Jonas early on had to fix issues in DreamCoder and add extensions to it to be able to use the software for his use case. While this was a great learning opportunity and required him to understand the internals of DreamCoder and how it is implemented, it decreased the time available for interpreting the results.

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

- **Further pursue our collaboration with Robert Trager with the goal of completing the first release of our web app implementing the**

Safety-Performance Tradeoff model: After building the first internal version of the web app, we plan to complete the first release in August 2021. In addition to showcasing the analytical results, we aim to reproduce the numerical results and implement functionality for automatically downloading and then showing the results in the web app. Furthermore, we plan to more than double the amount of currently available charts and limit the custom parameters users can enter to valid values. To help the users understand and explore the Safety-Performance Tradeoff model, we plan to add an explanation of the model, a definition of its parameters, three preset scenarios, and a summary of its key insights to the web app. To complete the first release, we also need to adapt the current layout and design to be able to accommodate the additional functionalities and information.

- **Start our collaboration with Shahar Avin where we plan to implement software to facilitate his and his team's Intelligence Rising workshop:** This workshop allows participants to explore AI governance, and in particular AI races, with other teams. We aim to implement the workshop's technology tree in a web app and additionally write a Monte Carlo simulation of the game in a notebook to find ways of improving robustness and help participants more deeply explore the consequences of their choices during simulation games like Intelligence Rising. The format also opens up opportunities for exploring how access to different tools can influence the decision-making of participants, who may hold key positions relevant to AI governance in the future.
- **Submit DreamCoder improvements implemented as part of the project "Using Bayesian ML to find more interpretable solutions to AI race models":** Having been awarded a personal grant, 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. Building on the achievements during this project, Jonas would like to contribute to the DreamCoder software 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 will help the researchers at MIT as well as other researchers outside of MIT to benefit from the DreamCoder ML system.

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 race 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:** 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 race 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 potentially 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.
- **Refine the interpretable strategies gained as part of the project "Using Bayesian ML to find more interpretable solutions to AI race models" and incorporate the solutions into our AI race simulations:** Having been awarded a personal grant, 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. Building on the achievements during this project, 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 race model 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 race 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 races, such as game theory and industrial organization, which we could write up tailored to other AI race 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.