Progress report 2020-07 – 2020-12

Modeling Cooperation

Paolo Bova, 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 primarily focused on our Baseline project which has our AI race model at its center. As part thereof, we iterated on our model by extending it with additional mechanics, released version 1.0.0 of our Monte Carlo simulation tool which allows anyone to run simulations of our model, and created a new and improved proof showing that our dynamic Baseline game can be reduced to the static game presented in the paper *Racing to the Precipice* (RTTP).

Additionally, we applied the Windfall Clause policy to a probabilistic version of the game presented in RTTP and have begun investigating which Windfall Clauses are rational for firms to join in this model. Modeling Cooperation consists of six team members corresponding to 2.2 FTE. For approximately 25% of the reporting period, three of the team members worked for $26/hour on a contract basis. During the remaining time, the team continued to work on a voluntary 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 2020 with the support of two grants—$20,000 from the Center on Long-Term Risk Fund (previously the Effective Altruism Foundation Fund) and $50,000 from the Survival and Flourishing Fund. After having employed five team members corresponding to 2.3 FTE during the first half of the year, the funding was nearing its end shortly before the beginning of July.

The team decided to continue its work on a voluntary basis to keep some money in reserve which was used in November to employ Paolo Bova and Jonas Emanuel Müller full-time as well as Vasily Kuznetsov part-time. During the reporting period, Ben Harack and Tanja Rüegg have been supporting the project part-time on a voluntary basis while Miles Tidmarsh has been on hiatus after starting his economics Ph.D. Overall, Modeling Cooperation consists of six team members corresponding to 2.2 FTE and all employed team members have been working for $26/hour on a contract basis. This means they only get paid when working and not when anything prevents them 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. In addition to advancing our Baseline project, we had identified a promising analysis opportunity during the first half of 2020: the examination 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

when applied to a probabilistic version of the game presented in the paper *Racing to the Precipice* (RTTP) by Armstrong, Bostrom, and Shulman.

We decided to continue to put our primary focus on our Baseline project because it seemed valuable to advance and build upon our previous work. In addition, we decided in favor of spending some time investigating the potential analysis of the Windfall Clause policy when applied to a probabilistic version of the game presented in RTTP. Since we consider the Windfall Clause a promising approach when it comes to reducing risks from AI races and we were already familiar with the relevant game, we assessed this to be a research idea worth exploring.

The goal of the Baseline project is to build an AI race model and analyze it using techniques from game theory, economics, and computational modeling. To achieve this goal, we decided to improve three elements of our project:

1. Re-evaluate the current model mechanics by assessing additional write-ups and papers within the fields of economics (with a focus on industrial organization), technological races, as well as AI safety and build upon our experience by improving the model based on the lessons we learned while analyzing it during the first half of 2020: This iteration on our AI race makes sure it builds more on existing academic research while also increasing its suitability for finding optimal solutions computationally.
2. Revise our proof showing how our dynamic game can be reduced to the static game presented in RTTP: This revision provides a deeper economic intuition as to why Baseline is an extension of the static game in RTTP.
3. Release version 1.0.0 of our Monte Carlo simulation tool: This tool makes it possible for anyone comfortable with command-line applications to run Monte Carlo simulations of our model and, in the long run, reproduce our results.

Accomplishments

Most of the following accomplishments are part of our Baseline write-up. It is the centerpiece of our Baseline project since it lays the foundations of our work by defining our AI race model and showing its relationships to other prominent models in the literature.

- **Iterated on our Baseline AI race model:** During the first half of 2020, we analyzed the reduced version of our Baseline model. As part thereof, we identified the removal of the catch-up mechanics and luck in production as the

steps necessary to increase the game's suitability for finding optimal solutions computationally. In addition to excluding these mechanics, we extended our model and write-up with the following mechanics and details to build more on existing academic research:

- **Different degrees of advantage that R&D investment brings for developing TAI:** This allows us to model everything from races that are a level playing field and where even small organizations have a shot at developing TAI to races where whoever invests the most in R&D certainly wins. A new parameter is used to specify the degree of advantage that R&D investment brings.
- **The depreciation of research knowledge over time:** This mechanic introduces that firms need to continue investing in capability to maintain their chance at winning the race. By changing the corresponding parameter, it is possible to specify how much a firm's knowledge stock depreciates each round.
- **A regulator whose goal it is to minimize damages:** We now model the goal of the AI governance community/regulators as utility maximization for a regulator who is part of the game. To make this possible, we also formalized the concept of damage, which quantifies the damage from different disaster scenarios. The goal of the regulator is to minimize this damage.
- **A market share mechanic:** This allows us to model different types of market share distributions at the end of the race—from being a winner-takes-all market to one where everyone gets a part of the resulting market share.
- **A hazard rate mechanic based on industrial organization literature:** Agents now have a probabilistic win condition which means they never know for sure in which round they will achieve TAI. Each firm's probability of winning the race is calculated every round using a logit discrete choice model which we derived from assuming consumers who have preferences over TAI services.
- **A disaster spillover mechanic:** This allows us to model different degrees of disaster—from being local to the firm that caused the disaster to affecting all firms equally.
- **More details about the grounding of our model mechanics in literature:** We now have explanations for how our new mechanics are grounded in existing literature and have added an initial section comparing our model to all other existing race models.

We expect to be able to build on the research results of the reduced version of our Baseline model to find optimal solutions for the improved model.

Furthermore, we plan to use this model to analyze the Windfall Clause policy in the future.

- **Created a new and improved version of our proof that the Baseline game can be reduced to the game presented in RTTP:** During the first half of 2020, we proved that the static game presented in RTTP is a special case of our dynamic Baseline game. To show this equivalence, we solved for the optimal Baseline value function and showed that it reduces to the RTTP payoffs. We then recovered the original static game presented in RTTP when everyone is guaranteed to discover TAI in the first period. We have since revised our approach to the proof. We instead reduce Baseline to a repeated game which is payoff equivalent to RTTP. The elegance of this approach means that we can show that the crucial difference between Baseline and the model in RTTP is the accumulation of R&D. The two models only become equivalent when we assume firms are unable to accumulate R&D knowledge over time. This framing also serves to distinguish our models from other recent models of AI races which do not allow R&D to accumulate over time.
- **Released version 1.0.0 of a Monte Carlo simulation tool to run simulations of our Baseline model:** We initially built our Monte Carlo simulation tool bl.cli to make it easier for our economists to produce data to quickly compare race results under heuristic strategies. The tool streams the results to a file which can then be loaded into R, Python, and many other programming languages since bl.cli supports multiple output formats. We then turned it into an easily usable tool that allows anyone comfortable with command-line applications to run Monte Carlo simulations of our model and, in the long run, reproduce our results. To do so, we ensured that the tool can be run without any prerequisites on any major operating systems. Despite requiring more work, we deliberately chose a way to build and deliver the application that can be expected to work and be supported for one or more decades to come. We see building research software tools that still work decades from now as an important step toward more reproducible science. Therefore, we also created a write-up elaborating on our approach to implementing models and addressing issues of reliability, composability, and sustainability in computational science.
- **Began evaluating the Windfall Clause policy when applied to a probabilistic version of the game presented in RTTP:** The Windfall Clause is a policy proposal for an ex-ante commitment by AI firms to donate a significant amount of any eventual extremely large profits garnered from the development of TAI. We applied the Windfall Clause to our probabilistic version of the game presented in

RTTP. We then began investigating which Windfall Clauses are rational for firms to join in this model.

Roadblocks

- **Due to financial constraints, the team members had to work on a voluntary basis for most of the reporting period and had to invest a significant amount of time to secure additional funding.** In 2019, Jonas Emanuel Müller, Vasily Kuznetsov, and Miles Tidmarsh laid the foundations of Modeling Cooperation at the AI Safety Camp. Shortly thereafter, we welcomed additional team members and started applying for grants to employ multiple researchers for Modeling Cooperation on a contract basis. Despite being a new research team, we were able to secure 60% of our budget for 2020. This resulted in the funding nearing its end shortly before the beginning of July. Therefore, the previously employed team members had to work on a voluntary basis for most of the reporting period. Additionally, this led the team to prioritize applying for funding for multiple weeks to increase the likelihood of securing funding for the entire year of 2021. During this time, Modeling Cooperation worked on putting their track record in writing, creating a Theory of Change as well as a Logical Framework, and incorporating feedback from stakeholders we received along the way.
- **We could have benefitted from a person with experience in hiring international contractors as well as familiarity with employment laws in the UK and Switzerland.** Having a person with international contractor agreement experience on our team would have allowed us to move our hiring process along more quickly. Instead, we had to contact local lawyers which was quite resource-intensive and led to contradicting information between local lawyers in different countries in one instance.

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

- **Author a paper about the evaluation of the Windfall Clause policy when applied to the game presented in RTTP:** After spending some time investigating the potential analysis of the Windfall Clause policy when applied to a probabilistic version of the game presented in RTTP, we consider this to be a promising research question. We aim to further pursue the evaluation of the Windfall Clause policy and to author a write-up discussing our results.
- **Use Bayesian ML to find more interpretable solutions to AI race models:** Jonas Emanuel Müller plans on translating Modeling Cooperation's numerical optimal solutions, which aren't interpretable by humans, to understandable concepts using MIT's new Bayesian program learning ML system. Currently, these solutions are captured in radial basis functions (RBFs) as is common in such a case and ML more generally. While one can plot the resulting n-dimensional space, it is difficult for humans to extract understandable concepts from function approximations like RBFs. Therefore, Jonas will use MIT's new Bayesian program learning ML system DreamCoder to translate the optimal solutions into a symbolic program using Bayesian program learning. For this project, Jonas has been awarded a grant from Survival and Flourishing (SAF).
- **Relaunch the Modeling Cooperation website:** In 2019, we launched our current landing page which contains a brief introduction to our research team as well as links to our model implementations. Since then, Modeling Cooperation has grown significantly and we aim to build a website that properly reflects our work and community. The website should allow anyone to get a first impression of Modeling Cooperation, look into our research as well as research support software in more detail, stay up to date with the work we are doing, and get in touch with us. Additionally, we want to make sure the new website is GDPR compliant.

Longer-term goals

- **Further investigate the Windfall Clause policy when applied to a model with more parameters than the one presented in RTTP:** 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.

- **Build a user-friendly UI as a research support software tool to help build intuitions for AI races:** When discussing the impact of potential research directions, our stakeholders repeatedly expressed interest in projects which enable other researchers 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 UIs based on AI governance research results. To explore this idea, we would like to collaborate with at least one of our stakeholders with the goal of transforming either an AI race model or AI development scenario workshop into a user-friendly UI.
- **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.