Progress report 2020-01 – 2020-06

# 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 solely focused on our Baseline project and, as part thereof, expanded our scope to include more game theory and investigate numerical solutions in addition to an agent-based modeling approach. We formalized and implemented an AI race model that combines best practices from game theory, economics, technological races, and AI safety literatures. We built our dynamic game so that it can be reduced to the static game presented in the paper *Racing to the Precipice* and have proven this to be the case.

Furthermore, we solved a reduced version of our model using numerical methods. We also developed a Monte Carlo simulation tool that enables our economists to analyze our agent-based model. Despite the pandemic taking place during the reporting period, we didn't have to slow down our core work. Modeling Cooperation consists of six team members corresponding to 2.3 FTE. All employed team members work 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 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 Effective Altruism Foundation Fund) and $50,000 from the Survival and Flourishing Fund. Since then, Paolo Bova and Jonas Emanuel Müller have been employed full-time while Ben Harack and Vasily Kuznetsov have been employed part-time. Miles Tidmarsh was employed full-time until he began his economics Ph.D. in March. Furthermore, we recruited a part-time volunteer project manager, Tanja Rüegg. Overall, Modeling Cooperation consists of six team members corresponding to 2.3 FTE and all employed team members work 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). In total, close to 80% of the funding has been spent during the reporting period.

- **Paolo Bova** is a recent Cambridge graduate with a bachelor's degree in economics. During his studies, he specialized in theoretical courses and wrote his dissertation on foundations for ethical AI which mixed network and set theory to characterize ethical and strategic interactions between meta-learners. He aims to extend his dissertation into a publishable paper.
- **Ben Harack** conducts research on the governance of transformative technologies that may exhibit race dynamics such as AI. After presenting the poster “Governing the emerging risk posed by asteroid manipulation technologies” at EAG 2019, he received (private) cooperation offers from researchers at CSER and GCRI. He also lead-authored “Ruling ourselves: The deliberate evolution of global cooperation and governance”—a semifinalist for the $5 million New Shape Prize.

- **Vasily Kuznetsov** has a master's degree in mathematics and over a decade of experience in software engineering. While working for the UN Climate Change Secretariat, he was the lead developer for adding support for Program of Activities to its Clean Development Mechanism information system. He currently works for eyeo GmbH, a company that aims to fix the monetization models of the web.
- **Jonas Emanuel Müller** has been involved in EA since 2012. For several years, he pursued earning to give as a software engineer and was a Scrum Master at a major Swiss bank. During this time, he donated 50%–70% of his annual salary and promoted both EA and earning to give through his appearance in numerous newspaper articles, TV reports, and radio programs. He has been on the board of multiple effectiveness-focused organizations and is currently on the board of Animal Charity Evaluators (ACE). He was the chair of the board of ACE for four years and was the lead for its executive director recruitment in 2018.
- **Tanja Rüegg** is earning to give as a project manager at a data science startup, where she is the project lead for five client projects while also being responsible for both the marketing and the customer service department. Previously, while working at the Effective Altruism Foundation, she successfully coordinated an initiative for an impactful law change affecting 400,000 people.
- **Miles Tidmarsh** is a Ph.D. candidate in economics who worked as a research economist for the Productivity Commission, a think tank advising the Australian government. He has conducted multiple literature reviews which resulted in published material, developed published recommendations for government action, and wrote chapters explaining the issues and justification for recommendations. Furthermore, he has replicated and disproved published papers and used hundreds of robustness checks on panel data.

## Strategy

After examining our planned work in light of our financial and human resources, we decided to put our sole focus on our Baseline project because the money granted by the Center on Long-Term Risk Fund was restricted to this type of research. This project aims to:

- Build an AI race model and analyze it using techniques from game theory, economics as well as computational modeling to find (multiple) equilibria.
- Test several hypotheses to investigate how different mechanics affect safety.

- Compare the expected safety under the derived strategies to the results from the *Racing to the Precipice* (RTTP) publication by Armstrong, Bostrom, and Shulman.

Overall, we are looking for ways to promote safety-enhancing cooperation among race participants and to turn this research into one game-theoretic and one computational write-up.

Accordingly, we needed to deprioritize our efforts to turn computational models into interactive simulations for researchers to explore AGI scenarios. Prior to its deprioritization, this workstream was motivated in part by the interest that was expressed by a workshop organizer in having us turn his scenario model into an application to increase his workshop's efficiency and thus amplify its impact.

To ensure that we work on the tasks contributing to our overall goal, we follow a process to map all of our tactical work to our strategic plan:

1. Our strategic plan defines our strategy for reducing risks from AI races.
2. We use the strategic plan to derive our project goals.
3. We use the project goals to identify our stakeholders.
4. We map the interests of those stakeholders to our project releases and user stories (collections of tasks) in our user story map.
5. We use the user story map to prioritize and estimate current and upcoming stories in our backlog.
6. We map each of the stories in the backlog to one or more GitLab issues.
7. Each GitLab issue is implemented by a team member, by default reviewed by another, and, where valuable, reviewed/approved by the whole team.

## Accomplishments

Since our goal is to turn our research into one game-theoretic and one computational write-up, our current Baseline write-up which describes both lines of work for now, is the centerpiece of our Baseline project. Because all of our work progresses through the process described in the strategy section, not all of the following accomplishments are part of the write-up yet.

- **Iterated on our Baseline AI race model:** We built an AI race model intended to be more representative of real-world AI development dynamics than the ones investigated in prior work in the AI race literature. Therefore, our goal was to combine best practices from game theory, economics (especially industrial organization), technological races, and AI safety literatures. After careful

deliberation, we decided to include production functions with decreasing marginal returns and catch-up mechanics into a game with multiple rounds that allows for strategic interactions between agents. Additionally, we included a set of win conditions as well as disaster calculations that are representative of the current discussions in AI policy.

- **Proved that the Baseline game can be reduced to the RTTP game:** The RTTP game is a static game whereas the Baseline game is a dynamic game. Therefore, showing that the RTTP game is a special case of the Baseline game is more involved than setting the values of Baseline's parameters to RTTP's parameters. Instead, we first solved for the optimal Baseline value function and then showed that it reduces to the RTTP payoffs. Moreover, we recover the original static RTTP game when everyone is guaranteed to discover TAI in the first period. Hence, we proved that the RTTP game is a special case of our Baseline game. This equivalence also helps us discuss how weakening the assumptions implicit in RTTP influences their safety findings.
- **Compared analytical methods for solving dynamic AI race games:** As part of our analytical research, we gained valuable insights into the suitability of different solution methods when applied to AI race models. While there are numerous techniques available for solving dynamic games, most have not yet been applied to the special case of AI races. Therefore, we investigated several such approaches. For each method, we summarized the corresponding literature and our specific utilization as well as the pros and cons relative to other methods.
- **Created a Monte Carlo simulation tool:** We implemented a production-quality version of our agent-based model and created a heavily parallelized Monte Carlo simulation tool enabling our economists to quickly compare race results under heuristic strategies. In addition to increasing productivity, this tool can be reused in the future for analyzing new models.
- **Found an optimal strategy for a reduced version of the Baseline model:** First, we refined the Baseline model using the industrial organization literature. These changes allowed us to compute an equilibrium of a reduced version of the model. So far, we have computed the two-player equilibrium for a subset of win conditions and disaster calculations. In particular, we solve for the Markov perfect equilibrium using the iterative method for solving dynamic games introduced in Cai et al. 2018. To help ensure convergence of the algorithm, our implementation uses radial basis functions when fitting our value function and the optimal policy.

## Roadblocks

- **Because of the pandemic, we had a reduced amount of resources available.** Ben had to take several weeks off to take care of his young daughter between March and May. Therefore, we needed to invest almost all of our remaining resources in research, which allowed us to move our core work forward without slowdown. During this time, we couldn't move our hiring process along as quickly as planned and we didn't reach out to our stakeholders and potential collaborators as regularly as we would have liked.
- **We could have benefitted from an additional computational economist.** Having an additional economist with computational economics expertise on our team would have allowed us to more quickly review the work exploring more advanced numerical solution methods.

## 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

- **Find additional optimal strategies for the reduced Baseline model for all information conditions described in RTTP:** In addition to our current results, we aim to further check for (multiple) equilibria in our reduced Baseline model. Furthermore, we would like to compare the expected safety under the derived strategies to the RTTP results.
- **Test the following hypotheses in our Baseline model:**
  - The catch-up term tends to make races more dangerous.
  - Safety being relatively more difficult tends to make races more dangerous.
  - Less variance for luck in production tends to make races safer.
  - The hard takeoff win condition makes races more dangerous while the soft takeoff win conditions make races safer.

- **Extend our write-up with the following details:**
  - Literature review of all current AI race models as well as economic races in industrial organization literature
  - More details about the grounding of our model mechanics in literature
  - A hazard rate mechanic based on industrial organization literature
  - More thorough heuristic strategies and a discussion of their effectiveness as well as the resulting consequences for safety
  - A discussion of the Markov perfect equilibria of a reduced version of our Baseline model
  - A discussion of how the Markov perfect equilibria of a reduced version of our Baseline model compare to the RTTP equilibria
  - A discussion of how disaster results under different information conditions compare to those of RTTP
  - A discussion of the tested hypotheses

## Longer-term goals

- **Turn our notes comparing analytical methods to solve dynamic AI race models into a series of blog posts:** Because most of the numerous techniques available for solving dynamic games have not yet been applied to the special case of AI races, our insights could help other researchers to decide on a suitable approach without them having to try out the various approaches themselves.
- **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.
- **Finish our review of literature relevant to AI races and turn it into a series of blog posts:** Paolo already completed a part of a review of literature relevant to AI races, such as game theory and industrial organization. We would like to finish reviewing the most valuable of the fields and turn the often technical and complex literature into a series of blog posts tailored to other AI race researchers.