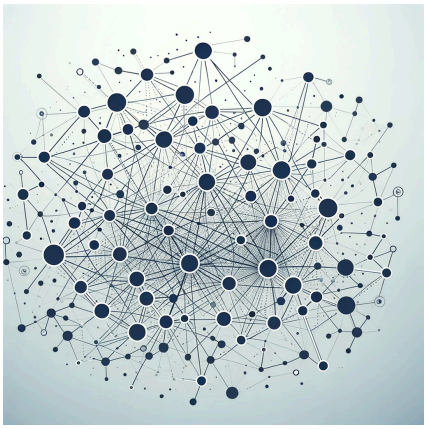
AI Safety Grant Awarded to TFI Research Project

We are pleased to announce an AI Safety Grant has been awarded by The Foresight Institute to Dr Toby D. Pilditch, the Senior Research Scientist at The Transformative Futures Institute for a project entitled "Cutting through the complexity of multi-agent AI scenarios: A computational tool." The project will begin in January 2024, and will be carried out over the course of the year.

The project will develop a multi-agent computational model, enabling the simulation of AI development races, and capturing key complex dynamics, such as the interdependent and game theoretic behaviours between and among groups of actors (e.g., developers, users, governments). This model will significantly advance our capability to project timelines and analyze outcomes in various potential AI scenarios, with an immediate application focus on understanding and projecting multi-polar scenarios.

Currently, AI safety research lacks a robust computational approach to address complex, multi-polar AI scenarios. Discussions on multi-polar AI scenarios are often limited to theoretical arguments, lacking computational realization of their complexities. While some existing frameworks (e.g., general morphological analysis by Kilian et al., 2023; AI Act risk assessment model by Novelli et al., 2023) aggregate expert opinions, they fall short in comprehensively realizing the complexity of these scenarios. Similarly, game theoretic models often simplify scenarios to two-player interactions, missing out on the richness of multi-agent dynamics.

As a result, we are currently only able to speculate about potential multi-polar AI scenario outcomes, how likely such scenarios may be (and how quickly), and critically, what can be done to change these projections. Critically, this means we are currently blind to how our current trajectories might intersect with these scenarios. Moreover, we are missing vital opportunities to understand the underlying causal mechanisms, such that policies can be designed that not only mitigate risk from the most harmful scenarios, but also increase our likelihood of flourishing futures.



This project seeks to provide clear, quantitative answers to these questions (and more). The developed model will focus on the computational dynamics of complex, interactive multi-agent systems. This method allows for a rigorous examination of causal relationships and potential scenarios in AI development and safety. By incorporating game theoretic behaviors and modeling the decision-making processes of individual agents, the project will enable a nuanced understanding of multi-polar scenarios and principal-agent problems as they evolve over time.

Keep an eye out for more announcements as the model is developed and refined in the coming months. Our thanks once again to The Foresight Institute for their generous grant.