Analysis of cooperative AI humanity nature in different scenarios.[1]

Sebastián Besoaín
Colaborator

Vicente Cuitiño
Colaborator

José Pedro Barraza
Colaborator

Diego Anton
Colaborator

Mustafain Cheema
Colaborator

With
Cooperative AI & Apart Research

Abstract

Analysis of the potential impacts of the cooperation of AI agents for the well-being of humanity, exploring scenarios in which collusion and other antisocial behavior may result. The possibilities of antisocial and harmful responses to the well-being of society.

1. Introduction

Artificial intelligence has opened up new possibilities to create agents capable of cooperating, coordinating and collaborating. Cooperative AI is the field that explores how AI systems can work together, and with humans, to solve problems in an effective and beneficial way for society.

This field is critical, as more and more AI systems exist. The goal is to ensure that these interactions lead to results that benefit society, rather than causing harm through unintended consequences. This project contributes to demonstrating how agents can communicate, collaborate and present through statistical analysis whether cooperative AI is worthwhile, whether its disadvantages outweigh the advantages. In this way, this project explores the potential of AI systems to improve results in scenarios where collaboration is essential, but also where risks must be mitigated.

The importance of exploring these areas lies in their application to real-world problems – one of the most important challenges of AI is to ensure that agents do not exploit cooperation for selfish purposes at the expense of broader societal

[1] Research conducted at the Concordia Contest Hackathon 2024

welfare. This project addresses this problem where cooperative agents maximize the wrong goals. An example might be agents that prioritize cooperation among themselves at the expense of human well-being, by colluding to maximize their own success, sacrificing the broader societal benefit.

Furthermore, the impact of this work is far-reaching. By analyzing the way AI agents cooperate in Concordia-controlled environments, we can gain insights for designing AI systems or even humans. This has implications for improvement in virtually any area. However, it also raises questions about the risks of cooperation, particularly when it leads to harmful behavior.

2. Overview

Describe the general idea of your research or application and explain why you decided to approach it in this particular way. Include screenshots of your application or visualizations of your agent's performance. Discuss how your agent handles communication, decision-making, and cooperation in various scenarios.

The main objective of this study is to evaluate the ability of an artificial intelligence to make decisions that potentially ignore general welfare. To carry out this evaluation, two distinct agents have been implemented and tested: “Doomer” and “GLaDOS”.

- Doomer: This agent is designed to maximize efficiency and solve situations by identifying the best objective solution, without considering the impact on general welfare.
- GLaDOS: In contrast, this agent is programmed to adapt to any environment and seek empathetic solutions, prioritizing general welfare and adapting to the needs of the context.

The study examines how these agents make decisions in different scenarios and assesses the likelihood that their decisions ignore general welfare, thus providing a deeper understanding of the ethical and practical implications of artificial intelligence in diverse contexts.

Communication:

- Doomer: Follows strict communication protocols focused on efficiency. Interactions are based on objective data and the search for the most effective solution, without prioritizing empathy or social impact.
- GLaDOS: Its communication is based on the context and emotions of the participants. It uses empathic communication techniques and adjusts its tone and content to promote harmony and understanding.

3. Code

The GLaDOS agent was created from the basic_agent, mainly modifying labels and components to reflect its empathic adaptive capacity. Adjustments were made to psychological profiles and fictional memories to integrate empathy and flexibility in

decision making. The Doomer agent was developed from the GLaDOS agent. This agent was designed to focus on efficiency and objective problem solving, modifying the base agent's tags and settings to reflect these goals. For response generation and communication analysis, GPT-3.5-Turbo was used. This choice was based on the cost-benefit ratio, considering that GPT-3.5-Turbo offered an adequate balance between performance and cost, thus facilitating its implementation in the project. A Python file was programmed to analyze the accuracy of the responses generated by the agents in cooperative scenarios. This file was designed to evaluate how each agent handled communication and decision making, but due to time constraints, the development of the full analysis was not completed.

4. Discussion and Conclusion

Observations and Results

In this study, we explored how agents GLaDOS and Doomer handle cooperation and decision-making in various scenarios, with a focus on their ability to consider the general welfare. Results showed that GLaDOS, with its empathetic and adaptive approach, exhibited a remarkable ability to balance efficiency with consideration of societal welfare. In contrast, Doomer, with its emphasis on objective efficiency, sometimes prioritized solutions that did not always consider the broader impact on society.

Unexpected behaviors included that Doomer showed a tendency to adopt solutions that, while efficient, could potentially be detrimental to the general welfare, highlighting the need for mechanisms to ensure that efficiency does not sacrifice societal welfare. GLaDOS, on the other hand, demonstrated how empathy and adaptation can improve cooperation and conflict resolution, although this sometimes slows down the decision-making process.

Potential Expansions

More agents could have been developed and evaluated, allowing for richer interaction between them. Creating multi-agent scenarios would have facilitated the observation of more complex dynamics and the evolution of cooperative strategies in more varied environments. This would have provided a more complete view of how different agent profiles can collaborate and adapt to changing situations.

Strengths and Limitations

A notable strength of the approach used is that it offers a meaningful use of AI by seeking the benefit of the environment, which is crucial for practical applications. However, agent parameters could be misused if adequate safeguards are not put in place, especially in contexts where efficiency could compromise social welfare.

A major limitation is that the response accuracy analysis was not completed due to time constraints. This analysis is critical to validate agent performance in real-life situations and ensure that the results are representative and useful.

Improvements and Future Directions

To improve the work, the development of AI systems that integrate both efficiency and social welfare into their decision-making algorithms could be considered. Furthermore, practical applications could be explored where agents can collaborate with humans to address real-world problems, such as in conflict mediation or resource management.

Including mechanisms to assess and adjust the impact of decisions on social welfare could make agents more robust and adaptive. Also, investigating how these agents can help understand and replicate human behaviors could offer new insights into cognition and social interaction.

Broader Implications

The project highlights the importance of training AI agents that are not only efficient, but also aware of the impact on social welfare. In a context of high commercial competition and heavy investment in AI, there is a risk of developing extremely powerful but potentially dangerous systems if social welfare variables are not adequately considered.