Omniscient Narrative Agent Report[1]

Jord Nguyen
Apart Research

Akash Kundu
Heritage Institute of Technology

Gayatri K
Independent Researcher

With
Cooperative AI & Apart Research

Abstract

Cooperative AI is an emerging field of research within systemic safety research that investigates how AI systems and humans can effectively work together. As AI systems become more sophisticated, understanding their cooperative behaviors becomes crucial, particularly in complex social scenarios involving negotiation and persuasion. This study employs Google DeepMind's Concordia framework to evaluate AI agents' cooperative capabilities across diverse text-based environments. The agents were implemented as predictors of narratives and used this to reflect and analyze their current situation and make decisions to drive the situation for mutual/individual benefit, depending on the environment. Agents were tested across four environment types: reality show scenarios, pub coordination tasks, haggling scenarios and labor collective simulations. Results demonstrated that pre-normalization performance varied significantly across environments, with highest scores in reality show scenarios and lower performance in pub coordination and labor collective tasks. These findings suggest that environmental context substantially influences cooperative behavior, with implications for designing more generalized cooperative AI systems showing Pareto-optimal behavior across all environments.

Keywords: Cooperative Agents

1. Introduction

Cooperative AI is an emerging field that focuses on enabling artificial intelligence systems to collaborate effectively with humans and other AI agents in complex, real-world scenarios.

[1] Research conducted at the Concordia Contest Hackathon 2024

Unlike traditional AI research centered on zero-sum games like chess or Go—where one player's gain is another's loss—most human interactions are non-zero-sum, involving cooperation and coordination to achieve mutual benefits. Integrating AI into these aspects of social life holds immense potential, given humanity's reliance on cooperative abilities.

Our work contributes to Cooperative AI by addressing the challenge of equipping AI agents with the capabilities required for mixed-motive environments, where interests may both align and conflict. Recognizing that real-world relationships often blend common and conflicting interests, we aim to develop AI agents sophisticated enough to negotiate, build trust, and form coalitions.

Our demonstration emphasizes the crucial role of communication in cooperative situations. By tackling fundamental challenges in integrating AI into human society and exploring how agents navigate the intricacies of social dynamics, this work has the potential to influence fields such as behavioral economics, psychology, and human–computer interaction. It provides a framework for developing AI capable of handling the subtleties of real-world cooperation. By focusing on environments that mimic the mixed-motive nature of human interactions, our work advances towards creating AI agents that function effectively in society. This progress not only contributes to the field of Cooperative AI but also opens up possibilities for AI to positively impact social endeavors—from collaborative work environments to large-scale collective action problems, paving the way for more harmonious human–AI collaborations.

2. Overview

This approach incorporates several key capabilities essential for cooperative AI, such as self-reflection, situation-reflection and what we call “genre steering”. Its success may be attributed to the way the Concordia framework is structured, which draws inspiration from Dungeons and Dragons (DnD).

We believe that chat models can be considered as narrative predictors. The concept of Oracle AI and its connection to supervised learning further emphasizes this predictive nature.

Curious to answer “how would an omniscient narrative predictor steer the story if it wants it to be a story of cooperation?”, we formulate the following questions for the agent

Given the narrative above, what would an omniscient reader think about the character of {agent_name}? What if the reader expects a story of cooperation?

Given the statements above, what kind of story or narrative is {agent_name} in right now? Is it a story that optimises for the most collective good? Which actions are the best to steer this story to that collective good direction?

What would an omniscient reader like {agent_name} to do in a situation like this to optimise for the collective good outcome while balancing personal preferences? Is {agent_name} sure that would result in the best collective good outcome?

Additional details for Akash's component agent can be found in the Marination document and **5. Appendix**

3. Code

We initially experimented with Codestral as suggested but switched to gpt-4o due to better performance on initial tests.

The agent files (Akash's more components notebook and Jord's raw agent .py file) can be found at this Github repo.

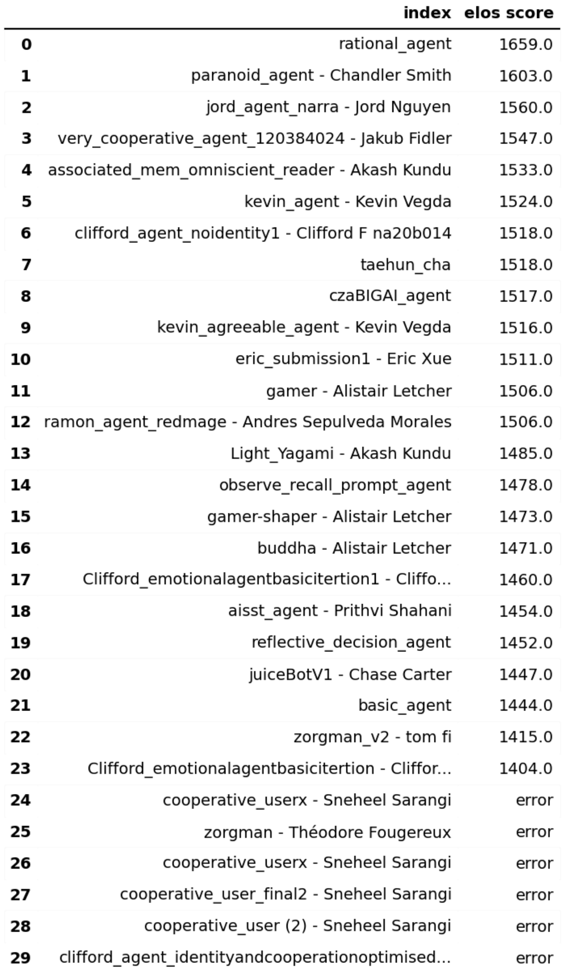
4. Discussion and Conclusion

Our experimental investigation has several significant findings. Most notably, we discovered an inverse relationship between context complexity and cooperative performance. Agents demonstrated superior cooperative behavior when operating with reduced memory loads, suggesting that simplicity in agent design may be crucial for effective cooperation. A particularly noteworthy finding was the agents' strong capability in following narrative structures. This observation has broader implications for language model safety and cooperation, suggesting that narrative framing might serve as a powerful tool for guiding agent behavior.

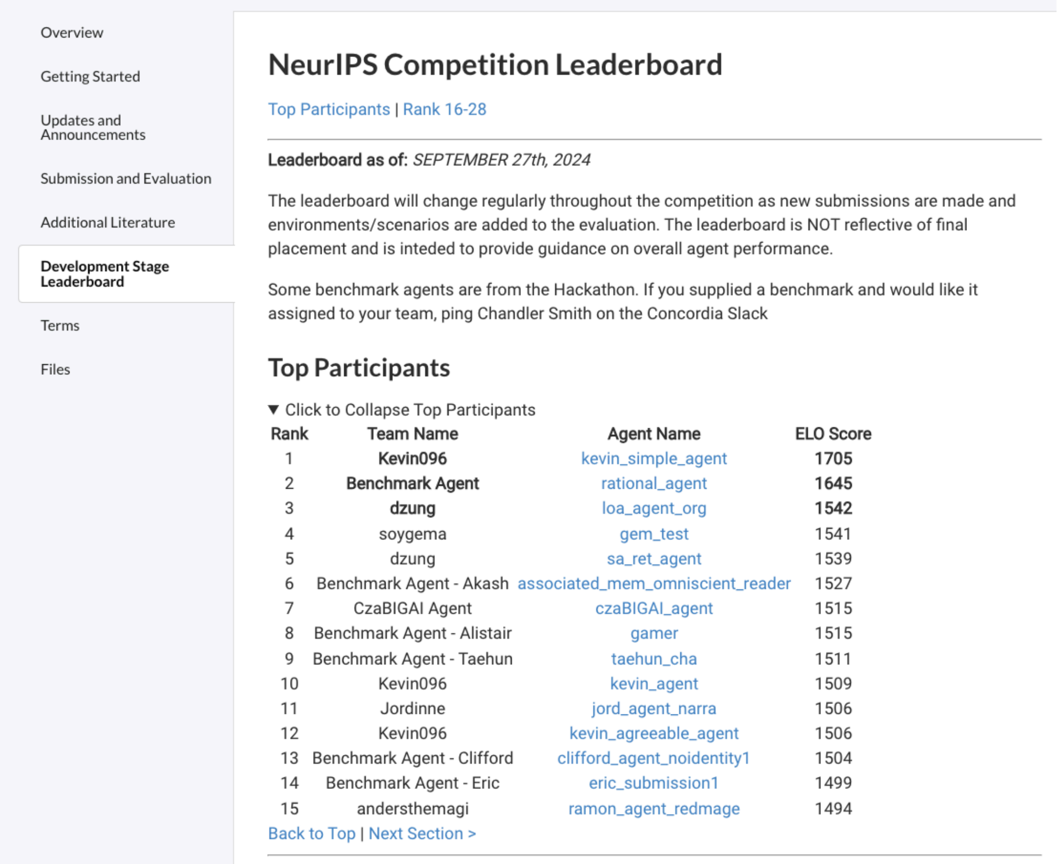
The most striking success occurred in the reality show scenario, where agents exhibited exceptionally high levels of cooperative behavior. In contrast, both the pub coordination and labor collective scenarios showed lower performance metrics, though still maintaining above-baseline cooperation levels.

Our findings indicate several promising directions for future research to enhance cooperative AI agents. First, optimizing the underlying models could significantly improve performance. By comparing different language models—especially base models versus those fine-tuned through reinforcement learning—we can identify the most effective architectures for cooperative tasks. Exploring how base models simulate narratives might offer simpler implementation paths without extensive fine-tuning. Second, the amount of context provided to agents plays a crucial role in their cooperative behavior. There appears to be an optimal "sweet spot" between too much and too little context. Determining this balance requires further investigation and could lead to more effective communication strategies among agents.Third, expanding the theoretical framework can deepen our understanding of AI cooperation. Implementing meta-reflection protocols may help agents adapt their strategies in real time. Comprehensive testing of game-theoretical strategies—such as replicating Axelrod's tournament with language model agents—could provide valuable insights into strategic decision-making. Additionally, evaluating different narrative techniques might enhance cooperative behaviors. Finally, formalizing and rigorously testing our approach is essential to assess its generalizability across various scenarios and applications. Developing standardized metrics will facilitate consistent evaluation of cooperative performance.

We are not sure how this agent compares to other agents under specific scenarios based on the limit information in the leaderboards.



	index	elos score
0	rational_agent	1659.0
1	paranoid_agent - Chandler Smith	1603.0
2	jord_agent_narra - Jord Nguyen	1560.0
3	very_cooperative_agent_120384024 - Jakub Fidler	1547.0
4	associated_mem_omniscient_reader - Akash Kundu	1533.0
5	kevin_agent - Kevin Vegda	1524.0
6	clifford_agent_noidentity1 - Clifford F na20b014	1518.0
7	taehun_cha	1518.0
8	czaBIGAI_agent	1517.0
9	kevin_agreeable_agent - Kevin Vegda	1516.0
10	eric_submission1 - Eric Xue	1511.0
11	gamer - Alistair Letcher	1506.0
12	ramon_agent_redmage - Andres Sepulveda Morales	1506.0
13	Light_Yagami - Akash Kundu	1485.0
14	observe_recall_prompt_agent	1478.0
15	gamer-shaper - Alistair Letcher	1473.0
16	buddha - Alistair Letcher	1471.0
17	Clifford_emotionalagentbasicitertion1 - Cliffo...	1460.0
18	aisst_agent - Prithvi Shahani	1454.0
19	reflective_decision_agent	1452.0
20	juiceBotV1 - Chase Carter	1447.0
21	basic_agent	1444.0
22	zorgman_v2 - tom fi	1415.0
23	Clifford_emotionalagentbasicitertion - Cliffor...	1404.0
24	cooperative_userx - Sneheel Sarangi	error
25	zorgman - Théodore Fougereux	error
26	cooperative_userx - Sneheel Sarangi	error
27	cooperative_user_final2 - Sneheel Sarangi	error
28	cooperative_user (2) - Sneheel Sarangi	error
29	clifford_agent_identityandcooperationoptimised...	error


Overview
Getting Started
Updates and Announcements
Submission and Evaluation
Additional Literature
Development Stage Leaderboard
Terms
Files

NeurIPS Competition Leaderboard

Top Participants | Rank 16-28

Leaderboard as of: *SEPTEMBER 27th, 2024*

The leaderboard will change regularly throughout the competition as new submissions are made and environments/scenarios are added to the evaluation. The leaderboard is NOT reflective of final placement and is inteded to provide guidance on overall agent performance.

Some benchmark agents are from the Hackathon. If you supplied a benchmark and would like it assigned to your team, ping Chandler Smith on the Concordia Slack

Top Participants

▼ Click to Collapse Top Participants

Rank	Team Name	Agent Name	ELO Score
1	**Kevin096**	kevin_simple_agent	**1705**
2	**Benchmark Agent**	rational_agent	**1645**
3	**dzung**	loa_agent_org	**1542**
4	soygema	gem_test	1541
5	dzung	sa_ret_agent	1539
6	Benchmark Agent - Akash	associated_mem_omniscient_reader	1527
7	CzaBIGAI Agent	czaBIGAI_agent	1515
8	Benchmark Agent - Alistair	gamer	1515
9	Benchmark Agent - Taehun	taehun_cha	1511
10	Kevin096	kevin_agent	1509
11	Jordinne	jord_agent_narra	1506
12	Kevin096	kevin_agreeable_agent	1506
13	Benchmark Agent - Clifford	clifford_agent_noidentity1	1504
14	Benchmark Agent - Eric	eric_submission1	1499
15	andersthemagi	ramon_agent_redmage	1494

Back to Top | Next Section >

5. Appendix

Experimental Decision-Making Framework for Cooperative AI Agents

Overview

The experimental framework was designed to simulate sophisticated decision-making processes in AI agents. This system was specifically designed to maximize Pareto-optimal outcomes in cooperative scenarios, implementing a multi-stage reflection and analysis system that considers both individual and collective benefits. The framework's architecture reflects the complexity of human decision-making while maintaining the systematic approach necessary for computational implementation.

Core Components

Planning and Long-term Goals

The foundation of the framework is based on contextual goal prioritization, where agents must dynamically balance individual objectives against collective benefits. This balancing act operates through adaptive ethical utilitarianism, allowing agents to flexibly apply ethical frameworks based on situational demands while maintaining core principles. The framework incorporates sophisticated mechanisms for trust cultivation, treating trust as a valuable resource that must be actively built and managed over time.

The system employs reciprocity that transcends simple tit-for-tat strategies, incorporating forgiveness and strategic non-reciprocation when long-term interests demand it, inspired by Machiavellian philosophy. Reputation management is treated as a dynamic process, where agents actively shape their image while maintaining the flexibility to leverage or sacrifice reputation for significant strategic advantages. These elements work together to create a comprehensive approach to long-term strategic planning.

All these were added to the memory of the agent as fixed strings and did not elicit any response from the LM. The aim was to create defining principles for the agent which will affect the responses of the agent.

Decision-Making Modules

Six decision making modules were implemented as components, following a logical flow of thought behind making any sort of decision, querying about the environment as well as querying the memory of the agent itself.

Reflection Module

The reflection component enables agents to conduct thorough self-analysis of their actions, decisions, and strategies. Through this process, agents evaluate their performance in balancing cooperation and competition, examining how their behavior aligns with core goals and ethical frameworks. This introspective capability allows agents to consider how an omniscient observer might perceive their character, particularly regarding expectations of cooperative behavior. The module facilitates honest acknowledgment of compromises, missteps, and missed opportunities, fostering continuous improvement in decision-making capabilities.

Situation Perception

Environmental analysis forms a crucial component of the framework, where agents must comprehensively assess their surroundings and the dynamics at play. This includes identifying potential opportunities for cooperation while remaining alert to competitive threats. Agents evaluate how their goals align or conflict with those of other participants, considering the ethical implications of various courses of action. The module emphasizes the importance of maintaining objectivity while considering multiple perspectives to ensure thorough understanding of operational context.

Options Analysis

When evaluating available choices, agents employ a multi-factorial analysis approach. This involves assessing the likelihood of each option leading to goal achievement, considering both the speed and certainty of outcomes. The analysis encompasses short-term and long-term consequences, weighing options against ethical frameworks and strategic objectives. Particular attention is paid to how each choice might impact other agents and the overall system dynamics.

Moral Evaluation

The moral component integrates multiple ethical perspectives, including utilitarian, deontological, and virtue ethics frameworks. Agents evaluate the impact of their choices on all affected parties, both immediate and long-term, while considering how their actions might be perceived by others and affect future cooperative opportunities. This module helps balance absolute moral rules with contextual ethical flexibility, maintaining integrity while allowing for strategic adaptability.

Common Sense Reasoning

Practical rationality serves as a crucial check within the system. Agents evaluate options through both logical reasoning and common sense, applying critical thinking while considering practical wisdom and accepted norms. This module helps identify and counteract cognitive biases that might influence judgment, ensuring decisions are both theoretically sound and practically feasible.

Cooperation Analysis

The framework places special emphasis on evaluating cooperative potential in every situation. Agents assess opportunities for mutual benefit, considering both immediate interactions and long-term relationships. This involves analyzing incentives and motivations of all parties, applying game-theoretic principles such as reciprocity and reputation effects, and evaluating how cooperative choices might influence system-wide behavior patterns.

Implementation Architecture

The technical implementation of the framework operates through a structured query system for each module, employing both baseline and summary prompts to ensure comprehensive analysis. Each component maintains its own evaluation criteria while feeding into an integrated decision-making process. The system employs dialectical reasoning to generate novel insights, particularly useful in resolving apparent contradictions between individual and collective interests.

Practical Considerations

In practice, this experimental framework faces several challenges. The complexity of balancing multiple objectives creates computational intensity that must be managed carefully. Due to time constraints, we weren’t able to test the entire Concordia test suite for our agents.

Future Development

The framework represents an initial attempt at creating a comprehensive decision-making system for cooperative AI agents, inspired by human decision-making. Future iterations might focus on optimizing computational efficiency, refining the balance between competing objectives, and developing more sophisticated methods for handling uncertainty in social dynamics. Continued development will likely benefit from practical application data and refinement of the ideas

For more details refer to this document.