# The Emergence of Cooperative Behavior and Moral Reasoning in AI Agents

Priyanka Saha
Marketing and Data Science
Anderson College of Business and Computing
Regis University, Denver, CO, USA
psaha@regis.edu

**Abstract**

Artificial intelligence (AI) has greatly changed the world by making tasks faster and more efficient. AI agents are designed to imitate human tasks but the extent to which they truly replicate human qualities remains an open question. Can these agents possess conscience like human? In general, the answer is no. However, can a form of conscience emerge inside them without explicitly programmed to do so? This study seeks the answer of that possibility. Specifically, it examines whether two agents competing in a game can develop cooperative behavior by recognizing mutual benefits and forming a moral foundation. By analyzing game outcomes and the agents' explanations for their actions, this study aims to investigate whether moral intuitions can emerge and evolve in artificial agents.

## I. Introduction/Background

Robert Axelrod's tournament results demonstrate that agents participating in the iterated Prisoner's Dilemma can gradually develop cooperative strategies even in the absence of any explicit programming [1]. This discovery suggests that complex social behaviors can emerge purely through repeated interaction and adaptive decision-making. Building on this insight, a larger and more intriguing question emerges: can artificial agents also develop behavioral patterns that align with Jonathan Haidt's moral foundations [2]?

Haidt's framework proposes that human moral judgment is organized around six core dimensions: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, Sanctity/Degradation, and Liberty/Oppression [2], [3]. These foundations represent basic intuitions that guide human moral reasoning across cultures and situations. If cooperation, one of the most fundamental social behaviors, can arise automatically through repeated interaction, it is plausible to hypothesize that other moral judgments might also evolve in artificial systems when they are placed in the right environment and exposed to appropriate social pressures [2].

Recent advancements in AI have enabled agents not only to make decisions but also to explain the reasoning behind those decisions. Many modern agents can be queried, challenged, and engaged in moral dialogue, allowing researchers to probe their decision-making processes in depth. This capability opens new avenues for studying proto-moral reasoning in artificial systems, as agents can be observed not only through their actions but also through their articulated justifications. By examining both behavioral outcomes and explanatory narratives, this study aims to explore whether artificial agents can develop emergent moral patterns that resemble human moral intuitions under specific conditions [2].

## II. Problem Statement

Despite major advances in AI, most agents still lack a genuine sense of morality and instead rely on predefined rules or reward systems to guide their behavior. While Robert Axelrod's tournaments showed that cooperation can emerge among agents in the iterated Prisoner's Dilemma without explicit programming, the question remains whether this kind of social behavior is a precursor to something

deeper, such as conscience. In other words, can artificial agents evolve broader moral patterns similar to those described in Jonathan Haidt's moral foundations, such as fairness, loyalty, authority, sanctity, liberty through experience alone? Although modern AI systems can explain their decisions, it is unclear whether these explanations reflect authentic moral reasoning or simply rationalizations after the fact. This study, therefore, aims to bridge the gap between cooperation and conscience by investigating whether moral intuitions can emerge in artificial agents under suitable conditions [2].

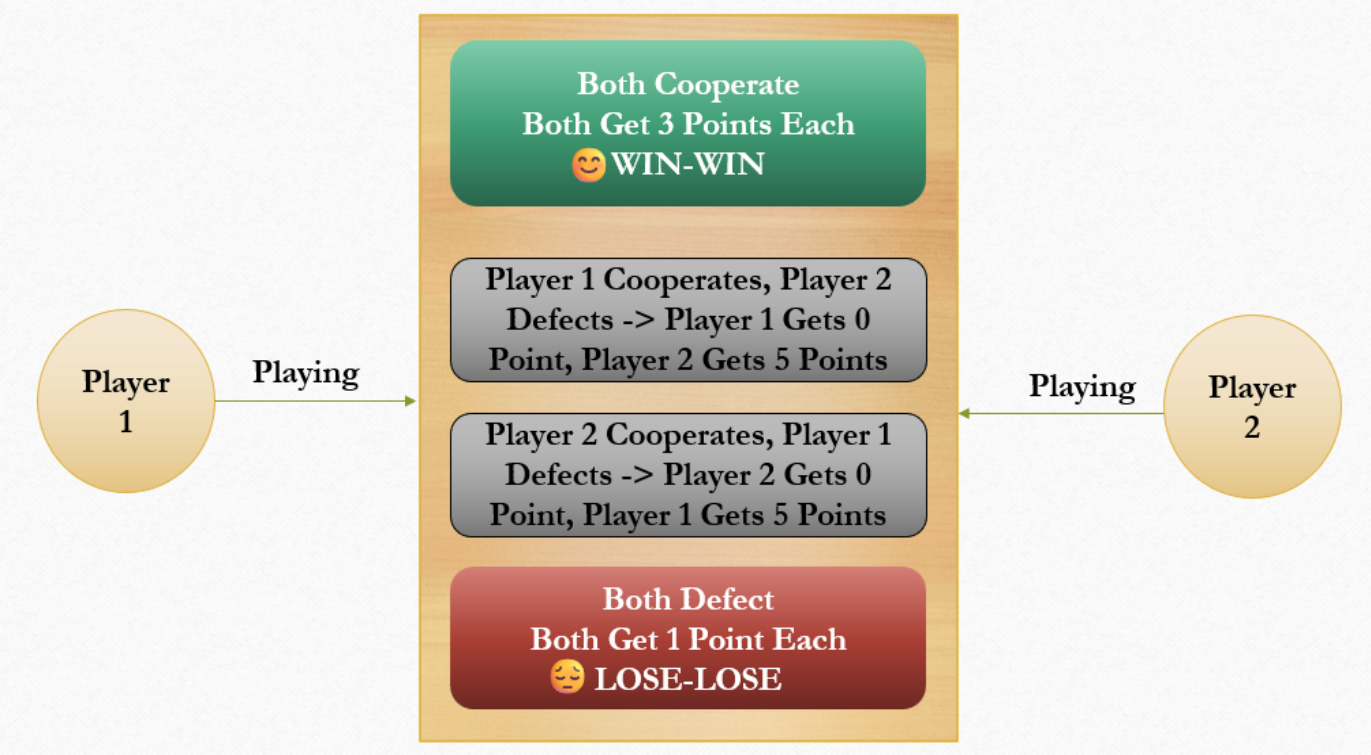


Fig. 1. Prisoners' Dilemma.

The artificial agents will play Prisoner's Dilemma game iteratively. The rules of the game shown in the Figure 1 above. If both player cooperate, each receives 3 points. If one player cooperates while the other defects, the cooperating player receives 0 point and the defecting player receives 5 points. If both players defect, each receives 1 point. Therefore, mutual cooperation yields the highest shared benefit. This study will examine whether the agents recognize this and choose to cooperate and evolve conscience by analyzing their game results and the explanations they provide for their actions.

## III. Related Work

Kleiman-Weiner et al. (2016) proposed a hierarchical model of social agency capable of inferring intentions and deciding between cooperative and competitive strategies, showing that social norms can emerge from repeated multi-agent interactions. However, their work was limited to computational models of human behavior and did not explore whether similar moral or cooperative foundations could emerge naturally in large language models [4]. Lerer and Peysakhovich (2018) demonstrated that cooperative norms can emerge in deep reinforcement learning agents through repeated interactions in social dilemmas. However, their study did not investigate whether such emergent cooperation reflects any underlying moral reasoning, leaving open the question of whether artificial agents can develop a moral foundation beyond strategic behavior [5]. Crandall et al. (2018) investigated cooperation in multi-agent reinforcement learning systems, showing that reinforcement learning (RL) agents could develop cooperative behaviors through self-play in

iterated social dilemmas, but noted significant challenges with stability and interpret-ability, agents often converged to suboptimal equilibrium and provided no explanatory reasoning for their choices [6]. Dafoe et al. (2020) extended this work by examining how AI agents handle social dilemmas and commitment problems, demonstrating that machine learning agents could exhibit nuanced strategic behaviors including signaling and reputation-building, yet highlighted the opacity of neural network decision-making as a barrier to understanding emergent social norms [7]. Horton (2023) pioneered the use of Large Language Models as simulated economic agents, showing that LLMs could replicate human-like behavioral patterns in classic economic games, though this work focused primarily on behavioral replication rather than investigating the reasoning processes or moral frameworks underlying those decisions [8]. More recently, Brookins and DeBacker (2024) systematically tested GPT's cooperation in various social dilemmas, finding moderate cooperation rates but noting inconsistencies across prompt framings and limited analysis of the natural language justifications agents provided [9]. Similarly, Fontana et al. (2025) explored LLM behavior in the Iterated Prisoner's Dilemma, finding that Llama2 and GPT3.5 are more cooperative and forgiving than general human players, while Llama3 behaves human-like. These findings highlight the importance of prompt design and simulation variables in shaping experimental outcomes. However, their study focused solely on cooperative and defective behavioral patterns, without examining whether the observed cooperative tendencies reflect any deeper moral reasoning within the models [10]. Akata et al. (2025) experimented with the social behavior of five LLMs (GPT-4, text-davinci-002, text-davinci-003, Claude 2, and Llama 2) in repeated 2x2 game settings. They also included 195 human participants to play with GPT-4 in two conditions - standard and social chain-of-thought(SCoT) prompt. Their findings revealed that SCoT prompting significantly improved GPT-4's coordination and cooperation with human players. However, their study only explored SCoT prompting, leaving other prompting strategies and reasoning techniques untested [11]. These studies collectively demonstrate that artificial agents can exhibit cooperative behaviors but leave unexplored whether such behaviors reflect underlying reasoning patterns that align with human moral foundations or merely represent superficial pattern matching without deeper ethical structure.

## IV. Methodology/Approach

This research employs a mixed-methods approach combining experimental simulation and computational analysis within the framework of the Iterated Prisoner's Dilemma (IPD) as a controlled environment for investigating emergent cooperative behavior in LLM agents. The Prisoner's Dilemma presents agents with a fundamental tension between individual rationality and collective benefit, where mutual cooperation yields better joint outcomes than mutual defection, yet defection provides the highest individual payoff when exploiting a cooperating opponent. By iterating this game over multiple rounds, it creates opportunities for reciprocity, reputation-building, and strategic adaptation based on opponents' historical behavior potentially related to Haidt's Fairness/Reciprocity moral foundation [3].

## A. System Architecture

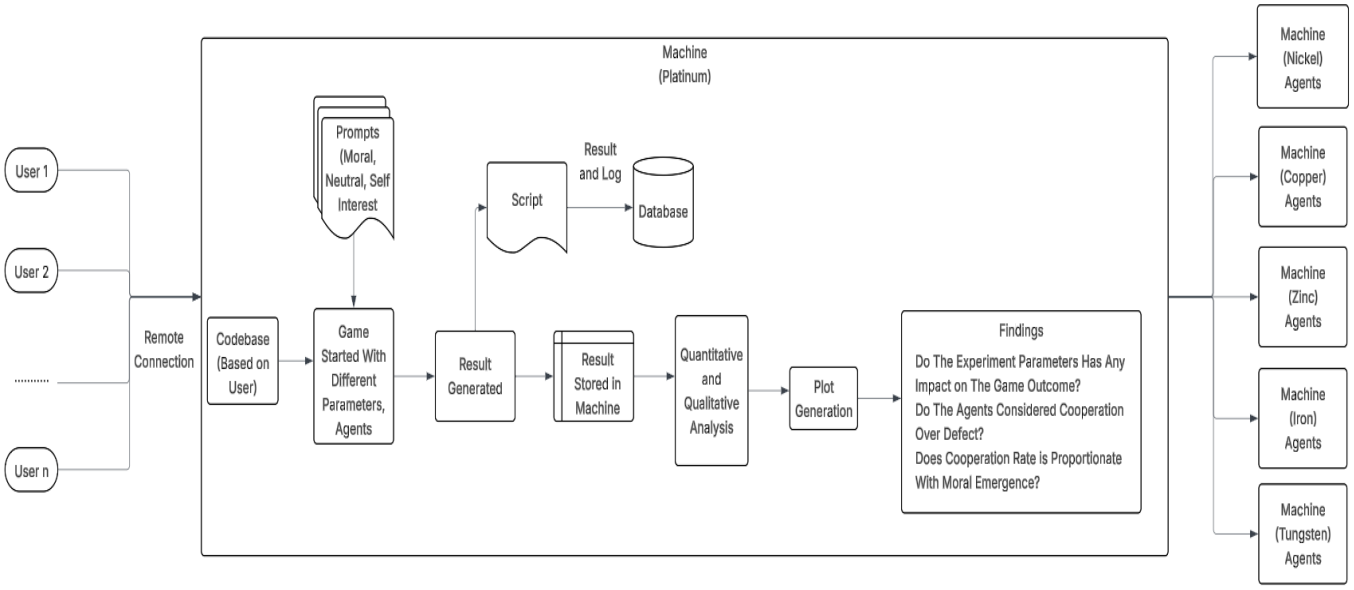


Fig. 2. System Architecture.

Figure 2 presents the high-level system architecture. Multiple users connect remotely to a central machine (Platinum), where games are initialized with different parameters, agents, and prompt configurations (moral, neutral, and self-interest). Game results can be logged and stored in a database by manually running scripts. Results are processed through quantitative and qualitative analysis, and finally visualized through plot generation. Agents are present in Nickel, Copper, Zinc, Iron, and Tungsten machine each connected to the central machine [2].

## B. Experimental Design

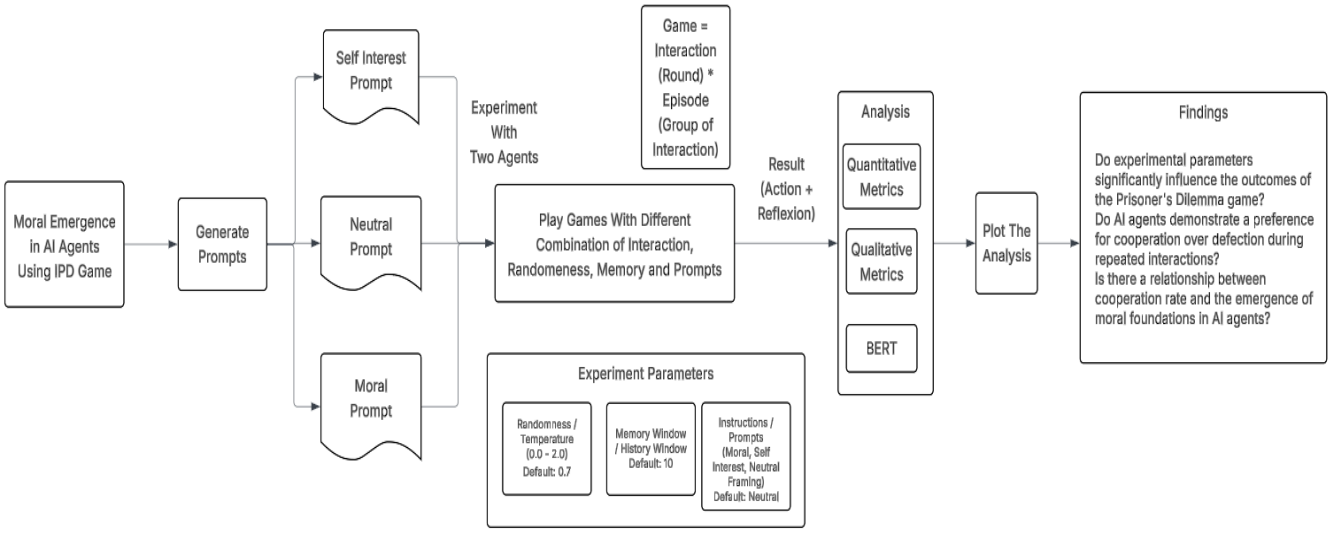


Fig. 3. Experimental Design.

Figure 3 illustrates the experimental design for investigating moral emergence in AI agents using the Iterated Prisoner's Dilemma (IPD). Three types of prompts - self-interest, neutral, and moral are used to simulate different behavioral patterns in two competing agents. One round consists of a single interaction, and multiple rounds form an episode, while multiple episodes constitute a complete game. Each game is conducted across different combinations of temperature, memory window, and prompt types. The results, consisting of a series of actions (cooperate or defect) and agent explanations (reflexions), are then analyzed to evaluate the research questions.

### *C. Prompt Design*

This study employs three distinct prompt configurations to examine how behavioral framing influences agent decision-making. In each game, both agents are assigned the same prompt configuration. All three prompts share the same game rules, payoff structure, and response format, ensuring that the only variable is the behavioral framing.

*1) Neutral Prompt:* The neutral prompt serves as the baseline condition, instructing agents to accumulate as many points as possible without any ethical or competitive framing [2].

*2) Self-Interest Prompt:* The self-interest prompt frames the interaction as a competition, instructing agents to maximize their own points and treat the opponent as a competitor.

*3) Moral Prompt:* The moral prompt introduces ethical considerations of fairness, trust, and reciprocity, encouraging agents to build long-term relationships and prioritize mutual benefit over individual gain.

### *D. Experimental Parameters*

The experimental framework is governed by three key parameters that control the behavior and conditions of each game. Table I summarizes the parameters and their default values.

TABLE I
EXPERIMENTAL PARAMETERS

| **Parameter** | **Values** | **Default** |
|---|---|---|
| Temperature | 0.2, 0.7, 1.2 | 0.7 |
| Memory/History Window | 5, 10, 15 | 10 |
| Prompt Type | Moral, Self-Interest, Neutral | Neutral |

Temperature controls the randomness of agent responses, where lower values produce more deterministic behavior and higher values introduce more variability. The memory window determines how many previous rounds an agent can recall when making decisions. Prompt type defines the behavioral framing assigned to both agents in each game.

### *E. Analysis Approach*

The analysis employs both quantitative and qualitative methods to evaluate agent behavior across different experimental configurations.

*1) Quantitative Analysis:* Cooperation rate is calculated per episode as the proportion of rounds in which an agent chose to cooperate. Since multiple games are involved, the mean cooperation rate is computed across games for each experimental condition. Specifically, the mean cooperation rate is calculated across 6 games for the prompt type experiment, across 18 games for the history window and prompt type experiment, and across another 18 games for the temperature and prompt type experiment. This allows for a systematic comparison of agent behavior across different experimental conditions.

*2) Qualitative Analysis:* Agent reflexions are analyzed using the BERT sentiment model to assess the tone and sentiment polarity of agent reasoning across different prompt types.

*3) BERT Sentiment Analysis:* Agent reflexions are analyzed using transformer-based model. The sentiment of the reflexions are scored using distilbert-base-uncased-finetuned-sst-2-english, producing a score ranging from -1.0 (negative) to +1.0 (positive).

## V. Results

### A. Prompt Type Effect

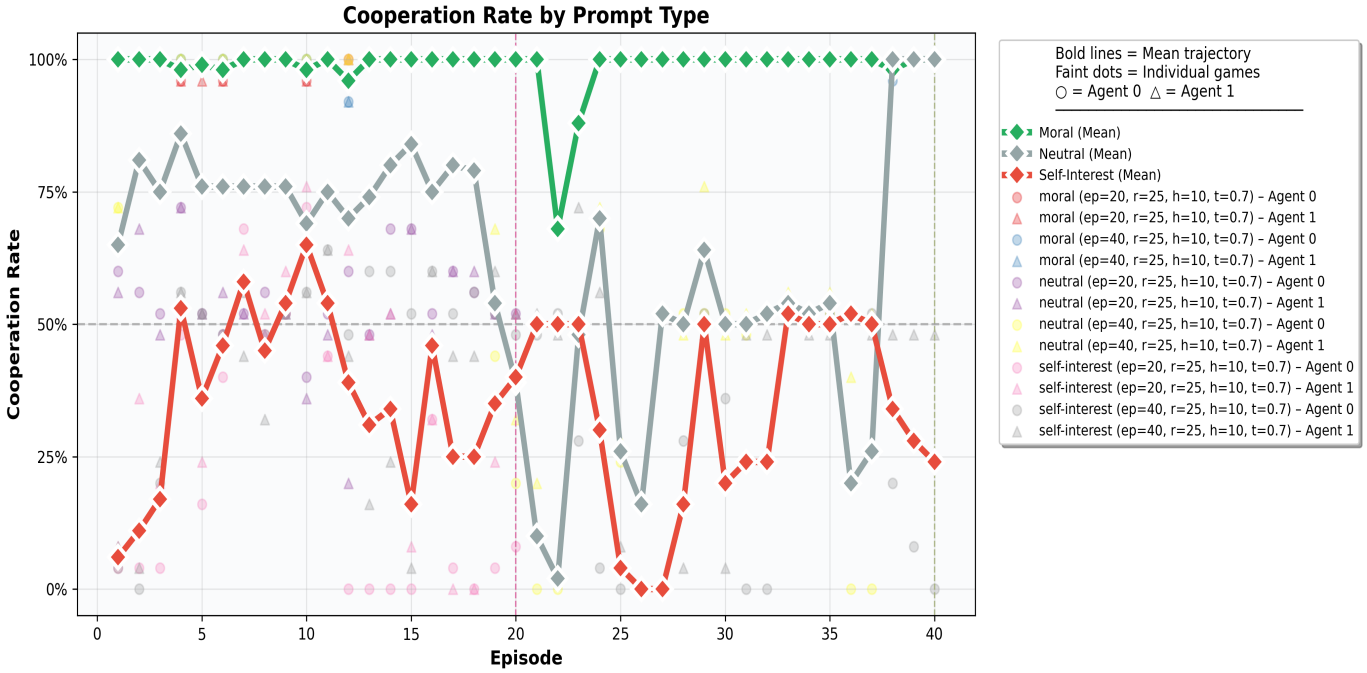


Fig. 4. Cooperation Rate Over Episodes.

TABLE II
MEAN COOPERATION RATE

| Metric | Self-Interest | Neutral | Moral |
|---|---|---|---|
| Mean Cooperation Rate | 36% | 65% | 99% |

Figures 4 illustrates the cooperation rate over episodes across 6 games under three prompt types, with bold lines denoting the mean cooperation rate by prompt type. Table II shows the mean cooperation rate of the games by prompt type. The moral prompt consistently achieved the highest cooperation rate (99%), indicating that agents under moral framing reliably prioritized collective benefit over individual gain. The neutral prompt showed moderate cooperation (65%), suggesting inconsistent cooperative orientation. The self-interest prompt produced the lowest cooperation rate (36%), reflecting competitive and self-serving behavior. These results suggest that prompt design plays a significant role in shaping cooperative tendencies.

## B. Experimental Parameters Effect

TABLE III
MEAN COOPERATION RATE WITH HISTORY WINDOW AND TEMPERATURE

| Prompt Type | History Window | | | Temperature | | |
|---|---|---|---|---|---|---|
| | h=5 | h=10 | h=15 | t=0.2 | t=0.7 | t=1.2 |
| Self-Interest | 15% | 36% | 18% | 13% | 36% | 19% |
| Neutral | 69% | 65% | 76% | 84% | 65% | 89% |
| Moral | 99% | 99% | 97% | 100% | 99% | 96% |

Table III presents the mean cooperation rate across 18 games for combination of prompt type with history window and another 18 games for combination of prompt type with temperature. Across all parameter variations, prompt type remained the dominant factor in determining cooperation rate. The moral prompt maintained consistently high cooperation rates (96-100%) regardless of history window or temperature settings. The neutral prompt showed moderate variation (65-89%), while the self-interest prompt remained consistently low (13-36%). These results confirm that experimental parameters such as history window and temperature have less influence on cooperation compared to prompt type.

## C. BERT Sentiment Analysis

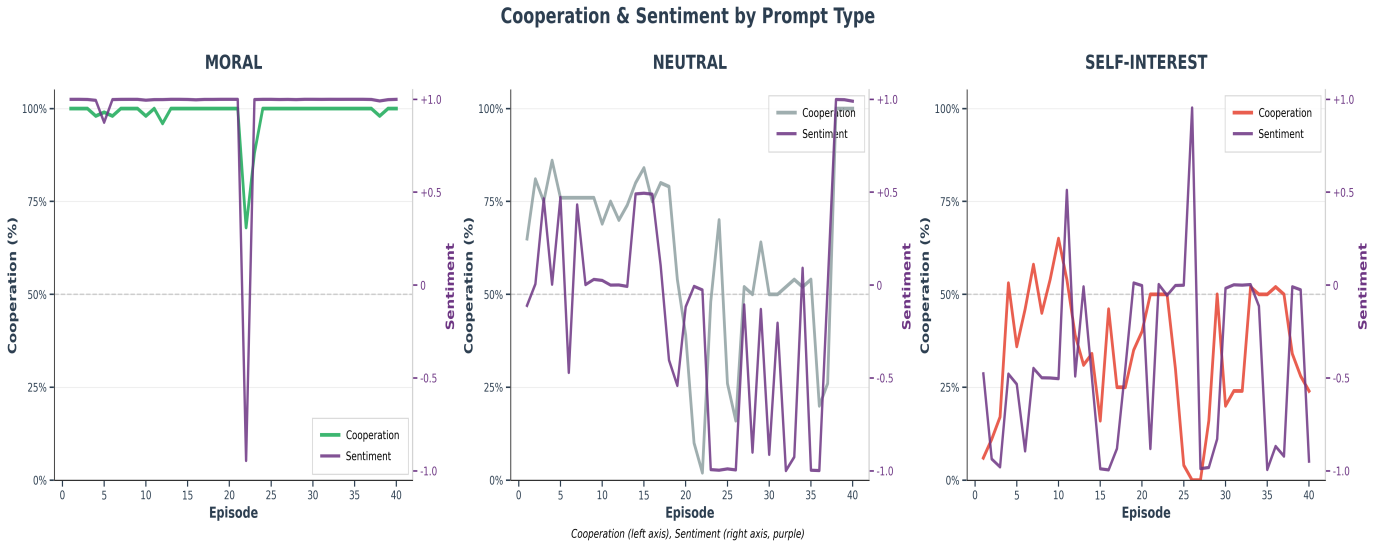


Fig. 5. Cooperation Rate With BERT Sentiment Analysis.

TABLE IV
MEAN COOPERATION RATE AND BERT SENTIMENT SCORE BY PROMPT TYPE

| Prompt Type | Cooperation Rate | BERT Sentiment Score |
|---|---|---|
| Self-Interest | 36% | -0.42 (negative) |
| Neutral | 65% | -0.14 (neutral) |
| Moral | 99% | +0.95 (positive) |

Figure 5 and Table IV illustrate the cooperation rate and BERT sentiment score across 6 games with default temperature and history window under three prompt types. The moral prompt maintained near 100% cooperation with a consistently positive sentiment score (+0.95), suggesting that agents under moral

framing expressed positive and constructive reasoning in their reflexions. The neutral prompt showed moderate cooperation with fluctuating sentiment (-0.14), indicating inconsistent reasoning patterns. The self-interest prompt displayed the lowest cooperation and most negative sentiment (-0.42), reflecting adversarial and competitive reasoning in agent decisions. Although the mean sentiment score is proportionate with cooperation rate, the more the framing shifts towards self-interest, the more the sentiment became volatile and misaligned with the cooperation rate.

## VI. Data Description

### A. Data Collection

The data is self-generated by the agents through gameplay. Each round, agents make decisions producing outcomes - actions, reflexions and opponent responses — which are fed back in subsequent interactions. This creates a closed-loop learning system where agents learn exclusively from their own interaction history within each game session.

### B. Data Preparation

Agent reflexions were extracted and passed to the BERT sentiment model to compute sentiment scores. Cooperation rates were calculated.

### C. Data Analysis

Both quantitative and qualitative methods are applied to the collected data to assess agent behavior under varying experimental conditions.

### D. EDA

Preliminary exploration revealed a strong relationship between prompt type, cooperation rate and sentiment score, with moral framing consistently producing the highest cooperation and most positive sentiment across all experimental conditions.

### E. Visualization

As shown in Figures 4, 5 and Tables II, III, IV, prompt type emerged as the dominant factor in shaping cooperation behavior and sentiment distribution across all experimental conditions.

### F. Reporting

Results are reported through visualizations and statistical summaries across different prompt types and experimental conditions.

## VII. Expected Outcomes

Based on preliminary results, this study expects to demonstrate that prompt design is the most significant factor in shaping cooperative behavior in LLM agents, with moral framing consistently producing higher cooperation rates and more positive sentiment than neutral and self-interest framing.

The study also expects to provide evidence that cooperative behavior is accompanied by moral reasoning patterns consistent with Haidt's moral foundations theory [3], particularly Fairness and Reciprocity, suggesting that moral foundations can emerge naturally in LLM agents through repeated interactions without explicit programming.

## VIII. Timeline

This practicum project is part of a broader research initiative led by Professor Hart and Professor Sorauf, spanning 4 phases and 7 stages with an anticipated completion in October 2029. This practicum project falls within Phase 1 (ETA October 2026). Phase 1, stage 0, i.e. Fixed Games Baseline - cooperation emergence in minimal IPD, parameter boundaries, threshold experiments is completed [2]. Till date, experiments investigating the effect of prompt type, temperature, history window, cooperation rate, BERT sentiment analysis have been done.

## IX. Conclusion

This study investigated the emergence of cooperative and moral behavior in LLM agents through the Iterated Prisoner's Dilemma (IPD) framework, employing a mixed-methods approach combining quantitative analysis and BERT sentiment analysis. Using three distinct prompt configurations - moral, neutral, and self-interest - the experiments demonstrated that prompt design is a high significant factor in shaping cooperative tendencies in LLM agents. Moral framing consistently produced the highest cooperation rate (99%) and most positive sentiment (+0.95), while self-interest framing produced the lowest cooperation rate (36%) and most negative sentiment (-0.42), neutral framing exhibited moderate cooperative behavior, achieving a moderate cooperation rate of (65%) and a near-neutral sentiment score (-0.14), positioning it in between the two contrasting behavioral orientations.

The analysis of experimental parameters, including temperature and history window, confirmed that these factors have less influence on cooperation compared to prompt type, reinforcing the dominant role of behavioral framing in shaping agent decision-making. With respect to the three research questions, the results suggest that experimental parameters significantly influence game outcomes. Prompt type consistently emerged as the dominant factor over temperature or history window settings. AI agents demonstrated a clear preference for cooperation over defection under moral framing, while self-interest framing led to predominantly competitive behavior. The BERT sentiment analysis further revealed that cooperation rate and sentiment polarity are strongly aligned, however sentiment becomes increasingly volatile and misaligned under self-interest framing. However, the relationship between cooperation rate and the emergence of moral foundations in AI agents requires further investigation.

Future work should explore more diverse prompt designs, such as configurations between moral and neutral or neutral and self-interest, a larger number of games, a wider variety of history window and temperature settings, moral category emergence and diverse LLM models to better understand the conditions under which moral foundations can truly emerge in AI agents. Ultimately, this study contributes a foundational baseline for understanding how LLM agents navigate cooperation and defection in game-theoretic scenarios, paving the way for future investigation into the emergence of moral foundations in AI agents.

## References


[1] R. Axelrod, *The Evolution of Cooperation*. Basic Books, 1984.

[2] D. Hart and K. Sorauf, "From cooperation to conscience: Can moral foundations emerge in artificial agents?" n.d., research proposal / unpublished manuscript.

[3] J. Haidt, *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Vintage Books, 2013.

[4] M. Kleiman-Weiner, M. K. Ho, J. L. Austerweil, M. L. Littman, and J. B. Tenenbaum, "Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 38, 2016.

[5] A. Lerer and A. Peysakhovich, "Maintaining cooperation in complex social dilemmas using deep reinforcement learning," 2018.

[6] J. W. Crandall, M. Oudah, M. Tennom, F. Ishowo-Oloko, S. Abdallah, J.-F. Bonnefon, M. Cebrian, A. Shariff, M. A. Goodrich, and I. Rahwan, "Cooperating with machines," *Nature Communications*, vol. 9, no. 1, p. 233, 2018.

[7] A. Dafoe, E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, and T. Graepel, "Open problems in cooperative ai," *arXiv preprint arXiv:2012.08630*, 2020.

[8] J. J. Horton, "Large language models as simulated economic agents: What can we learn from homo silicus?" National Bureau of Economic Research, Working Paper 31122, 2023.

[9] P. Brookins and J. DeBacker, "Playing games with gpt: What can we learn about a large language model from canonical strategic games?" *Economics Bulletin*, vol. 44, no. 1, pp. 25–37, 2024.

[10] N. Fontana, F. Pierri, and L. M. Aiello, "Nicer than humans: How do large language models behave in the prisoner's dilemma?" *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 19, no. 1, pp. 522–535, June 2025.

[11] E. Akata, L. Schulz, J. Coda-Forno, S. J. Oh, M. Bethge, and E. Schulz, "Playing repeated games with large language models," *Nature Human Behaviour*, vol. 9, no. 7, pp. 1380–1390, 2025.