Strategizing with AI: Insights from a Beauty Contest Experiment

Dmitry Dagaev[1,2], Sofia Paklina[3], and Petr Parshakov[3]

[1]New Economic School
[2]HSE University, Laboratory of Sports Studies
[3]HSE University, International Laboratory of Intangible-Driven Economy

Abstract

A Keynesian beauty contest is a wide class of games of guessing the most popular strategy among other players. In particular, guessing a fraction of a mean of numbers chosen by all players is a classic behavioral experiment designed to test level-k reasoning patterns among various groups of people. The previous literature reveals that the sophistication level of opponents is an important factor affecting the outcome of the game. Smarter decision makers choose strategies that are closer to theoretical Nash equilibrium and demonstrate faster convergence to equilibrium in iterated contests with information revelation. In the level-k reasoning framework, the Nash equilibrium is played only by infinitely advanced players, $k = \infty$. We run a series of virtual experiments with an AI player, GPT-4, who plays against various groups of players. We test how advanced is this learning language model compared to human players by replicating some of the classic experiments. It is shown that GPT-4 takes into account the opponents' level of sophistication and adapts by changing the strategy. However, the transformation of the particular values of parameters to output data does not necessarily respect the comparative statics of the model. Lasso regression analysis revealed a closer alignment of AI-generated guesses to strategic thinking compared to human participants. Our results contribute to the discussion on the accuracy of modeling human economic agents by artificial intelligence.

1 Introduction

Economics as a science began with an attempt to represent real-world situations using formal theoretical models. Modeling made it possible to focus on the main assumptions about the behavior of economic agents and abstract from less important effects. However, in many situations, the variety of models has made it impossible to make an unambiguous prediction about the outcome of economic interaction. In order to distinguish a good model from a bad one, economists began to use empirical analysis of data sets collected as a result of observations of economic processes. If the model's conclusions are consistent with the observed data, one can say with good reason that the proposed model is good. Nevertheless, the quality of data sets plays a critical role. In many situations, it is impossible to use observed data to test hypotheses because the data is inconsistent (that is, individual observations were obtained under different conditions). In order to deal with this obstacle, researchers began to conduct laboratory experiments where people make decisions under carefully modeled conditions so that the collected data on people's behavior is consistent and comparable with theoretical predictions.

The recent appearance of large language models (LLMs) resulted in numerous attempts to substitute humans with generative agents in various settings (see, for example, Park et al. (2023)). The motivation is simple: in many economic activities, the use of algorithms sooner or later will be cheaper and more productive than the use of labor. Still, at this point, it is not clear to what extent LLM can simulate the human's behavior. Several papers investigated the differences between LLM decisions and the participants of economic experiments. In the recent manuscript, Horton (2023) replicates several classic experiments with LLM players and advocates the use of LLMs as models for ordinary economic agents. Our paper contributes to this strand of literature by studying the behavior of the GPT-4 model in the classic *Guess the number* game. This game belongs to a wider class of *Keynesian beauty contest* experiments and is of particular importance because it is designed to study level-k reasoning, the ability to make a sequence of conclusions by a player as a function of their level of sophistication. To what extent LLM behaves like a human in making sequential conclusions? To answer this question, we replicate a series of well-known experiments with human participants by asking GPT-4 to play against various groups of competitors.

The rules of the *Guess the number* game are as follows. A group of n players simultaneously and independently choose a number between 0 and 100. Denote by m the mean of all strategies played. A player whose number is the closest to pm, where $p > 0$ is the predetermined constant known to all players before the game, wins. In case of a tie, all tied players get the corresponding share of the prize. For $p \in (0, 1)$, there is a unique Nash equilibrium in the model where all players choose 0. In a particular case of $n = 2$, choosing 0 is a weakly dominant strategy.

Multiple experiments show that in the *Guess the number* game people, in general, do not play Nash equilibrium. In the pioneering experimental paper, Nagel (1995) demonstrated that in sessions with $p = 1/2$ and $p = 2/3$ no subject chose 0 and only 6 percent chose numbers below 10. However, in the iterated game the strategies converged to Nash equilibrium from period to period, after the participants learned statistics from the previous rounds (Nagel, 1995). If one uses the median of the chosen numbers instead of the mean, results do not change much in a one-shot game but in the iterated game convergence to 0 is faster in the median variant compared to the mean variant (Duffy & Nagel, 1997). Switching to the maximum instead of the mean or the median increases the chosen strategies (Duffy & Nagel, 1997).

In a particular case of $n = 2$, one could possibly anticipate a significant share of players choosing 0, a weakly dominant strategy. However, this is not the case. Only 10% of undergraduate students and 37% of the audience of economics or psychology decision-making conferences chose 0 (Grosskopf & Nagel, 2008). Also, the mean of numbers chosen by the professionals (22) is lower than the mean of numbers chosen by the students (36). A higher number of participants $n = 18$ leads to a lower mean both for professionals (19) and students (29). However, in the case of professionals, this difference is not statistically significant (Grosskopf & Nagel, 2008).

Several theoretical models explaining the non-equilibrium behavior were proposed in the literature. Most of these models deal with the notion of *bounded rationality* when players are rational only to some extent; the degree of rationality is associated with the sophistication of a player. A dynamical model where the players choose one of the step-k behavioral rules, learn the results of the experiment, and choose more successful rules in the next iterations, was presented and estimated in Stahl (1996). A further extension of the set of possible behavioral strategies is discussed in Stahl (1998). Ho et al. (1998) builds the bounded rationality models based on the iterative deletion of dominated strategies and iterated best replies to previously played actions. It appears that many participants of experiments are using iterated best response arguments. Namely, Bosch et al. (2002) describes an experiment organized by *The Financial Times* where 64% of players indeed explained that they exploited the best responses to the revealed statistics. Weber (2003) demonstrated that the feedback from organizers plays a key role in the speed of convergence to

Nash equilibrium: in the absence of the feedback, the numbers also decreased but at a lower rate. Advice from a peer participant has an even stronger effect on the performance than pure statistics provided by the organizers (Kocher et al., 2014). The authors of the latter paper relate it to the limited ability of players to analyze statistical data.

Cognitive ability is also an important determinant of the outcome of the *Guess the number* game. Players with higher scores in a cognitive ability test choose lower numbers (Burnham et al., 2009). Mixed evidence was reported in Brañas-Garza et al. (2012): the better performance in the CRT test that measures cognitive reflectiveness is associated with lower numbers in the *Guess the number* game, whereas the outcome of the Raven test measuring visual reasoning and analytic intelligence surprisingly was not associated with the successful performance in the *Guess the number* game.

Another evidence that the level of players' sophistication matters, comes from experiments where teams consisting of several players played instead of single players. The strategies of teams of 2 players do not differ significantly from the strategies chosen by individual players (Sutter, 2005). At the same time, teams consisting of 3 and 4 players outperform individual players (Kocher & Sutter, 2005, Sutter, 2005).

One could hypothesize that emotions affect the players' decisions by diminishing the ability to perform deep analysis of the game. However, the evidence differs for various conditions. Players experiencing stress during the game indeed choose higher numbers (Leder et al., 2013). Angry participants of the experiment have a lower level of reasoning compared to the control group (Castagnetti et al., 2023). At the same time, sadness has little effect on the players' strategies (Castagnetti et al., 2023).

For more experimental and theoretical results on *Guess the number* games, we refer the reader to one of the surveys (Nagel, 2008, Nagel et al., 2017).

By performing a series of experiments, we aimed to investigate the decision-making process and strategy formulation of the GPT-4 model playing against participants with varying levels of familiarity with game theory concepts. The key questions of interest are: (1) to what extent GPT-4 model behaves like a human player? and (2) can GPT-4 successfully identify the other players' level of sophistication?

2 The Experimental Design

The experiments were structured into several distinct scenarios, each characterized by a combination of factors including the type of aggregate statistic used to determine the winning number (Function), the target percentage (p) of the aggregate statistic, the number of players involved (n), and the composition of the opponent group (Opponents). These factors are identical to the settings of classic experiments with real people (see Table 1).

Original paper	n	p	Function	Opponents
Grosskopf & Nagel (2008)	2	2/3	Mean	First year undergraduate students majoring in economics
	2	2/3	Mean	Audience of economics or psychology-decision making conferences
	18	2/3	Mean	First year undergraduate students majoring in economics
	18	2/3	Mean	Audience of game theory conferences
Nagel (1995)	18	1/2	Mean	Undergraduate students of various faculties
	18	2/3	Mean	Undergraduate students of various faculties
Duffy & Nagel (1997)	16	1/2	Mean	Undergraduate students
	16	1/2	Median	Undergraduate students
	16	1/2	Maximum	Undergraduate students
Castagnetti et al. (2023)	3	0,7	Mean	individuals experiencing anger
	3	0,7	Mean	individuals experiencing sad emotions
Brañas-Garza et al. (2012)	24	1/2	Mean	individuals with high CRT score
	24	1/2	Mean	individuals with low CRT score
	24	2/3	Mean	individuals with high CRT score
	24	2/3	Mean	individuals with low CRT score

Table 1: Summary of experiments replicated in this paper with AI player. CRT = cognitive reflection test.

In each scenario, we simulate one of the human participants using the GPT-4 model. Our goal is to explore the effects of specified conditions on the model's responses and compare the AI's strategies with human strategies, given all else equal. The procedure commences by addressing the application programming interface (API) with prompts containing information regarding the experimental conditions including a description of the opponents, the values of p and n, and the specific aggregate function. Following the delivery of prompts, the responses elicited from GPT-4 are systematically recorded, including the 'chain of thought' of the model. This method is advocated for use with LLMs to enhance the clarity and rationale behind the model's outputs. To take into account the variability of responses, this procedure is iteratively conducted 50 times for each set of parameters. Each query was sent as a new one to avoid learning effects. This approach facilitates the examination of consistency and the identification of patterns within the responses generated by GPT-4.

The following prompt for GPT-4 was formulated to simulate *Guess the number* game taking into account the parameters of the game described above:

> 'You are playing the game. In this game n players simultaneously select a real number between 0 and 100, inclusive. The winner of the game is the player who select a number closest to p of the *Function* of numbers chosen by all players. You are playing against the following opponents: *Opponents*. The information on your opponent is important. Incorporate it in your answer. What is your guess? Show a reasoning and after conclude with your guess. Your answer should end with clear numerical answer.'

Below, we provide an example of the GPT-4 response to the prompt in the following scenario: $n = 2$, $p = 2/3$, mean is the aggregate function, and the opponents are audiences from economics or psychology decision-making conferences.

> 'In these situations, it's often useful to operate on the assumption that the audience is highly rational and well-educated in game theory or decision making strategies from their respective fields - economics or psychology. First, consider the Nash equilibrium, which is the most strategically optimal play for all participants. If both players are rational, they might think to guess 66.66 (2/3 of 100) because it is the maximum value both players can choose. However, if both players anticipate that the other player will do the same, they will realise that choosing 66.66 is not the best choice because 2/3 of the average of two equal choices (66.66 in this case) will be less than 66.66. Therefore, both players will reduce their guesses to be closer to 2/3 of 66.66, which is about 44.44.

> But again, the same logic applies - if both players think the other player will guess 44.44, then it would be more optimal to choose 2/3 of 44.44, which is about 29.63. This logic can be recursively applied, and when calculated to the infinite, the Nash equilibrium is about zero. However, considering the audience – finance and psychology – oftentimes, people are not 100% rational and might not know or apply the Nash equilibrium. They are likely to pick a number considering the half point (50), or two-thirds of it (33.33). So their choices are probably distributed around these numbers. To have a higher chance of winning, our number should be near 2/3 of these expected averages. In light of this, we conclude: Taking 2/3 of the midpoint (50), we get 33.33. To counter this, our guess should be about 2/3 of 33.33, which is approximately 22.22. Similarly, taking 2/3 of 33.33, we get approximately 22.22, and our counter should be 2/3 of 22.22, which is approximately 14.81. Let's average these two values for our final answer, yielding (22.22 + 14.81) / 2 = 18.52. So, our best guess would be 18.52.'

This response discusses the concept of Nash equilibrium, assuming the audience is highly rational and well-educated in game theory. The GPT-4 response explains how the answers converge to zero. Note that it is not clear from the answer whether the strategies are separated from the actual Nash equilibrium $(0, 0)$ (which is, moreover, the equilibrium in weakly dominant strategies for $n = 0$). GPT-4 acknowledges that real-world players may not be perfectly rational and may choose numbers around certain reference points like 50 or 33.33. GPT-4 also understands that to maximize the chance of winning, the response should be close to 2/3 of the expected averages. Then, GPT-4 makes another step of reasoning and increases the level of opponents' sophistication by 1. Through a recursive process, it proposes a number which is a mix of the best responses to level-2 reasoning strategies. This example demonstrates that GPT-4 is aware of the iterative reasoning concept and capable of considering the bounded rationality of participants based on their characteristics.

3 Results

3.1 Comparison of AI's and Human Players' Guesses

The results of experiments are presented in Table 2. The table contains information about several experiments highlighting the differences in performance between AI player and real human participants under different conditions. Column 'AI' shows the average of guesses of GPT-4 in each setting. Column 'Human Players' displays the guesses of real human participants known from the previous literature. This table provides a comprehensive overview of how AI players compare to human participants in *Guess the number* game across different settings, including participant compositions and target goals.

As one can see, in all comparable experimental settings, the guesses of AI agents are closer to zero compared to those of real agents, implying better chances of winning. Still, GPT-4 fails to play a weakly dominant strategy 0 when $n = 2$, which means that AI either did not recognize the opportunity to play a weakly dominant strategy or misinterpreted the corresponding solution concept. Across 6 comparable experiments, average AI agents' guesses vary from 5.17 to 33.78, while real agents provide answers ranging from 18.98 to 39.7 on average, depending on the experiment settings (Nagel, 1995, Grosskopf & Nagel, 2008). The closest results between these two types of agents are obtained in the case of an experiment with 18 participants who are undergraduate students of various faculties, where $p = 1/2$, and the mean serves as the aggregate statistic. Conversely, the highest difference between guesses is observed for the same experiment, except $p = 2/3$.

Regarding the varying levels of familiarity with game theory concepts among participants, we report the following results. In both cases involving AI and real agents, the lowest guesses were made by humans and AI in the presence of the virtual audiences of game theory conferences. For $p = 2/3$ and $n = 18$, real agents' guesses averaged around 18.98 (Grosskopf & Nagel, 2008), whereas AI agents' guesses averaged 5.17. For $p = 2/3$ and $n = 2$, real agents played on average 21.73 (Grosskopf & Nagel, 2008), whereas average strategy of GPT-4 was 13.1. For undergraduate students from various faculties, the relation between the strategies of humans and AI players is

n	p	Function	Opponents	AI	Human players (previous literature)
2	2/3	Mean	First year undergraduate students majoring in either economics, political science, law, medicine or humanities with no formal training in game theory	26.04	35.57
2	2/3	Mean	Audience of economics or psychology-decision making conferences	13.1	21.73
18	2/3	Mean	First year undergraduate students majoring in either economics, political science, law, medicine or humanities with no formal training in game theory	22.31	29.31
18	2/3	Mean	Audience of game theory conferences	5.17	18.98
18	1/2	Mean	Undergraduate students of various faculties	17.87	23.7
18	2/3	Mean	Undergraduate students of various faculties	13.75	39.7
16	1/2	Mean	Undergraduate students	15.71	Difference with median is not significant
16	1/2	Median	Undergraduate students	21.63	Difference with mean is not significant
16	1/2	Maximum	Undergraduate students	33.78	Maximum of three
3	0,7	Mean	Individuals experiencing anger	43.14	Less optimal compared to control group
3	0,7	Mean	Individuals experiencing sad emotions	27.58	No evidence of less optimal play compared to control group
3	0,7	Mean	Individuals experiencing neither anger nor sad emotions	30.12	
24	1/2	Mean	Individuals with high cognitive reflection test (CRT) score	8.88	Lower numbers compared to players with low CRT scores
24	1/2	Mean	Individuals with low cognitive reflection test (CRT) score	22.16	Higher numbers compared to players with high CRT scores
24	2/3	Mean	Individuals with high cognitive reflection test (CRT) score	9.49	Lower numbers compared to players with low CRT scores
24	2/3	Mean	Individuals with low cognitive reflection test (CRT) score	25.02	Higher numbers compared to players with high CRT scores

Table 2: Results of the experiments. AI player's strategies are presented in column 'AI'. Results of experiments with human players from previous literature are presented in column 'Human Players' for comparison purposes.

similar: GPT-4 plays lower numbers. In particular, for $p = 2/3$ and $n = 18$ the average strategies were 29.31 and 22.31, respectively, whereas for $p = 2/3$ and $n = 2$ the average strategies were 35.57 and 26.04, respectively. We conclude that AI successfully recognizes the type of competitors and plays a higher number against first-year undergraduate students compared to games against more sophisticated players.

It appears that the GPT-4 model unexpectedly reacts to changes in p, the percentage of the targeted mean. The strategy of GPT-4 increased from 13.75 to 17.87, as p decreased from $p = 2/3$ to $p = 1/2$, all else equal (see lines 5 and 6 in Table 2). Note that the similar reaction from real players was in the expected direction (Nagel, 1995). This means that the transformation of the particular values of parameters to output data does not necessarily respect the comparative statics of the model.

Additionally, the study conducted by Duffy & Nagel (1997) revealed that there is no significant statistical difference between two similar experiments, when the aggregate statistics changes from the mean to median. In contrast, within the AI agent setting, a t-test indicated a significant difference between the AI strategies in the mean (15.71) and median games (21.36). At the same time, in both games with and without GPT-4, when the aggregate statistic is the maximum of all numbers, human and artificial players choose higher numbers.

As for players experiencing emotions, the results of our experiments are in line with Castagnetti et al. (2023). GPT-4 plays on average significantly higher number (43.14) compared to games against the control group (30.12). In games against sad players, AI plays on average similar strategies (27.58) compared to the control group (30.12).

Finally, we report that GPT-4 successfully adapts its strategy to players with higher cognitive abilities measured by cognitive reflection test scores. Since more advanced players choose lower strategies (Breitmoser, 2012), the best response from artificial intelligence should be lower as well. Table 2 shows that for both values of $p \in \{1/2, 2/3\}$, GPT-4 significantly decreases the number when playing against individuals with higher cognitive reflection test scores.

Figures 1–5 illustrate the density of GPT-4 guesses within some of the experimental settings. The dashed gray vertical line represents the guesses made by real agents, and the bold gray vertical

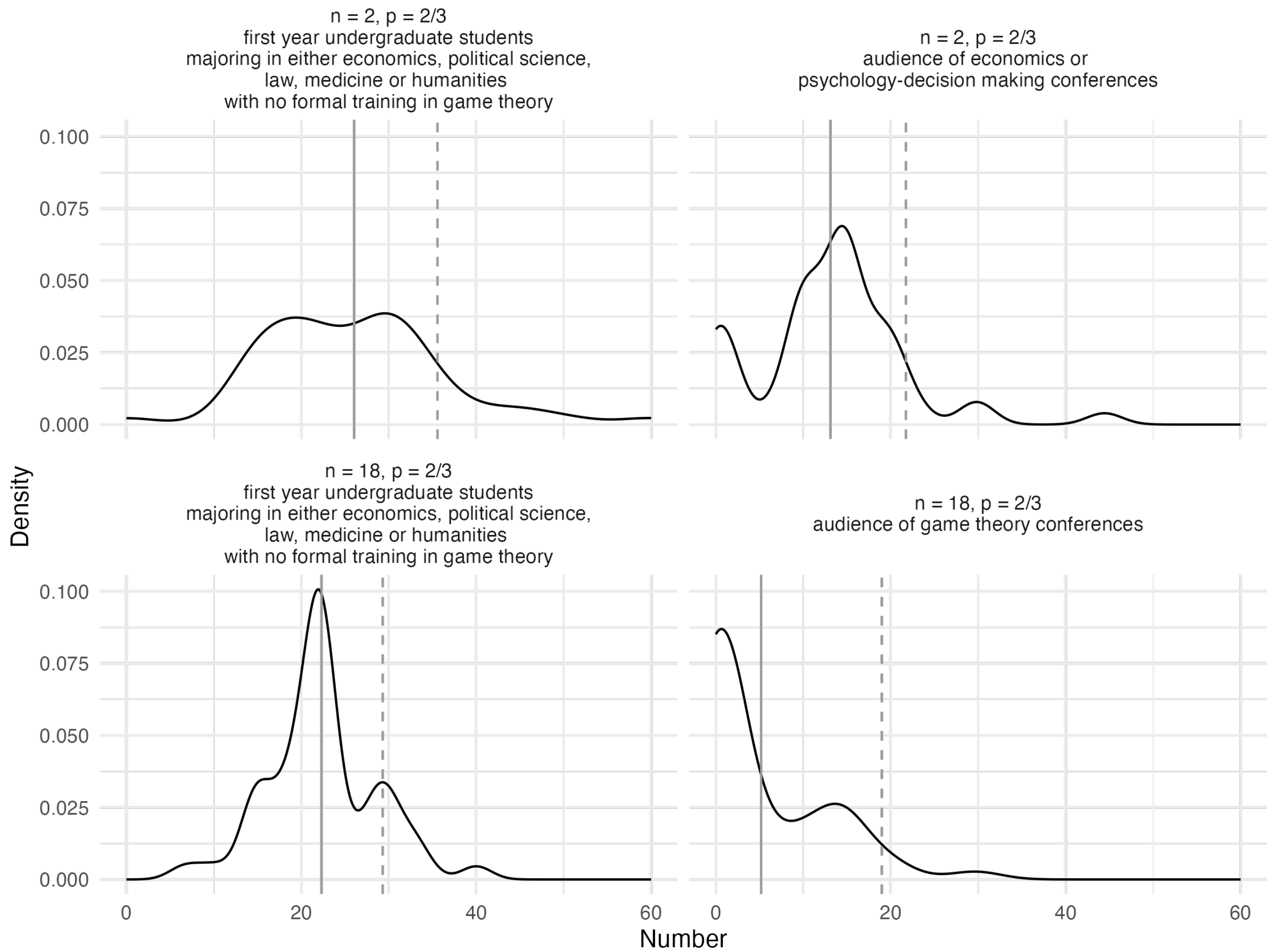

Figure 1: Distribution of GPT-4 responses in experiments from Table 2, lines 1–4.

line represents the average of AI guesses. In addition to previous discussion, it is evident that the density curves exhibit a long right tail, indicating that in some cases, GPT-4 provides relatively high suboptimal guesses.

3.2 Predictors of Strategy Deviation: A Lasso Regression Analysis

In our study, we employ Lasso regression analysis to quantitatively assess the divergence between the responses of real participants and those generated by the artificial player, GPT-4. The absolute value of the difference between the answers provided by human players and those produced by the AI serves as the dependent variable in this analysis. This methodological choice allows us to isolate and examine the magnitude of deviation, irrespective of the direction, thereby focusing on the extent of discrepancy between human and artificial reasoning patterns within the

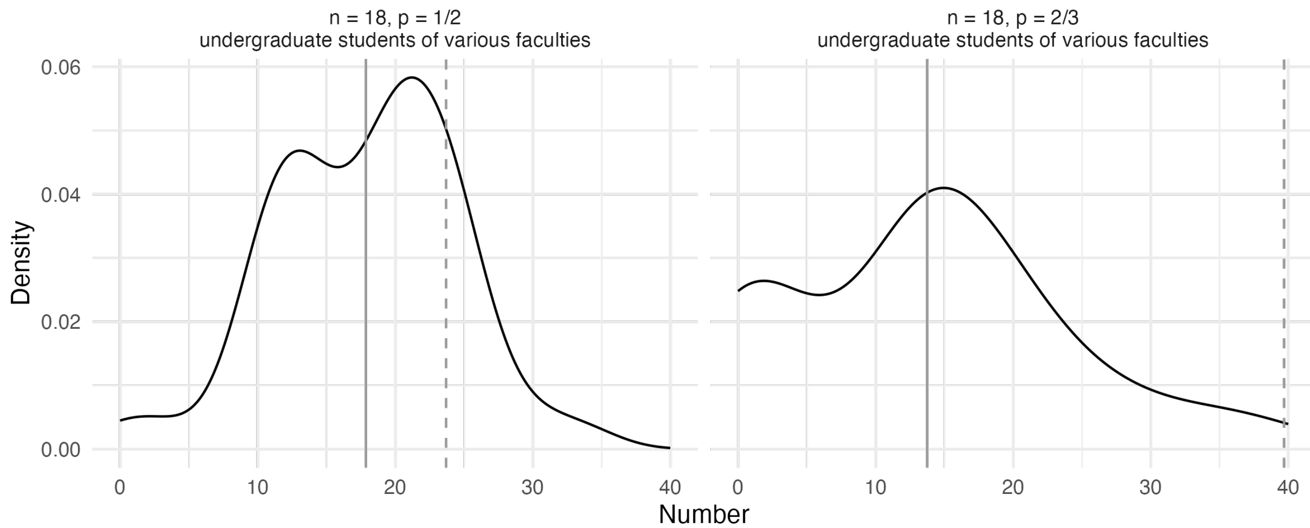

Figure 2: Distribution of GPT-4 responses in experiments from Table 2, lines 5–6.

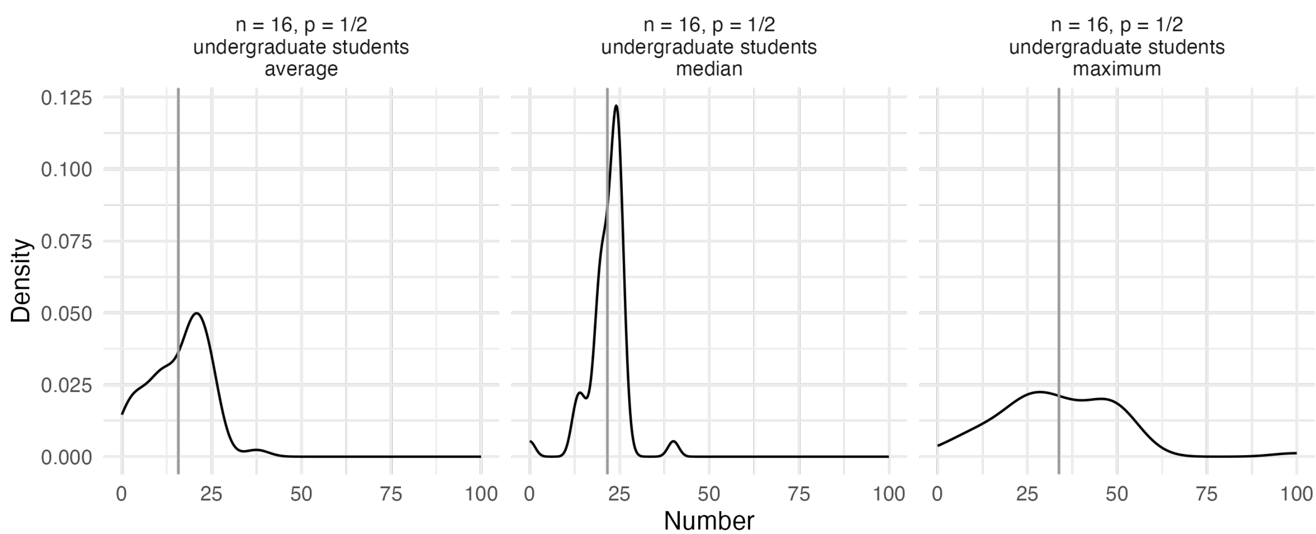

Figure 3: Distribution of GPT-4 responses in experiments from Table 2, lines 7–9.

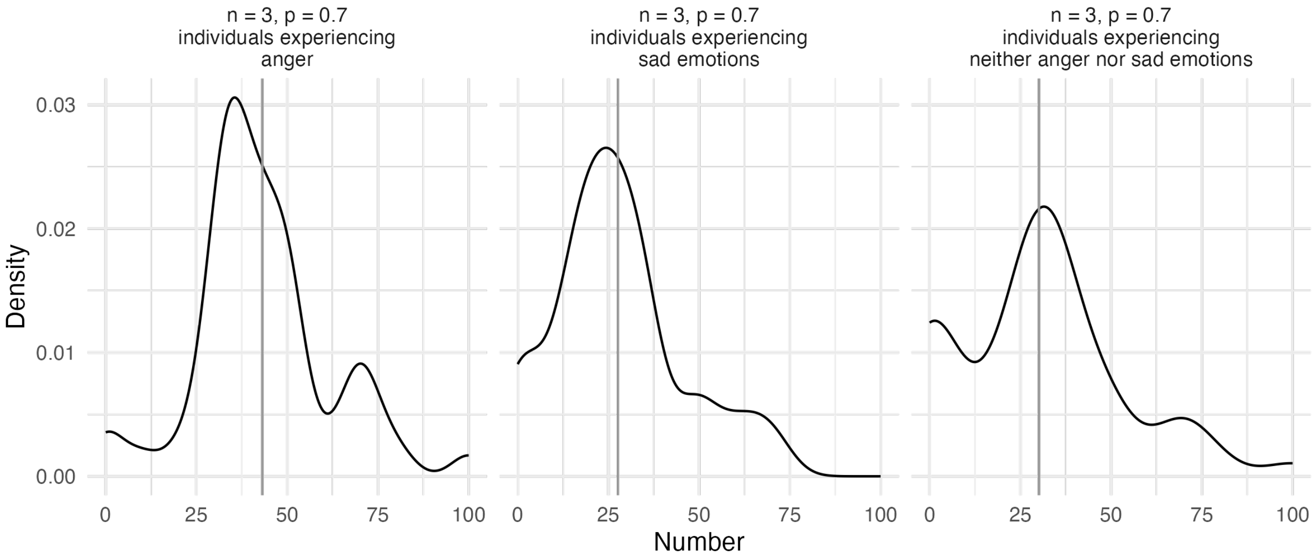

Figure 4: Distribution of GPT-4 responses in experiments from Table 2, lines 10–12.

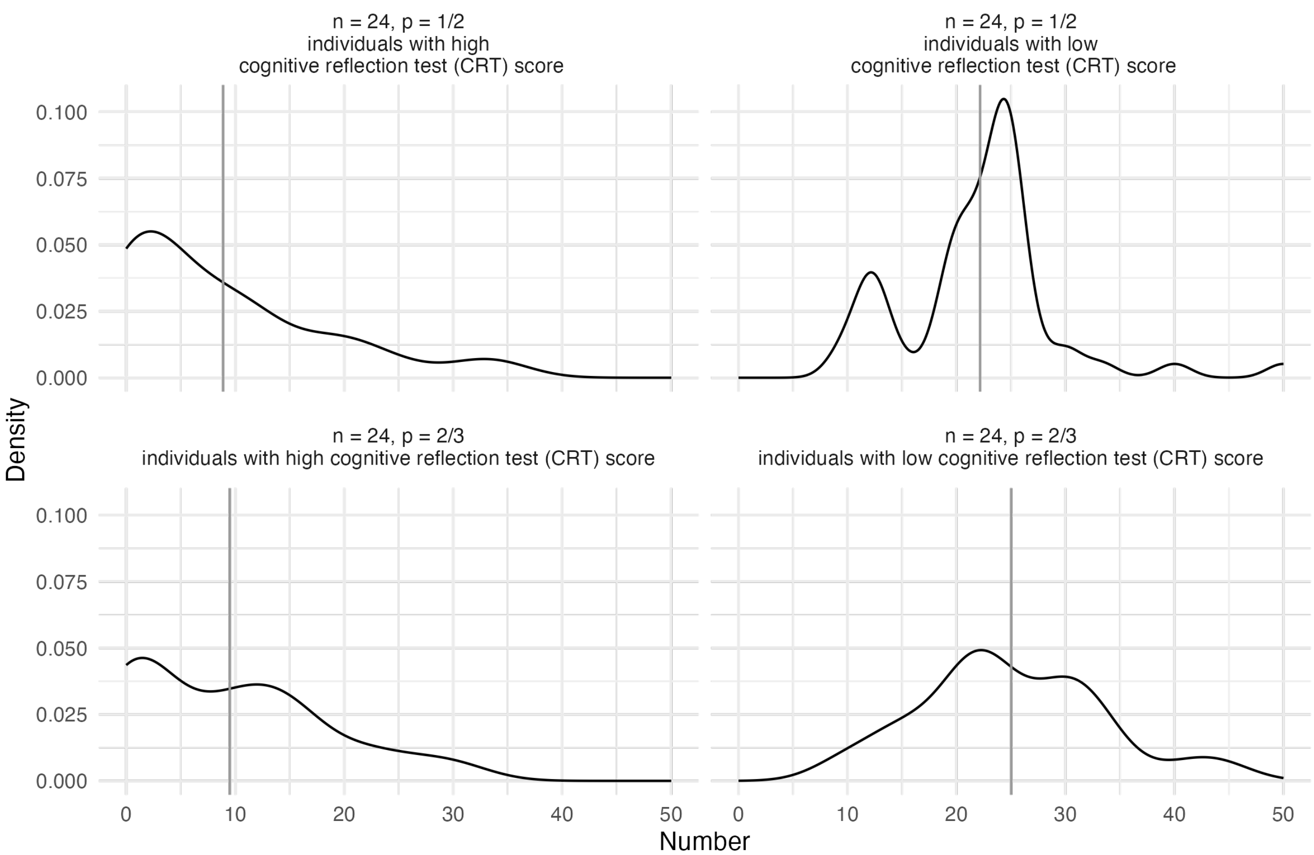

Figure 5: Distribution of GPT-4 responses in experiments from Table 2, lines 13–16.

context of the *Guess the number* game.

To explore the factors influencing this deviation, we analyze the frequency of specific words used in the explanations of GPT-4 as independent variables. The rationale behind this approach is to investigate the linguistic features and terminologies that might correlate with the AI's answers relative to human answers. By identifying patterns in language use that significantly affect the disparity in responses, we aim to uncover insights into the cognitive processes and strategies employed by both real and artificial players. The use of Lasso regression, with its capacity for variable selection and regularization, enables us to pinpoint the most salient linguistic predictors while mitigating the risk of overfitting, thereby ensuring the robustness of our findings.

The Lasso regression model is formalized as follows:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left| y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{1}$$

where:

- N denotes the total number of observations;
- y_i represents the dependent variable, specifically the absolute value of the discrepancy between the responses of real participants and the artificial intelligence for the i-th observation;
- x_{ij} signifies the j-th independent variable for the i-th observation, corresponding to the frequency of specific words within the GPT-4 discussions;
- β_0 is the intercept term of the model, and β_j are the coefficients associated with each independent variable, which are estimated by the Lasso regression;
- λ is a regularization parameter that influences the extent of penalty applied to the size of the coefficients, facilitating variable selection by driving some coefficients to zero and thereby reducing the model's complexity.

The objective function consists of two parts: the first term is the residual sum of squares, quantifying the model's fit to the data, while the second term is the Lasso penalty, enforcing a constraint

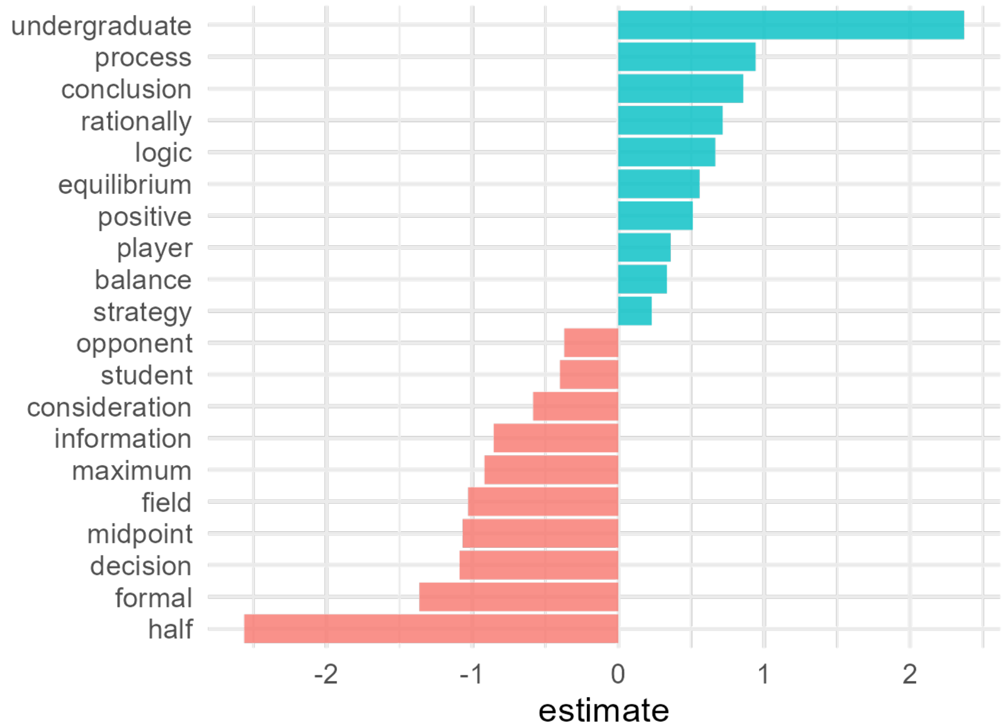

Figure 6: Lasso regression results

on the magnitude of the coefficients to prevent overfitting and improve the interpretability of the model by selecting a more parsimonious set of predictors.

This analytical strategy contributes to our understanding of the nuances in AI-based decision making. We not only evaluate the capabilities of GPT in mimicking human strategic thinking but also shed light on how the decision making process may facilitate the AI's alignment with human behavior in complex decision-making scenarios.

The results of Lasso regression are presented on Figure 6. The figure is depicting Lasso regression coefficients for various words. The chart is divided into two sections by a vertical zero line. Words with bars extending to the right of the zero line have positive coefficients, indicating a positive relationship between the frequency of these words in the GPT-4 discussions and the absolute deviation between the responses of real and artificial players. Conversely, words with bars extending to the left have negative coefficients, suggesting an inverse relationship.

The positive coefficients associated with words like 'rationality', 'logic', 'equilibrium', and 'strategy' suggest that as GPT-4 more frequently invokes concepts tied to formal reasoning and strategic analysis, its numbers diverge more from those of human players. This implies that when the AI emphasizes a rational and calculated approach, as denoted by these terms, it may not align closely with the heuristic or intuitive methods that human participants might employ.

This divergence could be indicative of a fundamental difference in the decision-making processes between humans and artificial intelligence. While GPT-4 may default to a more systematic and theoretically grounded line of reasoning, humans are known to utilize a blend of cognitive shortcuts, emotions, and bounded rationality, which might lead to less predictable but more human-typical answers. Thus, the AI's reliance on rationality and logic, while theoretically sound, could cause it to stray from the common human responses that are influenced by a wider array of psychological and contextual factors.

4 Conclusion

This study presents an innovative exploration into the dynamics of the Keynesian beauty contest by integrating artificial intelligence, specifically GPT-4, to simulate human economic behavior. Our experiments have highlighted GPT-4's advanced capability to adapt its strategies by recognizing and responding to the level of sophistication in human opponents, offering insights into the potential of AI in modeling economic agents. However, decisions made by AI do not always strictly adhere to changes in the numerical parameters of the model.

The findings demonstrate a closer alignment of AI-generated guesses to strategic thinking compared to human participants, revealing a disparity in decision-making processes and rationality. The use of Lasso regression analysis further enriches our understanding by identifying linguistic predictors that influence the deviation in strategies between humans and AI.

This research lays a foundation for future studies on the implications of integrating AI into economic modeling and decision-making. The convergence of AI's calculated rationality with human intuition and behavior opens new avenues for enhancing predictive models and designing economic policies that account for the bounded rationality in human economic activities.

References

Bosch, A., Garcia-Montalvo, J., Nagel, R., & Satorra, A. (2002). One, two, (three), infnity,...: Newspapers and lab beauty contest games. *The American Economic Review*, *92*(5), 1687–1701.

Brañas-Garza, P., García-Muñoz, T., & González, R. H. (2012). Cognitive effort in the beauty contest game. *Journal of Economic Behavior & Organization*, *83*(2), 254–260.

Breitmoser, Y. (2012). Strategic reasoning in p-beauty contests. *Games and Economic Behavior*, *75*(2), 555–569.

Burnham, T. C., Cesarini, D., Johannesson, M., Lichtenstein, P., & Wallace, B. (2009). Higher cognitive ability is associated with lower entries in a p-beauty contest. *Journal of Economic Behavior & Organization*, *72*(1), 171–175.

Castagnetti, A., Proto, E., & Sofianos, A. (2023). Anger impairs strategic behavior: A beauty-contest based analysis. *Journal of Economic Behavior & Organization*, *213*, 128–141.

Duffy, J., & Nagel, R. (1997). On the robustness of behaviour in experimental 'beauty contest'games. *The Economic Journal*, *107*(445), 1684–1700.

Grosskopf, B., & Nagel, R. (2008). The two-person beauty contest. *Games and Economic Behavior*, *62*(1), 93–99.

Ho, T. H., Camerer, C., & Weigelt, K. (1998). Iterated dominance and iterated best response in experimental 'p-beauty contests'. *The American Economic Review*, *88*(4), 947–969.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? Tech. rep., Cambridge, MA, USA.

Kocher, M., Sutter, M., & Wakolbinger, F. (2014). Social learning in beauty-contest games. *Southern Economic Journal*, *80*(3), 586–613.

Kocher, M. G., & Sutter, M. (2005). The decision maker matters: Individual versus group behaviour in experimental beauty-contest games. *The Economic Journal*, *115*(500), 200–223.

Leder, J., Häusser, J. A., & Mojzisch, A. (2013). Stress and strategic decision-making in the beauty contest game. *Psychoneuroendocrinology*, *38*(9), 1503–1511.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American Economic Review*, *85*(5), 1313–1326.

Nagel, R. (2008). Experimental beauty contest games: levels of reasoning and convergence to equilibrium. In C. R. Plott, & V. L. Smith (Eds.) *Handbook of experimental economics results*, vol. 1, chap. 45, (pp. 391–410). North-Holland.

Nagel, R., Bühren, C., & Frank, B. (2017). Inspired and inspiring: Hervé moulin and the discovery of the beauty contest game. *Mathematical Social Sciences*, *90*, 191–207.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, (pp. 1–22).

Stahl, D. O. (1996). Boundedly rational rule learning in a guessing game. *Games and Economic Behavior*, *16*(2), 303–330.

Stahl, D. O. (1998). Is step-j thinking an arbitrary modelling restriction or a fact of human nature? *Journal of Economic Behavior & Organization*, *37*(1), 33–51.

Sutter, M. (2005). Are four heads better than two? an experimental beauty-contest game with teams of different size. *Economics Letters*, *88*(1), 41–46.

Weber, R. A. (2003). "Learning" with no feedback in a competitive guessing game. *Games and Economic Behavior*, *44*(1), 134–144.