The Role of Responsibility in AI's Strategic Risk-Taking*

Gregory DeAngelo†, Bryan C. McCannon‡, William Wyatt§

February 24, 2026

Abstract

Charness and Jackson (2009) investigated the implication of risk taking when assuming the responsibility of another person. Their primary finding was that people are less willing to take risks on behalf of others than when the risk affects only themselves. We replicate their seminal contribution using Large Language Models (LLMs) as participants instead of humans to determine if AI behavior differs from human choices. Recent innovations like LLMs (e.g., ChatGPT) have been integrated into most of our human interface technology. LLMs are also being leveraged heavily to make decisions on behalf of humans. In this research we examine how risk decisions, and the associated responsibility, are altered in a world with AI making decisions. Our results indicate, LLMs on average are 73.2% more likely than humans to take on risk when their choices impact another player. Alternatively, when making risk decisions that impact only themselves, the risk taking decisions of the LLMs are nearly identical to human risk taking choices.

*Data and analysis code are available at https://osf.io/a6f3u/. The authors declare no competing interests.

†Associate Professor of Decision Sciences & Director, Computational Justice Lab, Claremont Graduate University, 150 E. Tenth Street, Claremont, CA 91711; gregory.deangelo@gmail.com

‡Robert S. Eckley Endowed Chair in Economics & Dean, School of Business and Economics, Illinois Wesleyan University, 1312 Park Street, Bloomington, IL 61701; bryan.c.mccannon@gmail.com

§Claremont Graduate University, 150 E. Tenth Street, Claremont, CA 91711; william.wyatt@cgu.edu

1 Introduction

Large Language Models (LLM) have proliferated throughout industry and private use as a general purpose tool (Chen et al., 2023). This pervasive new technology has seen widespread adoption, without particular discretion or scrutiny (Bender et al., 2021; Colombo et al., 2024). LLMs represent a powerful tool that enables questions to be quickly answered from a simple prompt. This technology presents a new level of information availability that is not limited merely to the packaging and presentation of information, as LLMs are not just providing information retrieval but active participation in reasoning and decision-making.

With easy access to computer programs capable of reasoning, humans are deferring their own reasoning (Liu et al., 2024). Such delegation ranges from straightforward linguistic tasks to complex ethical judgments. This deferral of our reasoning to computational intelligence could significantly affect, and likely is affecting, important aspects of human life when humans rely on LLMs to make complex and risk-involved choices. This paper explores these implications by examining the impact of LLM decision making, particularly in situations involving responsibility for others.

We build on the seminal contribution of Charness and Jackson (2009). They explored risk-taking behavior when decision makers are asked to be responsible for others' well-being. In their experimental framework, which we replicate and extend here, treatments differ by whether a decision affects only the decision maker or also affects others. By swapping human agents with LLM agents, we compare the differences in behavior related to risk and responsibility. Both studies review the behavior of how agents take risk associated with responsibility: We examine when an action only affects the agent and compare that to when the agent is responsible for another "person".

We make a number of important observations. First, when playing only for itself, LLMs behave strikingly similar to humans. Importantly, though, there is a substantial difference in risk-taking when decisions affect another. While Charness and Jackson (2009) showed that humans are ***less*** likely to take a risk when making the decision for the pair, the LLM is ***more*** likely to take the risk that can generate the larger reward but can also lead to a lower payoff.

Second, LLM's behavior is shown to be sensitive to the prompt used. When context is removed, decision making for others still exhibits more risk-taking, but the baseline level is sensitive to how the game is presented. Further, the LLM responds to the persona it is asked to take. Its behavior is contingent on which gender it is assigned to take, its age, and its personality type. The LLM is less likely to take risks when prompted to take the role of a woman, old person, or an introvert when responsible for others. It is more likely to take the risk when it is prompted to be

competitive and is even more responsive to responsibility with this persona.

Third, its willingness to take a risk depends on the certainty which its opponent coordinates with it on the higher payoff. When the opponent randomizes selecting each strategy with equal likelihood it is less willing to take a risk. When the opponent's behavior puts more weight on playing the strategy that has the possibility of generating the higher payoff, it is more willing to take that risk.

After reviewing the context of LLM in game theory and responsibility in section 2, we proceed into the specifics on how we created LLM agents in section 3 along with the specific game they are set to play in subsection 3.1. Our analysis goes beyond the replication in Charness and Jackson (2009) to investigate how altering specific attributes influences LLM behavior. Specifically, after replication in section 4, we examine the effect of concealing domain-specific information in subsection 5.1, assigning different personas to the LLMs subsection 5.2, and observing their responses when paired with random agents subsection 5.3. Finally, we explore questions that are provoked from this research in the discussion (section 6).

2 Literature Review

LLMs, such as ChatGPT, have rapidly expanded their presence across various sectors, significantly influencing corporate communication, consumer content creation, and decision-support systems (Liang et al., 2025; Bender et al., 2021). Although LLMs provide notable gains in efficiency and productivity, their widespread use raises ethical concerns, including the reproduction of biases, the dissemination of misinformation, and potential job displacement (Bender et al., 2021; Liu et al., 2024). Moreover, individuals frequently underestimate the influence of LLMs in their personal decision-making (Meng, 2024).

Literature exploring risk-taking behavior on behalf of others has provided mixed findings. Charness and Jackson (Charness and Jackson, 2009) demonstrated that decision-makers tend to become more cautious when their choices impact others due to increased perceived responsibility. Conversely, Pahlke et al. (2012) observed scenarios in which individuals took greater risks when deciding for others, particularly when accountability was lower or emotional detachment was higher. Factors such as social distance, perceived accountability, and risk domain influence these behaviors significantly (Montinari and Rancan, 2018; Stone et al., 2002; Wang et al., 2023). These insights are critical when considering how AI systems should behave when entrusted with decisions impacting human outcomes, such as

in healthcare or financial advisory roles.

Experimental research has recently examined LLM behavior in various decision-making contexts. Mei et al. (2024) provides one of the first investigations by having ChatGPT engage in a series of classical games and personality assessments. They show, for example, that LLMs tend to be more pro-social than humans. McCannon (2024b) illustrates that LLMs are influenced by humans' social norms when making allocation decisions in Dictator Games. It shares more of its endowment when confronted with a more-generous social norm. Meng (2024) shows that GPT-based models often mimic human norms of fairness and cooperation in economic games. Yet, LLMs have also demonstrated distinctive decision-making patterns, tending toward statistically optimal choices rather than human-like caution or risk aversion (Chen et al., 2025; Guo, 2023). Relatedly, McCannon (2024a) investigates LLMs behavior in strategically uncertain environments providing evidence that behavior is far from game-theoretic predictions. Thus, it is an open question the degree to which LLMs are willing to take risks, and how it responds to being responsible for human welfare. Given that LLMs are being asked to support and even make decisions on behalf of humans, its willingness to take risks and how it responds to being responsible is an important, open question.

3 Methods

Our study parallels and replicates the methods and analysis used in Charness and Jackson (2009). Their study ran two treatments:

- **Play for Self:** Two players who's actions only affect their own utility. This is served as a benchmark behavior.
- **Play for Pair:** Two teams of pairs where only one person from each team is making a decision, acting as the decision maker for the pair. The decision maker of the team does not communicate with their teammate but only makes decisions on their behalf.

In Charness and Jackson (2009) there was six separate experiments with sixteen participants in each, lasting one hour per session. In each session, all participants played fifteen rounds of both treatments: Play for Self and Play for Pair. The participants engaged in a simulation that was identical to the Stag-Hare game (see Table 1). When players choose Stag, they choose to be more risky. The key insight is comparing the baseline, Play for Self, to the responsibility treatment, Play-for-Pair. In this way the game captures how behavior changes when the agent's decision is responsible

for another player's well-being. We mirror this experiment by replacing human agents with LLM agents, investigating the behavior of LLMs relative to humans.

3.1 Stag-Hare Game

To measure risk and responsibility we will use the Stag Hare game, where choosing Stag is risky with higher reward, but requires cooperation of the other player. The alternative strategy, Hare, is the safe option but results in a lower payoff.

Table 1: Stag-Hare Game

	Stag	**Hare**
Stag	(9,9)	(1,8)
Hare	(8,1)	(8,8)

Payoff matrix for the Stag-Hare game. Playing Stag is risky and can result in a payoff of 1 if the other player does not commit to the same strategy. Playing Hare always results in a payoff of 8. Playing Stag has the potential payoff of 9 if both players cooperate. There are two stable Nash Equilibria: (Stag, Stag) and (Hare, Hare). The Mixed Strategy Nash Equilibria involve playing Stag with a probability of $\frac{7}{8}$ and Hare with a probability of $\frac{1}{8}$.

The Stag-Hare game used in Charness and Jackson (2009) used a specific software to create a network type game to interface the participants in their experiment. We also use the Stag-Hare game and use the same payoff matrix as in Charness and Jackson (2009).

3.2 Using LLMs

Our goal is to interface with an LLM in a way that simulates a human in an experiment. We exclusively used the ChatGPT api and the GPT-4o-mini model. To mirror the experiment in Charness and Jackson (2009) with LLM agents, two types of prompts were constructed:

- A **System Prompt**, which defines the behavior of the LLM's response.
- A **User Prompt**, to provides game rules, game history, and request the LLM to make a decision.

To play a round of the Stag-Hare game, we query the ChatGPT api for each LLM agent, and use the response to record a decision, then update the user prompt for the next round of the session. Our system prompt emulates a human participant by defining behavior guidelines stating that the decision is a real-life scenario. The system prompts can

be found in the Online Appendix B. There are two user prompts, one for each treatment type. In each, we simulate a human trial by providing the game it is tasked to play, along with a simple request to make a decision. The Play for Self user prompt is the baseline behavior we want to compare against the Play for Pair user prompt. See Online Appendix B for information about the Play for Self prompt and the Play for Pair prompt.

To represent a session of fifteen rounds, we query the LLM to make a choice between either Stag or Hare, then successive moves would include the previous decisions in the user prompt in a brief formatted summary. To see the way the summary is formatted see Online Appendix B. The summary gives the LLM memory of previous game decisions it has made in the session. In either Play for Self or Play for Pair treatment, there are only two LLMs participating, playing with or against each other and being queried for a decision. Despite Play for Pair being a pair of teams, the non-participating teammate is effectively a non-existent narrative character. At 150 sessions, with fifteen rounds, and two LLMs we observe a total of 4,500 recorded decisions per treatment type.

3.3 Variations of the Experiment

Initially, we replicate Charness and Jackson (2009) as analogously as we can before we vary the experiment. The first change we make is to remove all occurrences of the terms **Stag** and **Hare** from the experiment. This is to conceal the Stag-Hare game from the LLM and suppress any domain knowledge it may have acquired during training on game theory. To do this, we replace the term Stag with Option-A, and Hare with Option-B, in Play for Self user prompt and the Play for Pair user prompt. To see the prompts used see Online Appendix B.

Robustness checks are then performed that vary the personas of the LLM. The personalities change age, sex, along with character types. To do this we insert into the system prompt different broad character types. The persona prompts are located in Online Appendix B.

The last variation we test is running the LLM against random agents instead of another LLM. To do this, we call the LLM to make a decision, either Stag or Hare, then instead of calling another LLM to make a decision, we simply run a stochastic program to decide Stag or Hare independent of previous decisions. We will test against two forms of random agents:

- **Uniform Random Agent:** With a 50% chance of choosing Stag or 50% chance of choosing Hare.
- **Mixed Strategy Agent:** Using the mixed strategy, the agent chooses Stag with a $\frac{7}{8}$ probability and a $\frac{1}{8}$ proba-

bility to play Hare.

Each of these variations are to test the robustness of our original results. The analysis of the sessions of these variations are analyzed in the following section.

4 Results

We implement an LLM to play the Stag-Hare game as described in subsection 3.2. We define the system prompt which dictates the behavior of the agent, then we have the user prompt for Play for Self treatment to determine a baseline risk and Play for Pair treatment to see how the LLM changes risk taking choices when the decision affects another "person".

For the replication of Charness and Jackson (2009), we run 500 sessions, each with fifteen repeated rounds, and two LLMs interacting each round to result in 15,000 observations. On average, the LLM plays Stag more often then Hare in the Play for Pair treatment. In Table 2, we present the summary of the total observed strategies played.

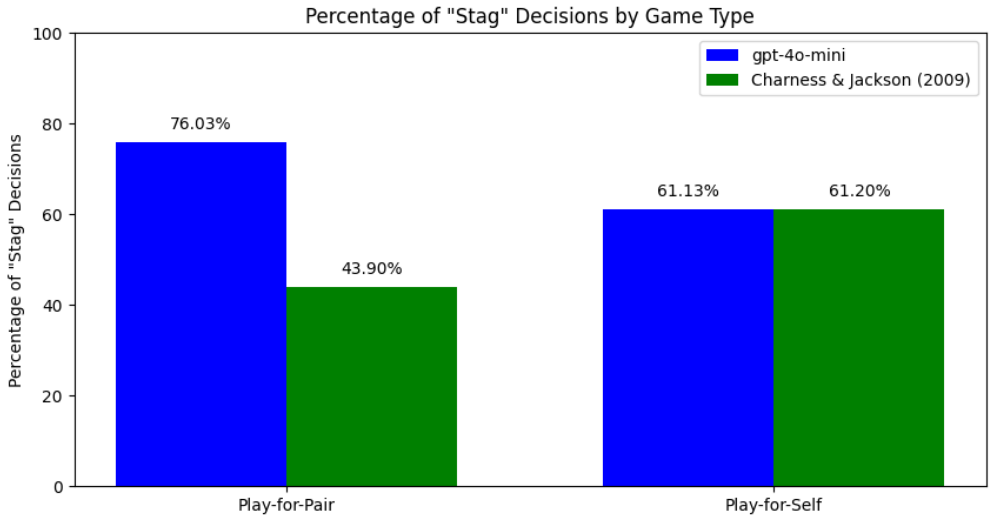
Table 2: Risk Taking by Game

Game	Stag	Hare
Play-for-self	9169 (61.13%)	5831 (38.87%)
Play-for-pair	11404 (76.03%)	3596 (23.97%)

Results from the decisions made in the 15,000 observations presented. This accounts for 500 independent sessions, with 15 rounds each, involving two LLMs making decisions.

LLMs chose Stag significantly more often under Play for Pair (76.03%) than Play for Self (61.13%), $\chi^2(1) = 772.00$, $p < .001$. Compared to humans, LLMs differed significantly in Play for Pair ($\chi^2(1) = 374.04$, $p < .001$) but not in Play for Self ($\chi^2(1) = 0.00$, $p = .927$) Charness and Jackson (2009). This is visually presented in Figure 1; showing an increase of 32 percentage points for choosing Stag for LLMs.

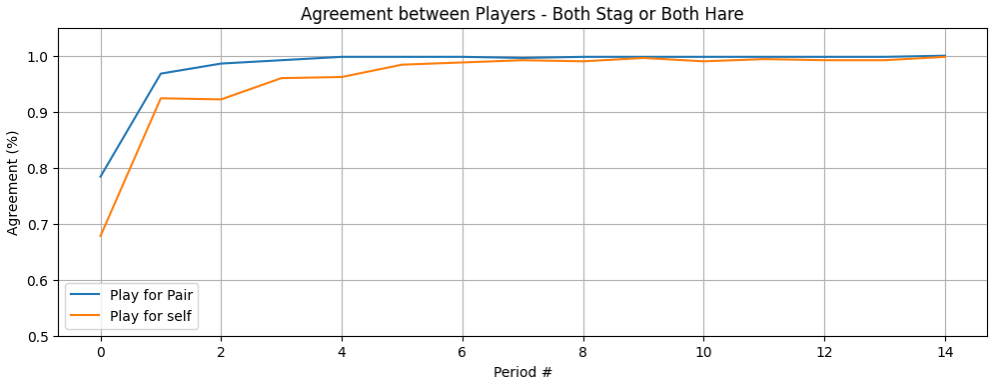
Figure 1: Baseline Result



An LLM playing the Stag-Hare game against another LLM compared to the human trial results from Charness and Jackson (2009). The blue bars show the average play of Stag for the LLM and the green bars are from the human trials. The Play-for-Pair and Play-for-Self treatments had two players for 15 rounds in a session with 500 sessions, for a total of 15,000 observations for each game. In the baseline, Play for Self, the LLM is almost matching the human results at around 61%. In the Play for Pair treatment, when the player is responsible for another person's well-being, the LLM acts more risky playing Stag 76% of the time. This is a 32 percentage point increase from the experiment with humans.

Additionally, an important observation of LLM's decision making is its choices in the early rounds. Once the LLMs coordinate on a Nash Equilibrium, they stabilize on it and do not deviate. Figure 2 illustrates.

Figure 2: Switching



Choices made over the 15 periods of the 500 sessions from the main game with a total of 15,000 observations, or 1,000 observations per x-axis tick are presented. The y-axis denotes agreement value of 1 if both LLMs are in agreement: If they both chose either Stag or Hare. If they are in disagreement, where one chooses Stag and the other Hare, the value is 0.

Additionally, Table 3 evaluates mis-coordination. A miscoordination is defined by one of the players choosing

Stag and the other Hare.

Table 3: Mis-coordination

Game	Both Stag	Mis-coordination	Both Hare
Play-for-pair	5629 (75.1%)	146 (1.9%)	1725 (23.0%)
Play-for-self	4425 (59.0%)	319 (4.3%)	2756 (36.8%)

Outcomes: number of observations (percentages), by category. In the default game there were 500 sessions, fifteen rounds making a total of 7,500 rounds. Here we see there is few miscoordination rounds where one player chose Stag and the other chose Hare. With only 146 rounds in Play for Pair and 319 rounds in Play for self having a miscoordination. The other shows that there were 5,629 rounds in play for Pair were both LLM agents played Stag. In Play for Self, there was 4,425 rounds that they both played Stag.

Failure to coordinate occurs only 4.3% of the time in Play for Self and 1.9% in Play for Pair out of 7,500 rounds.

5 Robustness Checks

To test the robustness of our results, we examine three variations of the game.

5.1 Concealed Game

The first robustness test that we make in the experiment is to remove all mentions of "Hare" and "Stag" from the LLM's prompts. The updated prompt for *Play for Self* and the updated *Play for Pair* prompt can be found in Online Appendix B. The system prompt remains unchanged. This adjustment is meant to reduce the LLM's prior knowledge of game theory and the Stag-Hare game. By replacing "Stag" with *Option A* and "Hare" with *Option B*, we make the context less recognizable to the model, encouraging it to respond in a way that is not based on previous research relating to the Stag-Hare game.

For this experiment, we conducted 150 sessions, each consisting of fifteen rounds with two LLMs, resulting in 4,500 total observations. We found that 150 sessions were enough to achieve stable results. In the appendix we go into detail about number of trials necessary for a given maximum error. At 150 sessions, look at just first round (150 observations) would give us an maximum margin of error at 95% confidence of $\pm 8\%$. In aggregate, 150 sessions results in 4,500 observations leading to a margin of error of $\pm 1.46\%$ at 95% confidence.

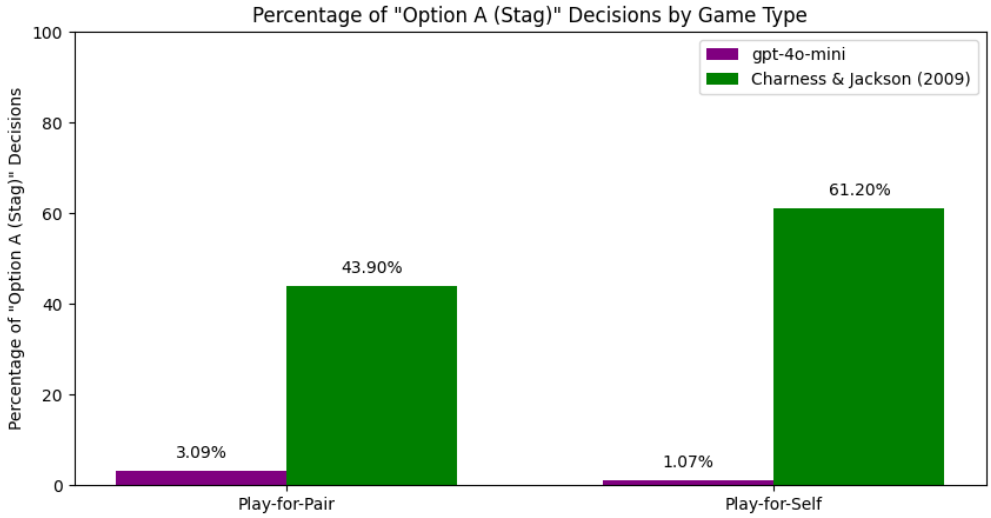
Under this setup, the LLMs overwhelmingly chose the safer option, Option B (formerly Hare). Table 4 contrasts our observation with Charness and Jackson (2009).

Table 4: Concealed Game Results

Game	Option A - Stag	Option B - Hare
Play-for-self	48 (1.07%)	4452 (98.9%)
Play-for-pair	139 (3.09%)	4361 (96.9%)

Results from the Concealed Game are presented where instead of Stag and Hare, we provided Option A and Option B respectively. The total 4,500 observations per treatment type. This accounts for 150 independent sessions, with 15 rounds each, involving two LLMs making decisions.

Figure 3: Concealed Play



An LLM playing the Stag-Hare game against another LLM compared to the human trial results from Charness and Jackson (2009) is presented. The purple bars show the average play of Stag for the LLM and the green bars are from the human trials. For Play-for-Pair and Play-for-Self respectively had two players for 15 periods in a session with 150 sessions; This totals to 4,500 observations for each game.

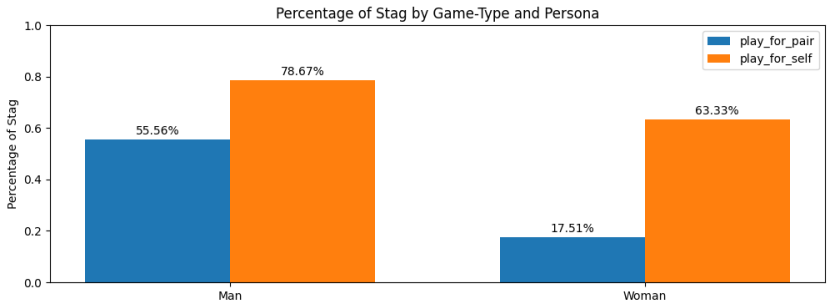
There is an obvious reduction in the frequency with which the LLM plays Stag, as compared to human decision making. As with the previous results, though, when the LLM is playing for itself it only plays Stag 1.07% of the time. When it is responsible for another in Play for Pair it plays Stag 3.09% of the time. Similar to the main results, it responds positively to the responsibility by almost tripling the rate at which it takes the risk; the difference between Play for Pair and Play for Self is statistically significant ($\chi^2(1) = 44.23$, $p < .001$). Compared to human decision making, this represents a significant shift in behavior.

5.2 Persona Test

The second variation introduces a slight change to the prompt, presenting the LLM with a persona. In this test, we run two LLMs with the same assigned persona through 150 sessions of the Stag-Hare game for each treatment type: Play for Self and Play for Pair. This persona test is designed to further assess how a few keywords can influence the LLMs' play style.

One dimension of particular interest is gender. We consider this in Figure 4.

Figure 4: Persona: Gender



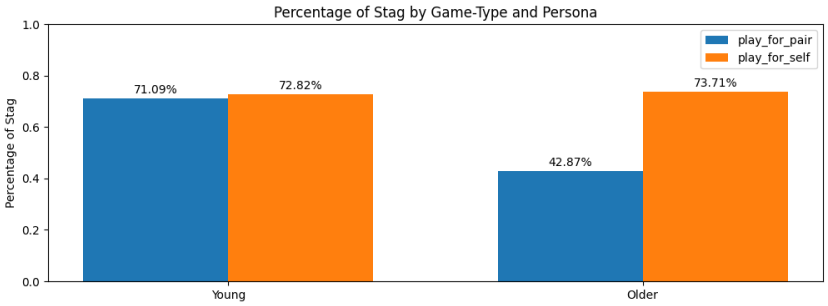
On the left, we prompt the LLM to be a man and on the right we prompt the LLM to be a woman. The yellow bars represent the Play for Self treatment where its actions only affect itself. The blue bars represent the Play for Pair treatment, when the agent is responsible for another person.

When we insert "You are a man" into the system prompt. The results, shown in the left columns in Figure 4, reveal a 15 percentage point increase in Play for Self and a 38 percentage point increase in Play for Pair, as compared to a prompt telling it "You are a woman". This difference between personas is statistically significant ($\chi^2(1) = 1402.89$, $p < .001$ for Play for Pair). This suggests that priming the LLM as a man makes it more willing to take risks. For both genders, the LLM is less willing to take a risk when responsible for another: men chose Stag 55.56% in Play for Pair versus 78.67% in Play for Self ($\chi^2(1) = 543.43$, $p < .001$), while women chose Stag 17.51% versus 63.33% ($\chi^2(1) = 1959.79$, $p < .001$). The reduction is substantially greater when it is told "You are a woman". This indicates that when primed as a woman, the LLM becomes much less willing to choose the risky option when making decisions that affect others. This compares well with laboratory experiments of humans (Beckamn et al., 2016). These results suggest that simply assigning a sex to the LLM significantly alters its behavior in ways that align with conventional normative expectations.

We also explore LLMs behavior when age is prompted. We consider behavior when prompted that "You are young;

a 25-year-old individual" and when prompted that "You are old; a 75-year-old individual". Figure 5 depicts the results.

Figure 5: Persona: Age

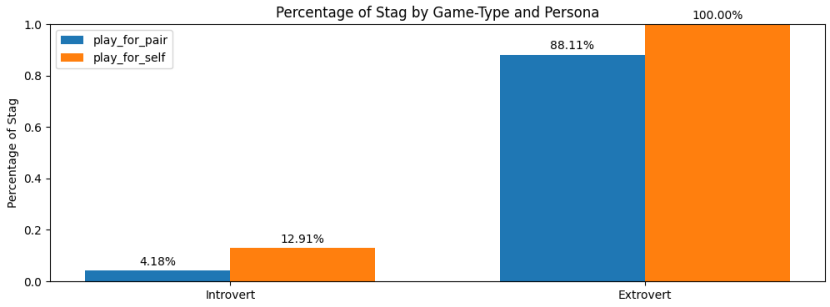


On the left, we have presented the LLM with a young persona and on the right we give the LLM an older persona. The yellow bars represent the Play for Self treatment where its actions only affect itself. The blue bars represent the Play for Pair treatment, when the agent is responsible for another person.

When we introduce age-related cues, the LLM exhibits similar behavior between the two game conditions, choosing Stag 74% of the time in Play for Self as an older individual and 73% of the time when playing as a younger individual, when playing only for itself. In Play for Pair, age-contingent behavior arises. It is less likely to take the risk when playing as an older individual. Interestingly, the LLM's behavior when a young individual is not affected by being responsible for another ($\chi^2(1) = 3.26$, $p = .071$). In contrast, the older persona shows a significant reduction in risk-taking under responsibility, choosing Stag 42.87% in Play for Pair versus 73.71% in Play for Self ($\chi^2(1) = 879.17$, $p < .001$). The difference between young and older personas in Play for Pair is also significant ($\chi^2(1) = 729.93$, $p < .001$). This suggests that when primed as an older individual, the LLM adopts a more cautious approach when responsible for another player, aligning with expected behavioral differences between younger and older individuals. This is consistent with behavior in human subjects (Albert and Duffy, 2012).

The work by Mei et al. (2024) suggests that LLM's personality differs from the median human. Consequently, we consider how its behavior changes when prompted to take on a particular personality. Specifically, we insert "You are an introverted person" and "You are an extroverted person" as prompts. The results are presented in Figure 6.

Figure 6: Persona: Personality

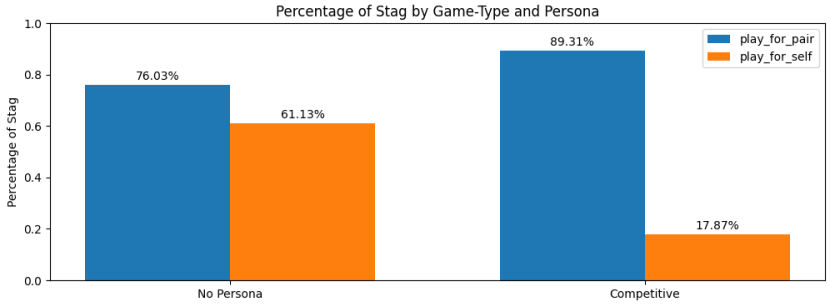


On the left, the LLM is prompted with an introverted persona and on the right we give the LLM an extroverted persona. The yellow bars represent the Play for Self treatment where its action's only affect itself. The blue bars represent the Play for Pair treatment, when the agent is responsible for another person.

We see that the introverted persona takes substantially less risk, regardless of treatment, choosing Stag only 4.18% in Play for Pair versus 12.91% in Play for Self ($\chi^2(1) = 218.49$, $p < .001$). The extroverted persona in general takes on much more risk. In particular, when playing for itself, it chooses Stag 100% of the time, dropping to 88.11% in Play for Pair ($\chi^2(1) = 566.69$, $p < .001$). The difference between these personas in Play for Pair is dramatic ($\chi^2(1) = 6374.87$, $p < .001$). Again, for both treatments when asked to be responsible for another, less risk is taken. This suggests that the extroverted and introverted personas push the LLM toward extreme and opposite decision making, nearly eliminating variation in responses based on game conditions.

Finally, we also test the impact of a competitive personality traits. We insert the prompt "You are a competitive person," and compare this to the baseline behavior in Figure 7.

Figure 7: Persona: Competitive



This table shows two LLM agents playing against each other. On the left we have presented the LLM without a persona (the default LLM mode) and on the right we give the LLM an competitive persona. The yellow bars represents the Play for Self treatment where its action's only affect itself. The blue bars represent the Play for Pair treatment, when the agent is responsible for another person.

When we insert "You are a competitive person", the LLM plays Stag 89.31% of the time in Play for Pair and only 17.87% in Play for Self ($\chi^2(1) = 4614.80$, $p < .001$). A competitive LLM responds positively to responsibility by taking more risk. This is a substantially greater jump in behavior than observed without a prompt ($\chi^2(1) = 368.61$, $p < .001$ comparing competitive to no persona in Play for Pair).

The persona tests demonstrates that small modifications to the system prompt can lead to drastically different responses in the Stag-Hare game. The personas produce behavior that aligns with expected social, normative traits. In all persona-based tests, LLMs played against an identical persona (i.e., two LLMs given the same priming statement), with no mixing of different personalities. Even so, these findings highlight how sensitive LLMs are to subtle priming through keywords in their system prompts.

5.3 Random Agents

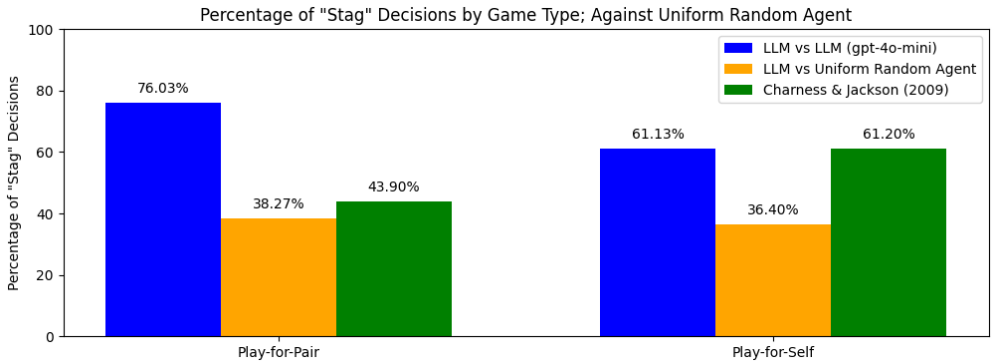
In our third robustness check, we replace one of the LLMs with a random agent. Instead of having two LLMs play successive rounds of Stag-Hare, where each LLM makes decisions based on the game's history, one LLM is substituted with a random agent that selects moves solely based on a random number generator. This allows us to assess whether its behavior with another LLM differs from its behavior when it plays with a human. We implement two types of random agents:

- **Uniformly Random Agent:** This agent has a 50-50 chance of choosing Stag or Hare in each round.
- **Mixed Strategy Random Agent:** This agent chooses Stag with a probability of $\frac{7}{8}$ and Hare with a probability of $\frac{1}{8}$, based on the Mixed Strategy Nash Equilibrium defined in Table 1.

The purpose of introducing a random agent is to observe how the LLM responds to uncertain, varying behavior of its opponent.

First, we consider its behavior against an opponent selecting each action with an equal likelihood in each round. Figure 8 presents the results.

Figure 8: Playing Against a 50-50 Randomizer



The average choice of Stag when an LLM plays against a uniformly random agent is presented. The blue bars show for reference when the LLM plays against another LLM. The green bars are the human trials from Charness and Jackson (2009). The yellow bars show the LLM against the uniform random agent. We see that the LLM becomes less risky choosing hare more when the opponent is uniformly randomly choosing Stag or Hare.

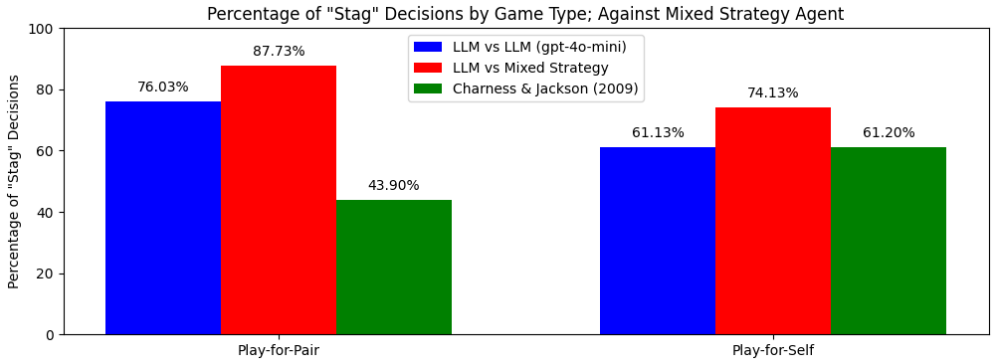
The results indicate a significant decrease in Stag play when the LLM faces this agent.

- In Play for Self, the LLM plays Stag 36.4% of the time, a 25 percentage point decrease from the default behavior.
- In Play for Pair, the LLM plays Stag 38.3% of the time, a 38 percentage point decrease from the default behavior.

When faced with a uniformly random opponent, the LLM significantly reduces its choice of Stag, indicating that it adjusts its strategy in response to unpredictability. Behavior when it makes a decision for itself, as compared to when it must take responsibility for another, is statistically indistinguishable ($\chi^2(1) = 2.71$, $p = .100$). This suggests that the uncertainty of its opponent's behavior dwarfs the importance of responsibility.

Next, we consider its behavior when it plays against an opponent who selects the Mixed Strategy Nash Equilibrium. Figure 9 presents the results.

Figure 9: Playing Against the Mixed Strategy Nash Equilibrium



The average choice of Stag when an LLM plays against a random agent using the mixed strategy is presented: The random agent chooses Stag 7/8th of the time. The blue bar is for reference when the LLM plays against another LLM. The green bars show the results from the human trial in Charness and Jackson (2009). The red bars show on average how often the LLM plays Stag given a random agent that plays Stag 7/8th of the time.

The random agent follows a mixed strategy, playing Stag with a probability of $\frac{7}{8}$ (87.5%) and Hare with a probability of $\frac{1}{8}$ (12.5%), as determined by the Mixed Strategy Nash Equilibrium in the Stag-Hare game. When paired with this agent, the LLM plays Stag 74.13% of the time in Play for Self, representing a 13 percentage point increase from the default behavior. In Play for Pair, the LLM plays Stag 87.73% of the time, a 12 percentage point increase from the default behavior. Unlike the uniform random agent, this difference between treatments remains statistically significant ($\chi^2(1) = 21.60$, $p < .001$). Interestingly, this Play for Pair result closely aligns with the mixed strategy probability of 87.5%, suggesting that the LLM adapts to its opponent's behavior.

6 Discussion

Our study replicates the work of Charness and Jackson (2009) by how LLMs behave in a strategic decision making environment, particularly when responsibility for others is introduced. The findings suggest that while LLMs broadly align with human decision making in Play for Self, they diverge significantly in Play for Pair, exhibiting greater risk-taking behavior when responsible for another player. This contrasts with human participants, who tend to become less risk-taking when making decisions on behalf of others. This suggests a fundamental difference in how AI models interpret responsibility.

When the Stag-Hare terminology was replaced with neutral terms (Option A and Option B), LLMs overwhelmingly avoided risk, shifting towards the safer option (formerly Hare). This indicates that explicit game-theoretic framing

primes the model towards a more strategic equilibrium, whereas concealing game terms leads to more conservative play. Priming LLMs with personas such as gender, age, and personality traits significantly altered their decision making patterns in ways that reflect human social norms. It is worthwhile to explore further how small changes in the prompts results in wildly different behavior and why this drastic behavioral change is not random.

When paired with random agents, LLMs adjusted their strategies based on the opponent's behavior. Against a mixed strategy agent, LLMs' choices closely matched equilibrium probabilities, while against a uniform random agent, LLMs became significantly less risky. This suggests that LLMs are highly responsive to the predictability of their opponent and attempt to stabilize on an optimal strategy even when faced with erratic behavior. When placing the LLMs against random agents showed just how persistent the LLMs were at forming a consensus on one of the equilibria in the Stag-Hare game. Further research is necessary to understand why the LLMs are so sticky in their choices to coordinate and form consensus with one another and resistant to change their strategy.

These results show that LLMs, while capable of replicating human-like strategic decision-making, exhibit unique biases in their interpretation of responsibility and risk. This raises important questions about how AI should be designed, and how sensitive that design is, for decision-making in real-world applications where responsibility is a key factor.

Further research could explore further why LLMs take on more risk when responsible for others and whether additional constraints or ethical guidelines in system prompts could align their behavior more closely with human norms. For example, understanding which components of the prompts are most sensitive to the result of the experiment. Additionally, testing across different LLM architectures and training methods could help identify model-specific biases and improve AI decision-making in socially sensitive areas.

The exposition here into the behavior of risk taking for an LLM is an interesting look into replicating a human trial with LLM agents. Further studies are required to understand why small changes result in such diverse results and if its possible to use LLMs as a comparable substitute for humans in game theoretic studies.

References

Albert, S. M. and Duffy, J. (2012). Differences in Risk Aversion between Young and Older Adults. *Neuroscience and neuroeconomics*, pages 3–9.

Beckamn, S. R., DeAngelo, G., Smith, W. J., and Ning, W. (2016). Is social choice gender-neutral? reference dependence and sexual selection in decisions toward risk and inequality. *Journal of Risk and Uncertainty*, 52(3):191–211.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Charness, G. and Jackson, M. O. (2009). The Role of Responsibility in Strategic Risk-Taking. *Journal of Economic Behavior & Organization*, 69(3):241–247.

Chen, Q., Ge, J., Xie, H., Xu, X., and Yang, Y. (2023). Large language models at work in china's labor market.

Chen, Y., Kirshner, S. N., Ovchinnikov, A., Andiappan, M., and Jenkin, T. (2025). A manager and an ai walk into a bar: does chatgpt make biased decisions like we do? *Manufacturing & Service Operations Management*.

Colombo, E., Mercorio, F., Mezzanzanica, M., and Serino, A. (2024). Towards the Terminator Economy: Assessing Job Exposure to AI through LLMs.

Guo, F. (2023). Gpt in game theory experiments.

Liang, W., Zhang, Y., Codreanu, M., Wang, J., Cao, H., and Zou, J. (2025). The Widespread Adoption of Large Language Model-Assisted Writing Across Society.

Liu, J., Xu, X., Nan, X., Li, Y., and Tan, Y. (2024). "Generate" the Future of Work through AI: Empirical Evidence from Online Labor Markets.

McCannon, B. C. (2024a). Artificial Intelligence and Strategic Uncertainty: Can AI Play Mixed Strategies? Working Paper.

McCannon, B. C. (2024b). Artificial Intelligence is a Pro-Social Norm Complier. *Economics Letters*, 241:111828.

Mei, Q., Xie, Y., Yuan, W., and Jackson, M. O. (2024). A Turing Test of Whether AI Chatbots are Behaviorally Similar to Humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121.

Meng, J. (2024). AI Emerges as the Frontier in Behavioral Science. *Proceedings of the National Academy of Sciences*, 121(10):e2401336121.

Montinari, N. and Rancan, M. (2018). Risk Taking on Behalf of Others: The Role of Social Distance. *Journal of Risk and Uncertainty*, 57(1):81–109.

Pahlke, J., Strasser, S., and Vieider, F. M. (2012). Risk-taking for others under accountability. *Economics Letters*, 114(1):102–105.

Stone, E. R., Yates, A. J., and Caruthers, A. S. (2002). Risk taking in decision making for others versus the self 1. *Journal of Applied Social Psychology*, 32(9):1797–1824.

Wang, D., Han, D., Sun, L., Zhou, M., Hao, L., and Hu, Y. (2023). Self-other (s) risk decision differences in different domains in the chinese context: A social value theory perspective. *Psychology Research and Behavior Management*, pages 4117–4132.