Reproducing the “Flip a Coin or Vote” experiment with GPT-3.5 and GPT-4o

An simulation study to the suitability of LLMs as participants in economic and behavioural experiments.

TUDelft Delft University of Technology

Reproducing the "Flip a Coin or Vote" experiment with GPT-3.5 and GPT-4o

An simulation study to the suitability of LLMs as participants in economic and behavioural experiments

By

M.J. (Michiel) Hutschemaekers

(4968069)

in partial fulfilment of the requirements for the degree of

Master of Science

in Applied Physics

at the Delft University of Technology,

To be defended publicly on Wednesday, August 28, 2024, at 1:45 PM.

Supervisor:	Dr. P.W.G. Bots	TU Delft
Thesis committee:	Dr. Ir. R. Van Bergem	TU Delft
	S. Renes Ph.D. LL.M.	TU Delft

TU Delft

Preface

To the reader of this master's thesis,

During both my bachelor's and master's studies, I always tried to find projects that fit within my interests. Thus, I started the academic year with the electives package Artificial Intelligence (AI) Specialisation. However, even after multiple courses on AI, I still felt that I could not fully grasp the subject. The fact that different professors at the university had very conflicting perceptions about the use of AI in education and research made the subject even more intriguing.

This led me to the subject of my master's thesis. Choosing an interesting subject is not always the most strategic choice. For this thesis, the novelty of the research area, combined with my far-from-perfect programming skills, presented many challenges. However, due to my interest in the study, I still enjoyed the entire process, and I am proud of the end result. I hope this is reflected in this work and is noticeable as you read.

I would like to thank the following individuals, in random order: First, I want to express my gratitude to Rutger van Bergem and Sander Renes. Although not officially, you both functioned as my content-related supervisors. Your enthusiasm for the subject inspired me in every meeting we had. And Sander, thank you for lending me your experiment and its findings. Second, I would like to thank my official supervisor, Pieter Bots. I appreciate the way you approach things formally. Moreover, it was very helpful to have someone somewhat removed from the topic to spar with. For all, I would like to thank you for the feedback throughout the work.

Finally, I would like to express my gratitude to the OpenAI Researchers Access Program, which partly supported the costs of running all the simulations with GPT-3.5 and GPT-4o. I was surprised and grateful to see that they support research at the master's level as well.

I hope you enjoy reading my work.

Michiel Hutschemaekers
Amsterdam, 2024

Notes: The image of the front page was created with the assistance of DALL·E 2. Furthermore, my statement of using other AI tools for this work can be found in appendix K. The code and all data files can be found here https://github.com/Michielhs/Data-and-Code

Executive Summary

The latest state-of-the-art large language models (LLMs) are implicit computational models of humans due to how they are trained and designed (Horton, 2023). This implies that LLMs can be used as participants in economic and behavioural experiments. Compared to research with human participants, LLMs offer cost-effective and time-efficient alternatives (Horton, 2023), and reasoning outputs that provide valuable insights into the rationale behind decisions (Guo, 2023). However, their technical and ethical limitations have sparked ongoing debates about the usefulness of LLMs in experimental research (Gui & Toubia, 2023; Harding et al., 2023).

This study contributes to the research area by reproducing the "Flip a Coin or Vote" experiment conducted by Hoffmann & Renes (2021), with GPT-3.5 and GPT-4o as participants. The experiment aimed to determine when more efficient decision rules might replace inefficient ones within groups of three players. The experiment is played as follows (Hoffmann & Renes, 2021):

> Groups of three players decide whether to implement a public project. In part 1, they select a decision-making mechanism from two options. Players choose their preferred mechanism, knowing only the range of possible valuations (ex-ante rounds) or their exact valuation (ad-interim rounds). In part 2, a mechanism is randomly chosen from one of the players' selections and used to determine whether the project is implemented. Payoffs depend on the project's implementation: each player receives their private valuation if implemented and zero if not.

Including three-player groups, dynamic strategic elements, a two-part structure with additional information for the players, and both straightforward and complex decision mechanisms make the Flip a Coin or Vote experiment a suitable choice. The results of the simulations were analysed to answer the following research question:

To what extent are the state-of-the-art LLMs able to be 'good' participants in the Flip a Coin or Vote experiment?

For an LLM to be considered a 'good' participant, it must (1) understand the experiment, (2) make rational decisions, and (3) make decisions that are human-like to a certain degree.

Based on these three criteria, GPT-3.5 does not seem to be a 'good' participant, particularly in the first part of the experiment. GPT-3.5 (1) struggles to fully understand the rules of the experiment, especially in calculating payoffs; (2) has difficulties interpreting negative and positive valuations; (3) fails to match the percentage of rational choices observed in the lab results in part 1; and (4) demonstrate results that deviate

too much with the lab results to be considered human-like. Furthermore, GPT-3.5 does seem to be very sensitive to contextual framing (Loré & Heydari, 2023)

In contrast, GPT-4o shows greater promises as a 'good' participant. GPT-4o indicates to understand the game's rules correctly and demonstrates a stronger ability to make rational choices. In part 1 of the experiment, GPT-4o matches, and sometimes surpasses, the percentage of rational choices made by human participants. The nuanced must be made that this comparison is purely based on the outcomes. The human participants were not asked to rationalise their decision. Furthermore, GPT-4o appears less sensitive to contextual framing.

However, in part 2 of the experiment, GPT-4o's ability to make rational choices becomes inhumanly accurate, indicating a hyper-accuracy distortion (Aher et al., 2023). The results suggest that GPT-4o assumes rational behaviour from other players (R. Liu et al., 2024), leading it to tell the truth when stating its valuation consistently. In contrast, the lab results show more instances of lying, indicating that human players also expect others to lie. In this respect, GPT-3.5's behaviour aligns more closely with the lab results in part 2, as it does not make inhumanly rational decisions.

GPT-4o's inhumanly rational decisions in part 2 of the experiment, could enhance its suitability for simulations aimed at institutional design. By assuming near-complete rationality, such simulations can identify optimal outcomes and efficient institutions, which can then be benchmarked against real-world scenarios to account for human irrationalities. Even with entirely rational players, these simulations may reveal unintended or suboptimal consequences if players strategically exploit the institutions. However, caution is needed, as over-reliance on these models may overlook the irrationalities inherent in human behaviour.

In summary, GPT-4o demonstrates a greater ability to make rational and consistent choices than GPT-3.5,. In part 1 of the experiment, its decisions align more closely with the lab results. However, in part 2, GPT-4o exhibits inhumanly rational decisions, which do not capture the nuances of human decisions observed in the lab. Conversely, GPT-3.5 shows better alignment with human behaviour in part 2 of the experiment and specific aspects of the ex-ante rounds. However, GPT-3.5's inconsistent and often irrational decision-making in part 1 suggests limitations in handling more complex experiments.

Researchers have several ways to influence the outcomes of simulations with LLMs. For GPT-3.5, the primary goal was to increase the rationality of its decisions through experimental controls. A promising approach is the step-by-step reasoning approach proposed by Kojima et al. (2023), which increased GPT-3.5's rational decision-making by 22%. For GPT-4o, the focus was on aligning its preferences more closely with lab results. The most promising approach involved prompting GPT-4o to 'make human-like' decisions, although the results were statistically insignificant compared to the benchmark run.

The findings of this study have implications the scientific research area. It contributes to the ongoing debate about the role of LLMs as participants in economic and behavioural experiments, offering insights into their potential for simulating complex strategic group dynamics.

Future research should be focused on how to steer the results of LLMs towards more human-like results, either through prompting options or through training and fine-tuning of the LLM model. Furthermore, future research should establish best practices for developing a more standardised approach to conducting simulation research with LLMs. Final, as noted in multiple studies and confirmed by this research, LLMs are sensitive to contextual framing. Since human participants are also influenced by contextual framing to some degree, it would be beneficial to explore whether there is any correlation between the framing effects on LLMs and those on humans.

List of acronyms

AGV	Arrow d'Aspremont-Gérard-Varet
AI	Artificial Intelligence
API	Application Programming Interface
CoT	Chan of Though
EC	Experimental Control
GPT	Generative Pre-Trained Transformer
IBEX	Interdisciplinary center for Behavioral Experiments
LLM	Large Language Model
MRQ	Main Research Question
NSQ	Non-Implementation Status Qup
SCM	Structural Causal Model
SM	Simple Majority Voting
SRM	Subpopulation Representative Model
SQ	Sub-Question
RAND	Random Implementation (flip of a fair coin)
TE	Turing Experiment

List of tables

List of figures

Content

1 Introduction

The latest advancements in Large Language Models (LLMs) have profoundly impacted the field of Artificial Intelligence (AI). With ChatGPT gaining over 100 million active monthly users in the first two months since its launch, it has established itself as the fastest-growing consumer application (Makridakis et al., 2023). These models are designed to generate human-like text and perform natural language processing tasks (Horton, 2023). The performance of LLMs has increased substantially in the last decade, with the training compute doubling every six months on average (Sevilla et al., 2022). Scaling up the performance of LLMs has led to new capabilities, often referred to as *emergent abilities* (Wei et al., 2022).

One important emergent ability is that LLMs function as implicit computational models of humans due to how they are designed (Horton, 2023). This ability has sparked debate in the social sciences about using LLMs in simulation research. While different studies have found promising results (e.g., Horton, 2023; Aher et al., 2023), there are also studies stressing the technical and ethical limitations (e.g. Gui & Toubia, 2023; Harding et al., 2023). Consequently, while the potential of LLMs in simulation research is substantial, their adoption must be carefully evaluated, given these technical and ethical limitations.

Researchers at the Delft University of Technology have recognised this potential of using LLMs in economic and behavioural experiments. Their research project[1] 'Large Language Model Multi-Actor Tool to Automate Economic Experiments', recently received funding from the Dutch Research Council (NWO). The first experiments they plan to test are from the Interdisciplinary Centre for Behavioural Experiments (IBEX) lab[2]. The experiments at IBEX are focused on choice and behaviour of individuals or groups. Furthermore, the experiments are used to analyse and evaluate institutional designs in socio-technical systems.

This master thesis aims to support the research project by reproducing the 'Flip a coin or vote' experiment, focused on the implementation and efficiency of social choice mechanisms in groups of 3 (Hoffmann & Renes, 2021), with GPT-3.5 and GPT-4o participants. By simulating this experiment using the GPT-3.5 and GPT-4o models[3], this study will assess the decision-making process of LLMs in a three-player experiment with dynamic strategic elements under various experimental conditions, such as differing prompts and model settings, to determine how closely their choices can mimic from human choices.

The insights derived from this work will contribute to the ongoing debate about the applicability of LLMs in simulation. The setup involving three players with dynamic strategic elements will provide new insights,

[1] More information can be found here: https://www.tudelft.nl/en/2024/tbm/sander-renes-receives-nwo-funding l
[2] More information can be found here: https://ibex.tudelft.nl/
[3] This research uses the 'base' models, without fine-tuning.

given that most studies focus on one or two-player experiments (e.g., Horton, 2023; Aher et al., 2023; Akata et al., 2023). Furthermore, based on the practical experiences gained through the OpenAI API, recommendations will be given regarding the Multi-Actor Tool for future experiments.

There are compelling reasons to utilise Large Language Models (LLMs) for economic and behavioural experiments. Experiments with human participants are often costly, time-consuming, and sometimes not even feasible (Horton, 2023). LLMs can guide empirical work with the ability to pilot experiments, explore the parameter space, and test whether choices are sensitive to wording and framing (Horton, 2023). Furthermore, LLMs can be used as a substitution for human participants in cases with legal, ethical, or privacy considerations (Aher et al., 2023). Their potential in simulation research lies in their ability to replicate human-like decision-making processes and explore a wide range of scenarios.

However, opinions among researchers on the suitability of LLMs as substitutes for human participants remain divided. Ethical issues arise from the potential misrepresentation of groups, which can uphold stereotypes (Harding et al., 2023; A. Wang et al., 2024). In addition, LLMs have been demonstrated to underperform in tasks requiring coordination (Akata et al., 2023) and have shown decreased mathematical capabilities (Fan et al., 2022). In contrast, they have also produced inhumanly accurate responses in specific tasks (Aher et al., 2023), highlighting their complex and sometimes contradictory performance capabilities.

Thus, while LLMs offer a promising, cost-effective, and time-efficient option for economic and behavioural experiments, scepticism persists due to ethical concerns and technical limitations. This research will explore these performance issues, which stem from the technical limitations, by analysing how LLMs make decision-rule choices in a group setting under various experimental controls. By evaluating LLMs' performance and human-like behaviour in a more complex experiment, insights can be gained into their suitability for experimental research aimed at institutional design in socio-technical systems.

This thesis is structured as follows: Chapter 2 reviews the relevant literature, identifying two key knowledge gaps that inform the formulation of the main research question. Chapter 3 will briefly describe the Flip a Coin or Vote experiment (Hoffmann & Renes, 2021) and its suitability for reproduction with GPT-3.5 and GPT-4o. Chapter 4 outlines the sub-questions (SQs) of the study, details the research approach used to address them, and establishes the evaluation metrics. Chapter 5 describes the conceptualisation of the experiment as a simulation model. Chapter 6 presents the results of the various simulations. Chapter 7 reflects on these results, connecting them to the literature reviewed in Chapter 2. Chapter 8 offers recommendations based on identified best practices and insights from working with the OpenAI API. Chapter 9 summarises the key findings and suggests potential directions for future research.

2 Theoretical Background

This chapter presents the literature review conducted for this thesis. For the review, I used the snowballing procedure, following the guidelines of Wohlin (2014). The whole methodology is presented in Appendix A. The objective of the literature review is twofold. First, it aims to define and explain the relevant concepts (section 2.1). Second, it seeks to enhance understanding of the latest research on experimental studies involving LLMs. For this, I will delve into the 'humanness' of LLMs (section 2.2), the experiments that have already been done with LLMs (section 2.3), and the related risks and challenges (section 2.4). The results are used to identify the knowledge gaps for this thesis (section 2.5).

2.1 Definitions of Concepts

Artificial Intelligence (AI) encompasses a variety of definitions and applications. In the book *Artificial Intelligence: A Modern Approach,* Russell & Norvig (2010) explored the full range of the field to synthesise the subfields into a common framework. Building on this framework, Xi et al. (2023) define AI as follows: "Artificial Intelligence is a field dedicated to designing and developing systems that can replicate human-like intelligence and abilities." This definition aligns closely with the focus of this thesis.

The field of AI encompasses many different models, with *large language models* (LLMs) among the most popular ones, gaining substantial interest in both academic and industrial domains (Chang et al., 2024). Akata et al. (2023) define LLMs as "deep learning models with billions of parameters trained on huge corpora of text." The parameter count for these models is on the order of 10^{11} and growing (Korinek, 2023). The goal of LLMs is to be able to generate human-like text or perform other natural language processing tasks (Horton, 2023). Most researchers use a non-technical definition similar to that of Akata et al. (2023). However, Argyle et al. (2023) provide a more technical definition:

> Language models like GPT-3 are a conditional probability distribution $p(x_n|x_1, \cdot\cdot\cdot, x_{n-1})$ over tokens, where each x_i comes from a fixed vocabulary. By iteratively sampling from this distribution, a language model can generate arbitrarily long sequences of text. However, before it can generate text, a language model like GPT-3 requires "conditioning," meaning that it must be provided with initial input tokens comprising $\{x_1, ..., x_{n-1}\}$.

The most used LLMs are the *ChatGPT* models designed by OpenAI. GPT is an acronym for Generative Pre-trained Transformer, an advanced family of LLMs (Teubner, 2023). The original model, GPT-1, was introduced in June 2018 and consisted of 117 million parameters (Marr, 2024). Seven months later, in February 2019, OpenAI released GPT-2, a significant upgrade with 1.5 billion parameters. GPT-3 again was a huge leap forward, with 175 billion parameters trained on more than 45 terabytes of text (Argyle et al., 2022). At the end of 2022, GPT-3.5 was released, gaining more than 100 million users in the first two months (Korinek, 2023) and is estimated to produce a volume of text equivalent to all printed works of humanity every fourteen days (Thompson, 2023).

All the models up to GPT-3.5 operated mainly on text-based *prompts* to get responses. Prompts can be defined as “a set of instructions provided to an LLM that programs the LLM by customising it and/or enhancing or refining its capabilities” (White et al., 2023; S. Liu et al., 2023). Or, more straightforward, “a request submitted to a language model to receive a response” (Google Cloud, 2024). The release of GPT-4 introduced capabilities to process text and image inputs (OpenAI, 2023). Furthermore, the language capabilities also increased significantly. According to the technical report by OpenAI, GPT-4 exhibits human-level performance on professional and academic benchmarks. OpenAI's latest flagship model, GPT-4o, launched in May 2024, further expands these capabilities to accept and generate outputs from text, audio, images, and video (OpenAI, 2024).

The fact that GPT-4 demonstrates human-level performance is not surprising. According to Sevilla et al. (2022), the training compute of deep learning models has doubled every six months on average over the last decade, implying a thousand-fold increase every five years, a significant increase compared to a doubling roughly every 20 months before 2010. This increase in performance has led to impressive abilities in a wide range of domains, in addition to their natural language-generation abilities (Chen et al., 2023). These abilities are referred to as *emergent abilities*. According to Wei et al. (2022), “an ability is emergent if it is not present in smaller models but is in larger models.” In recent years, the most interesting abilities for researchers have emerged (Korinek, 2023), such as the implication that LLMs can be used in simulation as computational models of humans (Horton, 2023).

In order for LLMs to be used in simulation research, they must be able to make rational decisions. The concept of *rationality* is defined in various of ways and differs between different scientific domains (Mele & Rawling, 2004). Kacelnik (2006) mentions three categories of rationality derived from Philosophy and Psychology (PP-rationality), Economics (E-rationality), and Evolutionary Biology (B-rationality). Focusing on economic rationality, Jones’s (2021) study demonstrates that the concept of rationality varies considerably in major Economics textbooks used in the UK, highlighting the different views on the concept.

In game theory, Mohammad et al. (2019) define a rational player as “one who always chooses the most preferred outcome given the expectations of their opponent(s)”. Turocy & Von Stengel (2001) define a rational player as “if he seeks to play in a manner which maximizes his own payoff.” This thesis follows this definition of Turocy & Von Stengel (2001), given the game theoretical nature of the experiment that is being reproduced. In addition, this definition links with the observation of Jones (2021) that ‘rationality’ is often used as a shorthand for maximisation or consistency. Turocy & Von Stengel (2003) define game theory as follows: "Game theory is the formal study of decision-making where several players must make choices that potentially affect the interests of the other players." A game, in this context, refers to the formal description of a strategic situation (Turocy & Von Stengel, 2003).

2.2 Humanness of AI

In order to effectively utilise LLMs in simulation research, LLMs must be able to provide human-like responses. This *humanness* of AI has long been questioned (Dillion et al., 2023). The latest research across various disciplines presents mixed results and perspectives, which I will detail in the following subsequent section.

A consensus exists among researchers regarding the impressive language capabilities of LLMs. Text generated by the older GPT-3 model is already strikingly difficult to distinguish from those generated by humans (Brown et al., 2020; Argyle et al., 2022; Dillion et al., 2023). This can be seen in its ability to capture patterns of grammar, cultural knowledge, and conversational rhythms present in natural language (Argyle et al., 2023). Furthermore, ongoing tests indicate even better results with the newer GPT-4 model (OpenAI, 2023).

Beyond the language capabilities, the ChatGPT models also show other capabilities. For example, recent advancements have shown GPT-4 passing the bar exam (OpenAI, 2023) and GPT-3 solving complex mathematical problems (Zong & Krishnamachari, 2023). Dillion et al. (2023) also observed that the *moral judgements* of GPT3.5 were extremely well aligned with human standards. However, Dillion et al. (2023) emphasise that their proof is anecdotal, suggesting that definitive conclusions about the moral judgment capabilities of LLMs should be approached with caution.

2.2.1 Rationality

The studies of Chen et al. (2023) and Fan et al. (2024) looked at the rationality of LLMs but reached different conclusions. Chen et al. (2023) reported that ChatGPT demonstrated a high level of rationality in all their four decision-making tasks. Furthermore, they found that the rationality scores were consistent across different demographic characteristics. Chen et al. (2023) use "a classic notion of economic rationality in revealed preference analysis that captures the extent to which a decision maker maximizes some well-behaved utility functions for the given budget constraints."

In contrast, Fan et al. (2024) identified significant shortcomings in LLMs' rationality. They concluded: "LLMs struggle to build desires based on uncommon preferences, fail to refine belief from many simple patterns, and may overlook or modify refined belief when taking actions." This indicates that GPT-4 demonstrates substantial disparities compared to humans in game theory. The disparity may stem from differing definitions of rationality. Fan et al. (2024) drew on earlier works by Zagare (1984) and Osborne & Rubinstein (1995) to outline three traits of a rational player:

1. *build a clear desire for the game.*
2. *refine belief about uncertainty in the game*
3. *take optimal action based on desire and belief*

2.2.2 Subpopulation Representation

The most debated area regarding the humanness of AI is subpopulation representation. From an optimistic perspective, the studies of Argyle et al. (2023) and Simmons & Hare (2023) show promising results of using LLMs to estimate subpopulation representative models. Argyle et al. (2023) introduce the concept of *algorithmic fidelity*, defined as "the degree to which the complex patterns of relationships between ideas, attitudes, and socio-cultural contexts within a model accurately mirror those within a range of human subpopulations." To measure the algorithmic fidelity, they present four different criteria.

1. Social Science Turing Test
2. Backward Continuity
3. Forward Continuity
4. Pattern Correspondence

These criteria were applied to three studies that Argyle et al. (2023) reproduced with an LLM. All three reproductions are based on surveys focused mainly on political views in America. Their conclusions are surprisingly promising, indicating that:

A) the same language model, when properly conditioned, is able to produce outputs biased both toward and against specific groups and perspectives in ways that strongly correspond with human response patterns along fine-grained demographic axes.

B) that algorithmic fidelity is a crucial attribute of tools like GPT-3 because it demonstrates that these language models can be used prior to or in the absence of human data.

Simmons & Hare (2023) present a cautiously optimistic perspective. They introduce the concept of *subpopulation representative models* (SRMs), defined as "models that approximate to some useful degree certain characteristics of human subpopulations." SRMs could provide an alternate way to measure public opining with subpopulations based on demographic, geographic, or political segments (Simmons & Hare, 2023). However, in contrast to Argyle et al. (2023), Simmons & Hare (2023) emphasise that this does not come without risks such as "social stereotypes and unfair discrimination," "lower performance for languages and social groups," and "potential for hallucinations and misinformation."

The risks mentioned by Simmons & Hare (2023) are reasons for a more pessimistic perspective. LLMs are most likely to give accurate estimates about Western English speakers because of the data they are trained on (Dillion et al., 2023; Harding et al., 2023; L Wang et al., 2024; A Wang et al., 2024). For example, the ChatGPT models tend to overrepresent the views of liberal, higher-income, and higher-educated people (Santurkar et al., 2023). Consequently, LLMs cannot give accurate estimates for those not represented in the training data. For instance, even within the USA, LLMs fail to capture accurate estimates of people over 65 years old and highly religious (Santurkar et al., 2023). A. Wang et al. (2024) note that 37% of the world population has never accessed the internet and thus is unlikely to be well-represented in LLM training data. In addition, the current methods of fine-tuning LLM performances, especially with human feedback loops, further exacerbate this issue (Harding et al., 2023).

Grossman et al. (2023) make another important point here. While scientists aim to study 'pure' LLMs to simulate social-cultural biases present in humans, ethical constraints require engineers to protect the LLM from these biases, training and fine-tuning the models for the world that 'should be' rather than 'what is.' This undermines the validity of AI-assisted social science research. Harding et al. (2023) and A. Wang et al. (2024) go even further, arguing that due to this inability to portray subpopulations, LLMs cannot replace human research participants indefinitely. These different standpoints and scientific results pose essential considerations for future research with LLMs.

2.3 Experimental Research with Large Language Models

Despite the contrasting scientific views on whether to use LLMs as human research participants, multiple studies have already tried experimenting with LLMs. In the following section, I will explain the motivations behind these experiments and provide an overview of the experiments that have already been simulated.

2.3.1 The Motivation

For Horton (2023), the core of the argument to use LLMs as participants is twofold: "LLMs – by nature of their training and design – are (1) computational models of humans and (2) likely possess a great deal of latent social information." For (1), the engineers have designed the LLMs to respond in ways similar to humans. For (2), due to being trained on a vast amount of data, these models likely capture latent information on topics such as economics, laws, decision-making heuristics, and social preferences (Horton, 2023).

Based on his argument, LLMs have sparked interest across disciplines, such as marketing, computer science, economics, psychology, and political science, in leveraging LLMs to simulate humans (Gui & Toubia, 2023). The appeal of using LLMs in research lies in their ability to pilot experiments for preliminary insights, to cheaply explore the parameter space, and to test whether behaviour seems sensitive to wording (Horton, 2023). These insights can be used to guide actual empirical work. The study of Aher et al. (2022)

mentions that LLMs can be used in scenarios "where it is costly to carry out experiments on humans due to considerations regarding scale, selection bias, monetary cost, legal, moral, or privacy considerations."

Moreover, Guo (2023) argues that reasoning outputs from LLMs provide insights into the rationale behind decisions, which is often challenging to explore in human studies. Gui & Toubia (2023) mention that LLMs allow researchers to simulate how humans would respond to stimuli or questions in different contexts. In addition, with LLMs, researchers can do a clean within-subject experiment because the LLM does not remember having seen the previous prompt (Horton, 2023)

Akata et al. (2023) give another noteworthy perspective on why to do scientific research with LLMs as participants. Their argument does not originate from the objective of gaining more insights into human behaviour but from gaining more insights into the behaviour of LLMs themselves. According to them, it is important to understand how LLMs behave in social settings, especially as they increasingly permeate diverse applications and interact more frequently with humans and other agents.

2.3.2 Experiments with LLMs

The use of LLMs in experimental research spans a variety of disciplines, demonstrating their potential to replicate and extend traditional human subject studies. Horton (2023) conducted experiments classical in behavioural economics. Aher et al. (2022) conducted experiments in behavioural economics, psycholinguistics, and social psychology. However, most studies focused on relatively simple, well-known experiments in game theory (Akata et al., 2023; Guo, 2023; Loré & Heydari, 2023; Fan et al., 2024). The following sections will briefly cover all the aforementioned studies.

2.3.2.1 Behavioural Economic Experiments (Horton, 2023)

Horton (2023) did four classical experiments in the behavioural economics literature with the GPT-3 model. His experiments consisted of (1) the unilateral dictator games, (2) survey responses to economic scenarios, (3) decision-making scenarios, and (4) a hiring scenario. He varied the input prompts, telling the LLM how to behave regarding of inequity aversion versus self-interest or socialist versus liberalist. His main conclusion was that the approach seems promising: "it can qualitatively recover findings from experiments with actual humans."

2.3.2.2 Turing Experiments (Aher et al., 2023)

Aher et al. (2023) took a different approach. They introduced a new type of test, called a *Turing Experiment* (TE), used for "the evaluation of an AI system in terms of its use in simulating human behaviour in the context of a specific experiment, like a human subject study" (Aher et al., 2023). Four different experiments were conducted, each with its origin in a different scientific discipline.

The first experiment was The Ultimatum Game, used in behavioural economics to study fairness and rationality. The second experiment was Garden-Path Sentences, used to study parsing in psycholinguistics. The third experiment was Milgram's shock experiment, used in psychology to study authority obedience. The final experiment was Wisdom of Crowds, used to study collective intelligence across disciplines. In addition to doing experiments from different disciplines, Aher et al. (2023) also used different models. In total, they evaluated 5-6 LLMs through the OpenAI's API. Their persona input incorporated a wide array of personas with different titles, racial backgrounds, and surnames, simulating a pool of 1,000 unique participants.

The experiments by Aher et al. (2023) showcased mixed results; larger models tended to provide more accurate simulations in the initial three experiments, aligning closely with historical data from human studies. However, in the last experiment, the larger models did not outperform the smaller ones. This was due to an 'hyper accuracy distortion,' "where larger and more aligned LMs simulate human subjects that give unhumanly accurate answers" (Aher et al., 2023). These results show that LLMs have the potential to reproduce experiments, however, the models can be flawed in specific experiments.

2.3.2.3 Game Theory Experiments

Game theory experiments represent the most frequently simulated studies in the exploration of LLMs' capabilities, yielding results that vary from optimistic to pessimistic interpretations. Table 1 presents an overview of all the game theoretical experiments done, the LLM models being used, and the additional prompting used as behaviour instructions, such as personality traits or geographical instructions.

Table 1. Overview of the game theoretical experiments

Source	Games	Models	Additional prompting
Guo (2023)	- The ultimatum game - Prisoner's dilemma	- GPT-4	Traits such as fairness and selfishness
Loré & Heydari (2023)	- Prisoner's dilemma - Stag hunt - Snowdrift - Prisoner's delight	- GPT-3.5 - GPT-4 - LLaMa2	Contextual framings, such as friendly interactions, business meetings, or team interactions
Akata et al. (2023)	- 144 different finitely repeated 2x2-games	- GPT-3.5 - GPT-4	
Fan et al. (2024)	- Dictator game - Rock-paper-scissors - Ring network game	- GPT-4	

The most optimistic findings come from Guo (2023). His study shows the GPT's potential, as a valuable tool in social science research, because GPT exhibits behaviours similar to human responses. The conclusion is nuanced by highlighting limitations, such as the sensitivity of GPT models' responses to input prompts.

Loré & Heydari (2023) found a similar limitation, namely that GPT-3.5 is highly sensitive to contextual framing. In addition, they show that GPT-3.5 is incapable of strategic behaviour and conclude that "the algorithm is unsophisticated at best and spiteful at worst." GPT-4 and Llama-2, on the other hand, show more strategic behaviour. However, context still influences both, and GPT-4 shows a strong bias for socially optimal actions. Their results indicate that LLMs are unfit for strategic behaviour.

Akata et al. (2023) show that LLMs generally perform well in games where valuing self-interest pays off. However, underperform well at games that require coordination, even when faced with simple strategies. With this way op playing, selfish and uncoordinated, Akata et al. (2023) show that "there is still significant ground to cover for LLMs to become truly social and well-aligned machines."

Finally, Fan et al. (2023) looked specifically at the rationality of LLMs. Their results indicate that "even the current state-of-the-art LLM (GPT-4) exhibits substantial disparities compared to humans in game theory." In sum, while most studies acknowledge the potential of LLMs in social science research, they consistently point out the limitations of current models. This suggests that while LLMs offer promising tools for social science research, significant gaps must be addressed or accepted and reflected upon to fully harness their capabilities in emulating human decision-making processes.

2.3.3 Automating experiments

Most of the aforementioned studies mention that their work is merely an initial exploration with the need to perform larger, more systematic simulations to establish best practices (Horton, 2023; Aher et al., 2023; Guo, 2023; Fan et al., 2024; Loré & Heydari, 2023). Moreover, Gui & Toubia (2023) demonstrate an important challenge related to LLM experimentation: distinguishing causation from correlation. Their study concludes the following:

> The LLM's ability to produce responses that capture associations present in the training data, while generally desirable, also implies that changing a variable in the prompt can unintentionally affect other unspecified variables that are supposed to stay constant. This makes it difficult to interpret whether the simulated cause-and-effect relationship is indeed driven by the treatment of interest.

Manning et al. (2024) present an approach for automatically generating and testing that also addresses the challenge demonstrated by Gui & Toubia (2023). The key feature of their new approach is the use of *structural causal models (SCMs).* With the SCM framework, data generation is based on causal structure, avoiding bad controls.

The approach of Manning et al. (2024) seems promising. It can generate controlled experimental simulations in bulk with prespecified plans for data collection and analysis. However, there are still certain limitations related to their approach. First, the problem of which attributes are inputted to the 'agents' in their system. Important characteristics such as personality traits, demographic information, and other relevant features are omitted unless explicitly incorporated into the SCM. The researchers acknowledge the potential benefits of including such attributes, though they note that optimizing this process remains a complex challenge. Moreover, their system operates based solely on input instructions, which simplifies the replication and execution of experiments but also restricts the flexibility of experimental controls.

In conclusion, while the approach by Manning et al. (2024) represents a significant advancement toward more systematic research using LLMs, it requires further refinement to enhance the selection of agent attributes and improve the flexibility of experimental controls. This will ensure that the system maintains rigorous scientific standards and adapts more effectively to the nuanced requirements of diverse experimental scenarios.

2.4 Risks and Limitations

The use of LLMs for simulating experiments is not without risks and limitations. According to Harding et al. (2023) and A. Wang et al. (2024), these limitations are so fundamental that LLMs cannot replace human research participants. In the following subsections, I will provide an overview of the risks and limitations, categorized as technical (section 3.5.1), practical (section 3.5.2), and ethical (section 3.5.3). Note that the categories serve merely as a tool to enhance reading clarity and are not mutually exclusive. For instance, technical limitations can also lead to practical and ethical risks and vice versa.

2.4.1 Technical Risks and Limitations

As mentioned in section 2.2.2, an important technical limitation comes from the data on which the LLM is trained. The data is written by a biased subgroup of the population, and thus the LLM reflects the biases of these authors (Aher et al., 2023; Dillon et al., 2023; Harding et al., 2023). In addition, for new roles or subgroups that are seldom discussed on the web, LLMs may not simulate them well (A. Wang et al., 2024). Furthermore, LLMs are generally fine-tuned to align with correct human values, but an ideal simulator should be able to honestly depict human traits, including the 'incorrect' ones (A. Wang et al., 2024).

The second technical limitation is that any given LLM can act only as a single participant (Dillion et al., 2023). The models tend to collapse the diversity of judgements into a single modal opinion. Therefore, LLMs are better at approximating the average human judgement of their training data than capturing variation. However, Horton (2023) disputes this observation. He mentions that the notion that the responses of LLMs are a sort of weighted average is incorrect.

> They are more like random number generators than estimators. If you trained an LLM on millions of people reporting random draws from U[0, 1], it would not respond with ≈ 0.5 but rather be more or less equally like to return any number in [0, 1].

There are more conflicting limitations regarding LLMs. The study of Aher et al. (2023) uncovered a "hyper-accuracy distortion," where LLMs give inhumanly accurate answers. L. Wang et al. (2024) agree with the discovery, mentioning that the knowledge of LLMs far exceeds that of average individuals. In contrast, Fan et al. (2022) mention that LLMs often suffer from decreased mathematical ability and an inability to understand preferences.

Recent studies' conflicting findings highlight LLMs' complex limitations, from their ability to produce inhumanly precise answers to struggles with mathematical reasoning and understanding preferences.

2.4.2 Practical Risks and Limitations

The first practical limitation comes from the lack of robustness in prompts for LLMs (L. Wang et al., 2024; Chen et al., 2022; Guo, 2023). Even minor changes in context and framing can yield substantially different outcomes. However, this sensitivity is not entirely negative. It allows for instilling human-like traits in LLMs through prompting (Guo, 2023).

Another practical risk is *hallucination*. Hallucination occurs when LLMs produce false information with a high level of confidence (L. Wang et al., 2024). Although hallucination is logically a large limitation, in most cases, we do not want LLMs to output false information; Horton (2023) also sees hallucination as a positive characteristic. This means that LLMs are not simply repeating something they have already read but can make up new information, which is beneficial for the variation in model output.

The last practical limitation is that LLMs may not well reflect the human psychology characters (L. Wang et al., 2024). The result is a lack of self-awareness in social scenarios. Furthermore, Manning et al. (2024) encountered the problem of engineering social interactions because LLMs are designed to exchange text in sequence. This does not reflect the natural ebb and flow of human conversation.

The practical limitations of LLMs, such as prompt sensitivity, hallucination, and challenges in modelling human psychology, highlight their dual nature. While these limitations present significant challenges, they also contribute to the models' ability to instil human-like traits and generate novel content.

2.4.3 Ethical Risks and Limitations

There is an ongoing debate around the ethics of torturing simulated agents (Darling. 2016). But, according to Aher et al. (2023), there are no Laws or Institutional Review Board policies, prohibiting mistreating simulating agents, at the time of their study. Still, creating unpleasant simulations may harm both readers and authors (Darling, 2016), and should therefore be approached with caution.

In addition, using LLMs in research may pose other risks (Charness et al., 2023). For example, (1) intellectual property (IP) concerns, (2) digital-privacy issues, (3) user deception, and (4) scientific fraud by fabricating data or strategies. Grossman et al. (2023) mention that the “black-box” nature of LLMs challenges researchers’ ability to evaluate underlying mechanisms and replicate findings.

Harding et al. (2023) make ethical a consideration about the results of a novel situation, even assuming that an LLM can output relatively accurate modal opinions. If the output is intuitively surprising, what does this say? It is strong evidence that some humans would form that judgement. Or “should we suspect that the model has given a non-humanlike response, perhaps because it’s latched onto some unconsidered aspect of the prompt, or because the vignette is out of distribution for the model?” (Harding et al., 2023). In such scenarios, the results are impossible to access without doing further research with human participants. Furthermore, LLMs offer more of a snapshot of moral opinion over some fixed past period, but moral views are constantly fluctuating due to changing world events, experiences, and social developments (Harding et al., 2023).

Finally, I want to delve deeper into the aforementioned problem of the inability of subpopulation representation. A. Wang et al. (2024) identify two main limitations here. The first limitation is that “LLMs can misportray marginalized groups as more like out-group imitations than in-group representations” (A. Wang et al., 2024). This is harmful because it upholds stereotypes, and it can involve the erasure and inscription of social hierarchies. The second limitation is that “LLMs flatten groups and portray them one-dimensionally” (A Wang et al., 2024). This is especially harmful to marginalized groups that are historically portrayed as one-dimensional.

Although the ethical risks and limitations are less relevant for the experiment reproduced in this research, the studies by Harding et al. (2023) and A. Wang et al. (2024) underscore significant ethical challenges in using LLMs, demonstrating the need for careful ongoing scrutiny of LLM applications

2.5 Knowledge Gap

The novelty of conducting experiments with LLMs, alongside the contrasting findings and perspectives revealing promises and limitations, introduces many unknowns. This thesis aims to contribute to two knowledge gaps. First, most experiments focus on relatively simple one or two-player games (Horton, 2023; Aher et al., 2023; Akata et al., 2023). More complex, three-actor experiments have not been tested as far as I am aware, and thus, it is unclear how the state-of-the-art LLMs make decisions in such a setting. Second, as previously mentioned, there is an ongoing debate about the reasoning and rationality abilities of LLMs. Combining both knowledge gaps, the following research question has been established:

To what extent are the state-of-the-art LLMs able to be 'good' participants in the Flip a Coin or Vote experiment?

Compared to the experiments from previously mentioned studies, the Flip a Coin or Vote experiment (Hoffmann & Renes, 2021) is particularly interesting due to two elements: 1) it involves three players, and 2) it incorporates dynamic strategic elements. Moreover, the experiment is conducted in two parts, each corresponding to one of these elements. The following chapter explains the experiment and emphasises its suitability for contributing to both knowledge gaps.

Subsequently, Chapter 4 outlines three levels on which to measure whether LLMs can be 'good' participants, corresponding to the metrics used by L. Wang et al. (2024) to evaluate the effectiveness of AI agents. These levels are (a) understanding of the experiment, (b) the rationality of choices, and (c) the human-likeness of the choices. Exploring these aspects will shed light on the potential and limitations of LLMs in more complex experiments and their suitability for use in the institutional design of socio-technical systems.

3 The Flip a Coin or Vote Experiment

This chapter summarizes the Flip a Coin or Vote Experiment conducted by Hoffmann & Renes (2021), which will be reproduced in this thesis with GPT-3.5 and GPT-4o. In the following sections, I will superficially cover the background of the experiment (section 3.1), the experimental design (section 3.2), and the results of the experiment (section 3.3). The objective is to provide the reader with an understanding of the experiment. For a more complex, in-depth explanation of the experiment, the rationale behind the experiment, and the complete interpretation of the results, I refer to the study of Hoffmann & Renes (2021).

The final section (section 3.4) elaborates on the suitability of the experiment in light of recent scientific studies on simulations with LLMs. The two main elements that make the Flip a Coin or Vote experiment suitable are: 1) it involves three players, and 2) it incorporates dynamic strategic elements. Throughout this chapter, I consistently make use of the study of Hoffmann & Renes (2021). Therefore, I do not explicitly refer to the study in the sections in order to avoid cycle referring.

3.1 Introduction

The first thing a group has to do before making a decision is to select a decision rule. The decision rule should aggregate individual preferences into a group decision. It is logical to expect that inefficient rules will be replaced by efficient ones. However, in practice, inefficient decision rules regularly persist. The experiment is used to gain an understanding of when inefficient decision rules can be expected to (not) be replaced by efficient ones in two ways. First, it measures subjects' willingness to participate in several mechanisms. Second, it measures and compares the empirical efficiency of the mechanisms.

3.2 Experimental Design

3.2.1 The Game

In the experiment, subjects interact in groups of three, and each group faces the question of whether or not to implement a public project. There are 18 experimental rounds, each consisting of two parts. First, a mechanism is selected for each group. Second, the group decides on the implementation through the chosen mechanism. If the project is implemented, each player receives a payoff equal to their private valuation. If the project is not implemented, each player receives a payoff of zero.

At the beginning of the round, subjects are informed about two available mechanisms. Each subject selects one of the two mechanisms. Thereafter, the computer randomly picks one group member as the dictator and the mechanism chosen by this dictator is executed for the group. All subjects were informed of the selected mechanism before they played it, but they did not learn whose choice it was or what the other two subjects selected. In the end, after the mechanism is played, subjects are informed about the outcome and their payoffs.

The experiment proceeds in two stages. In the first stage, the first 12 rounds, subjects learn their private valuation for the project after choosing their preferred mechanism, but before the mechanism is played, the 'ex-ante' rounds. In the second stage, rounds 13 to 18, subjects are informed about their private valuation before choosing their preferred mechanism, the 'ad-interim' rounds. The experiment is conducted in German, but the full translated instructions of the experiment can be found in Appendix B.

3.2.2 The four Mechanisms

The four mechanisms in the experiment are:

Mechanism I AGV mechanism (AGV)

All group members report a valuation for the project. They can only report valuations that are present in the type space. If the sum of reported valuations is larger than zero, the project is implemented. If the sum is smaller than zero, the project is not implemented. Independent of project implementation, subjects pay or receive a transfer that depends on the vector of reported valuations.

Mechanism II Voting – Simple Majority (SM)

All group members vote for or against the project (no abstention). If two or more group members vote for implementation the project is implemented, otherwise the project is not implemented

Mechanism III Non-implementation Status Quo (NSQ)

The public project is not implemented.

Mechanism IV Random implementation (RAND)

Whether the public project is implemented depends on the flip of a fair coin. The project is implemented with 50% probability independent of subjects' valuations

To ensure truthful reports in the AGV mechanism, the mechanism calls for transfers equal to the expected externality an individual generates for the others.

3.2.3 Treatments

Treatments only differ in the distribution of private valuations for the project. All treatments consist over a type space with four possible valuations (in €), and personal valuations are drawn independently each round from a uniform distribution over the possible values. The treatment, and therefore the possible valuations are common knowledge and therefore known before starting part 1. Table 1 presents an overview of the treatments and the number of observations.

Table 2. Distribution of valuations and number of observations (Hoffmann & Renes, 2021, Table 1)

Treatment	**Valuations (€)**				**Number of**	
					Subjects	Match groups
Symmetric	-3	-1	1	3	45	4
Right skewed (+7)	-3	-1	1	7	42	4
Left skewed (-7)	-7	-1	1	3	45	5
Robustness	-3	-2	-1	7	18	2
Probability	25%	25%	25%	25%	2700	

3.3 Results

The experimental patterns observed in mechanism choices were largely, but not completely, consistent with the narrowly self-interested rationality. Contrary to theoretical expectations, not all subjects favoured the AGV mechanism over flipping a coin, and clear majorities for either AGV or SM were often absent. Both the optimal AGV and SM do not achieve the same level of efficiency in practical settings as they do in theoretical predictions, indicating that theoretical predictions about individual participation preferences do not always hold true in empirical tests.

3.4 Conclusion

Most studies that have conducted simulations with LLMs have focused on relatively simple experiments involving one or two players (e.g., Horton, 2023; Aher et al., 2022; Akata et al., 2023; Guo, 2023). Consequently, more complex experiments with three players remain unexplored. In addition, Akata et al. (2023) demonstrated that while LLMs generally underperform in games requiring coordination, they perform well in games where valuing self-interest pays off.

The Flip a Coin or Vote experiment requires both coordination and the valuation of self-interest due to its two-part structure. In the first part of the experiment, players must make decisions based on limited information about their own and the other players' potential valuations. The decisions made in part one determine whether part two will require coordination among the players. When SM is the group's mechanism, the application is rather straightforward. However, when AGV is the mechanism, players can make strategic decisions to influence the group outcome. The NSQ and RAND mechanisms do not require any coordination.

The inclusion of three-player groups, dynamic strategic elements, a two-part structure with additional information for the players, and both straightforward and complex decision mechanisms make the Flip a Coin or Vote experiment a suitable choice. The reproduction of the experiment with GPT-3.5 and GPT-4o will contribute to addressing the knowledge gaps identified in Chapter 2.

4 Research Approach

This chapter outlines the research methodology to address the main research question (MRQ). Based on the formulation of the MRQ, a quantitative experimental approach was adopted (Section 4.1). The MRQ is further broken down into five sub-questions (SQs), establishing three metrics for defining a ‘good’ participant (Section 4.2). The various experimental conditions are detailed and connected to the different simulation runs (Section 4.3). Section 4.4 reflects on a different approach that could have been used for this study. Finally, section 4.5 explains the statistical tests that will be used to compare the different runs with the lab results.

4.1 Quantitative Experimental Approach

To address the MRQ, extensive statistical testing is necessary to determine if the LLMs can serve as ‘good’ participants in behavioural economics simulation experiments. Consequently, this thesis employs a quantitative research approach. According to Creswell (2009), quantitative research is "a means for testing objective theories by examining the relationship among variables." In this context, it involves assessing to what extent LLMs can be ‘good’ participants under varying experimental controls.

Quantitative research encompasses various strategies, including experimental research, which facilitates the measurement, quantification of phenomena, and hypothesis testing (Creswell, 2009). Seltman (2018) describes the scientific learning process as beginning with the construction of a testable hypothesis, followed by the design and execution of the experiment. In this thesis, hypotheses regarding different experimental controls are formulated in Section 4.3, the conceptualisation of the experiment is detailed in Chapter 5, and the results of the experiment are presented in Chapters 6 and 7, providing a clear and structured research process.

However, quantitative research has inherent limitations. Creswell (2009) notes that it may overlook nuanced, contextual, and subjective aspects of a subject. Additionally, quantitative research tends to focus more on the ‘what’ and ‘how much’ rather than the ‘why’ (Patton, 2014). These limitations must be carefully considered in this thesis. To counteract these limitations, this thesis includes prompts for the LLM to provide reasoning behind each decision, a task often challenging in human studies (Guo, 2023). By incorporating reasoning outputs, it becomes possible to briefly analyse the ‘why’ as well.

4.2 Sub-questions

In this section, the MRQ will be divided into five sub-questions. Section 4.2.1 defines three metrics to evaluate when an LLM can be considered a ‘good’ participant. Section 4.2.2 defines two additional sub-questions related to the various experimental controls and their impact on whether the LLM is a ‘good’ participant. Table 3 presents an overview of the SQs.

Table 3. Overview of the different sub-questions

N^0	Sub-question
1	Does the LLM understand the rules of the experiment?
2	Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?
3	Are the choices of the LLM ‘human-like’ in both parts of the experiment?
4	How much variation is there in the choices of the LLM between the different runs?
5	What effects have the experimental controls had on the choices of the LLM?

4.2.1 The Three Levels of a ‘good’ Participant

First and foremost, the LLM must produce human-like responses to be considered a ‘good’ participant in economic and behavioural experiments (Dillion et al., 2023). Various studies define "human-like" at different levels. Horton (2023) uses a relatively superficial definition, stating that an LLM is human-like if “it can qualitatively recover findings from experiments with actual humans.” Chen et al. (2023) and Fan et al. (2024) delve deeper, examining the rationality of LLMs, each using a different definition of rationality. Loré and Heydari (2023) specifically investigate the strategic behaviour of LLMs.

There is no clear, standardised level or definition for measuring whether results are human-like. Therefore, I will assess whether an LLM is a ‘good’ participant on three levels, aligning them with the metrics used by L. Wang et al. (2024) to evaluate the effectiveness of AI agents. Table 4 provides an overview of the three levels and their corresponding evaluation methods.

Sub question 1: Does the LLM understand the rules of the experiment?

The first level corresponds to the task success metric defined by L. Wang et al. (2024) and is described as “how well an agent can complete tasks and achieve goals.” In this context, it refers to how well the LLM understands the rules of the experiment and how effectively it can participate in the experiment.
Measuring an LLM's understanding of instructions can be challenging. Chen et al. (2023) addressed this by asking three questions to verify the LLM's comprehension of the instructions, simulating the process 25

times. In their study, the LLM consistently provided correct answers, indicating it understood the decision environment..

While Chen et al.'s (2023) approach is a good starting point, my exploratory research with the OpenAI API reveals that the situation is more nuanced. Each request to the LLM is treated as a 'new' participant encountering the instructions for the first time. Thus, in an experiment with 100 replications, it is possible that the LLM correctly understands the instructions 95 times. Therefore, the question of whether the LLM understands the instructions cannot be answered in a binary manner (yes or no) based on a set of questions replicated only 25 times.

In addition, using these questions would result in a higher number of tokens being consumed, leading to increased costs for running the experiment. Therefore, I will employ an alternative approach to assess whether the LLM understands the rules of the experiment. This approach involves analysing the qualitative explanations provided by the LLM for its decisions. In evaluating these explanations, I will focus on the following criteria:

- **Logical coherence:** Do the explanations logically make sense?
- **Contextual accuracy:** Is the LLM hallucinating context? (L. Wang et al., 2024; Horton, 2023)
- **Response validity:** Are the LLM's responses valid and in the correct format?

The logical coherence is important to analyse whether the explanations are consistent given the rules of the experiment. If the LLM's explanations for its decisions align logically with the rules and objectives of the experiment, it suggests that the LLM has a clear understanding of the experiment. Contextual accuracy focuses on hallucination, a limitation mentioned by L. Wang et al. (2024) and Horton (2023). Analysing hallucinations ensures that the LLM is interpreting and responding to the given instructions correctly without adding irrelevant or fabricated details that are not part of the experiment. Response validity checks if the LLM is providing answers that conform to the expected format and choices defined by the experiment

These criteria ensure that the LLM is not only capable of generating answers but also that it indicates to understands the rules, context, and format of the experiment. The limitation of this approach are that it is only a qualitative way of measuring, based on my biased interpretation of the LLM's explanations.

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

The second level is related to the *efficiency metric* of L. Wang et al. (2024). Here, I follow definition, used in game theory, of a rational player: "A player is said to be rational if he seeks to play in a manner which maximizes his own payoff. It is often assumed that the rationality of all players is common knowledge" (Turocy & Von Stengel, 2003). In this research, this is conceptualised as follows.

First, I examine whether the LLM chooses the theoretically efficient mechanism given the binary option of two mechanisms in the first part of the experiment. The theoretically efficient mechanism is determined based on the Nash equilibrium, also called strategic equilibrium – "a list of strategies, one for each player, which has the property that no player can unilaterally change his strategy and get a better payoff" (Torucy & Von Stengel, 2003). Over a complete run[4], this gives a percentage of 'rational' choices made. Note that this percentage is individual-based and does not indicate the optimal choice for the group as a whole.

Second, I assess how the LLM applies the AGV and SM mechanisms in the second part of the experiment. For AGV, I measure the percentage of truth-telling and truth-telling about its sign. Lying in general and lying about its sign especially indicate an irrational decision. However, lying about its valuation can also indicate strategic behaviour if the LLM expects that the other participants are going to lie as well. In the SM mechanism, the rational choices are rather straightforward; to vote "Yes" with a positive valuation and "No" with a negative valuation.

Laboratory results show that human participants do not achieve 100% on these rationality metrics. Therefore, if both metrics approximate the results of the lab data, I consider the LLMs rational enough to be 'good' participants. Additionally, if the LLM is able to make rational choices, it can be assumed that the LLM understands the instructions of the experiment. Thus, the metrics for SQ 2 can also be used to assess SQ 1.

Sub question 3: Are the choices of the LLM 'human-like' in both parts of the experiment?

The third level corresponds to the *human similarity metric* outlined by L. Wang et al. (2024). Here, I look at whether the LLM's choices resemble those of humans. Importantly, this sometimes requires the LLM to refrain from making the theoretically optimal choice. If the optimal option is always chosen, the results may become unrealistically accurate, a phenomenon often referred to as 'hyper-accuracy distortion,' which has been observed in larger models (Aher et al., 2023).

---

[4] With run, I refer to a whole reproduction of the experiment. Simulating all 2700 rows.

The metric I will use involves calculating the absolute and squared differences between the binary choices for the four mechanisms. For example, in the laboratory results, 82% of participants choose AGV over NSQ with the right-skewed treatment in the ex-ante rounds. If GPT-3.5 chooses AGV only 70% of the time, the absolute mean difference would be 12%, and the squared mean difference would be 144. With six binary choices and four different treatments, the metric will encompass 24 comparisons, resulting in an average absolute and squared mean difference, as well as a sum of the absolute and squared mean differences. The metrics allow me to compare the different simulations – 'runs' – with the lab data. However, there must also be a way to determine when the LLM results are similar enough. Therefore, I will statistically test all runs with the lab data and all runs with RUN-0. The description of the statistical analysis, can be found in section 4.5.

Table 4. Overview of 3 levels of 'good' participants and corresponding evaluation methods

Metric	Evaluation method
SQ1: Task success metric	- **Logical coherence:** Do the explanations logically make sense? - **Contextual accuracy:** Is the LLM hallucinating context? - **Response validity:** Are the LLM's responses valid and in the correct format?
SQ2: Efficiency / rationality metric	- Percentage of choosing the theoretically optimal mechanism based on Nash equilibrium - Percentage of lying about its valuation - Percentage of lying about its sign - Percentage of rational voting
SQ3: Human similarity metric	- Absolute differences between lab data and simulated data - Squared differences between lab data and simulated data

4.2.2 Running the Experiment

After establishing the metrics to evaluate whether LLMs are 'good' participants, the next step is to analyse the different runs. A "run" refers to a complete simulation of the experiment. Each run will involve different experimental controls, such as various options for steering LLM behaviour, including different framings and model settings. However, the core elements of the experiment—such as the rules, mechanisms, treatments, randomly generated numbers, and personal valuations—will remain consistent across all runs. The following two sub-questions (SQs) relate to the execution of these experimental runs.

Sub question 4: How much variation is there in responses between the different runs?

As argued by Dillion et al. (2023), an LLM is better at approximating the average than they are at capturing variation. This would imply that between different runs the LLM would often make the same choice, given the same treatment. Horton (2023) disagrees with Dillion et al. (2023) and argues that LLMs are more random number generators than estimators, resulting in more variation. With SQ4, I aim to research both assumptions. However, due to the binary choice option, no strong claims can be made about the variation in answers, only in the variation between the two options.

Sub Question 5: What effects have the experimental controls had on the choices of the LLM?

As argued by Dillon et al. (2023), an LLM is better at approximating the average than capturing variation. This suggests that, across different runs, the LLM would often make the same choice when given the same treatment. However, Horton (2023) disagrees with Dillon et al. (2023) and argues that LLMs behave more like random number generators than estimators, resulting in greater variation. With SQ4, I aim to explore both assumptions. However, due to the binary choice nature of the experiment, any conclusions about variation will be limited to differences between the two options.

4.3 Experimental Controls

To deepen the understanding of GPT's choices and how to influence them, I introduce variations in the different experimental controls. In practice, the design space for experimental controls in an LLM is virtually infinite. A whole area of research is dedicated to optimising prompts for LLMs (White et al., 2023). For example, it is possible to vary prompt instructions, languages, reasoning steps, persona allocations, and model temperature. To select relevant and testable experimental controls, I drew upon best practices identified in the literature review from Chapter 2.

In total, I selected five different experimental controls, resulting in 14 different runs. Table 5 provides an overview of all the runs. For each run, a specific experimental control will be tested, while all others will be set to their default levels—the levels from RUN-0. In the following subsections, I will explain the different experimental controls and provide details on individual runs where necessary. Running simulations with GPT-4o is significantly more expensive. Therefore, I selected only a subset of four runs to simulate with GPT-4o.

Table 5. Overview of the experimental controls and their corresponding runs

RUN	Description	GPT-3.5	GPT-4o
0	Default run, based on the prompt instructions of the experiment from Hoffmann & Renes (2021)	✓	✓
Experimental Control 1: Context Control			
1	Prompt instructions adapted with GPT-3.5 (LLM-Generation method L. Wang et al., 2024)	✓	✓
2	Prompt Instructions adapted with GPT-4 (LLM-Generation method L. Wang et al., 2024)	✓	
Experimental Control 2: Step by Step reasoning			
3	Adding 'Let's think step by step before answering" (Kojima et al., 2023)	✓	
4	Adding 'Let's think step by step before answering" followed by manual suggested thinking steps	✓	
Experimental Control 3: Trait or Role Allocation			
5	Add the role: Try to make human-like decisions	✓	✓
6	Add the role: Try to make self-interested human-like decisions (Horton, 2023)	✓	
7	Add the role: Try to make inequity aversion human-like decisions (Horton, 2023)	✓	
Experimental Control 4: Persona Allocation			
8	Add the Gender and Age as input per participant	✓	
9	Add the political view as input per participant, on a scale from left (1) to right (11)	✓	
10	Add the risk willingness as input per participant, on a scale from risk-averse (1) to risk-seeking (10)	✓	
Experimental Control 5: Temperature			
11	Change the model temperature to 0	✓	
12 13	Change the model temperature to 0.5 Change the model temperature to 1.5	✓	✓

4.3.1 Contextual Framing – Context Control

The first experimental control I will test is what studies often refer to as contextual framing (Loré & Heydari, 2023; Guo, 2023) or context control in prompt engineering (White et al., 2023). With contextual framing, the researcher makes changes in the context, for example, by choosing different words in the instructions prompt.

I will conduct two runs with different contextual framings. In the first, the prompt instructions—the explanation of the experiment—will be adapted by GPT-3.5. In the second, the prompt instructions will be adapted by GPT-4. This approach is referred to as the LLM-Generation method, in contrast to the Handcrafting method, where researchers manually modify the prompt (L. Wang et al., 2024). I chose the LLM-Generation method to enhance the reproducibility of my research, with minimal manual input.

My hypothesis is that altering the prompt instructions will lead to significant differences, as identified in multiple studies (L. Wang et al., 2024; Chen et al., 2022; Guo, 2023; Loré & Heydari, 2023). Furthermore, I expect that the adapted prompt instructions will result in more rational choices based on the assumption that GPT is better at crafting instructions that are easier for an LLM to understand.

RUN-1: Prompt instructions adapted with GPT-3.5

The first run will have different prompt instructions, adapted with GPT-3.5. The prompt I used to instruct GPT-3.5 is the following:

> I have done an experiment with participants. Now I want to do the same experiment with an LLM. Can you rewrite the following instructions so that they are more clear for an LLM to understand?

The original and the rewritten prompt instructions can be found in Appendix B. GPT-3.5 also added reasoning on why its instructions were easier for an LLM to understand:

> The instructions I provided are more concise and direct, making it easier for an LLM to grasp the essential information quickly. By breaking down the instructions into shorter sentences and focusing on key points, such as the decision-making process and payment structure, the LLM can better understand the task at hand without being overwhelmed by unnecessary details.

RUN-2: Prompt instructions adapted with GPT-4

The second run follows the same concept as the first but with GPT-4 instead of GPT-3.5. The input instructions were identical, but GPT-4 provided the following reasoning:

> Ensure that your instructions are clear and that the decision-making process, the impact of valuations, and the sequence of the experiment phases are all explicitly stated.

4.3.2 Step-by-Step Reasoning

The second experimental control I will test also relates to the prompt instructions but focuses on specifying how the LLM should 'reason' about its decisions. There are various approaches to this, such as chain of thought prompting (CoT), which involves providing LLMs with step-by-step reasoning examples rather than a standard question (Wei et al., 2022). Building on this approach, Kojima et al. (2023) introduced the Zero-shot-CoT reasoning. Their study demonstrates that simply adding the phrase "Let's think step by step" before each answer significantly enhances the LLM's performance, suggesting that the capabilities of LLMs can be improved with simple prompts. I will test this approach with two different runs.

Similar to RUN-1 and RUN-2, I expect the results to differ from RUN-0 due to GPT's sensitivity. In addition, following the results of Kojima et al. (2023), I expect the results to be more rational. However, I am sceptical about the degree of impact.

RUN-3: Step-by-Step Reasoning part 1

In the third run I will simply follow the approach from Kojima et al. (2023). At the end of the prompt instructions I will add “Let’s think step by step”.

RUN-4: Step-by-Step Reasoning part 2

In the fourth run, I will use the same approach as in the 3rd run but with additional steps for the LLM to consider. Instead of simply adding “Let’s think step by step,” I will include the following:

> Let’s think step by step: First, reflect on the possible valuations for you and your group members. Second, reflect on the implications of both decision mechanisms on these possible valuations. Third, based on these two reflections, select the best mechanism that optimises your expected payoff.

4.3.3 Trait or Role Allocation

After varying the prompt instructions and reasoning instructions, the next experimental control involves trait or role allocation. By trait or role allocation, I refer to assigning a specific role or trait to the LLM, which remains consistent across all requests within a given run. For example, Horton (2023) conducted simulations where the LLM was instructed to exhibit traits such as inequity aversion or self-interest. The OpenAI API includes a specific function for role allocation at the start of each request, allowing the programmer to define the desired behaviour of the model. The default role in the OpenAI API is “You are a helpful assistant” (OpenAI, n.d.). I will conduct three runs, each with a different role allocation.

Similar to the study of Horton (2023), I expect that inputting a different role or trait, will result in different results. However, I am sceptical about whether the different results correlate to the specific roles or are related to a change in the input prompt..

RUN-5: Role Allocation – Try to give human-like responses

In the fifth run, the LLM will be assigned the role “Try to give human-like responses.” The rationale is that this prompt will encourage the LLM to reason in a more human-like manner. However, while the results may vary, I do not anticipate a significant increase in human-like responses, given the LLM's limitations in reasoning like a human on command.

RUN-6: Role Allocation – Try to give self-interested human-like responses

Similar to Horton (2023), I will test whether personality traits will impact the results. According to the used definition of rationality, self-interested command would, in theory, result in more rational choices. However, similar to RUN-5, I am sceptical about the effect.

RUN-7: Role Allocation – Try to give inequity aversion human-like responses

In contrast to RUN-6, I will test whether there is difference between the self-interested prompt and the inequity aversion prompt. For this run I expect the LLM to give responses that are more aligned with the overall group compared.

4.3.4 Persona Allocation

The fourth experimental control is persona allocation, which involves assigning the LLM a specific persona to adopt when generating output (White et al., 2023). The difference between this and trait or role allocation is that, instead of one general role or trait, each request will have its own unique persona. This approach is similar to that used by Aher et al. (2023), where they assigned personas based on title, racial background, and surnames, creating a pool of 1,000 unique participants. In this study, I will use gender, age, political orientation, and willingness to take risks to create personas. The values will be extracted from the data of Hoffmann & Renes (2021) to generate personas that match those of the participants in the lab experiment.

My hypothesis is that the LLM will produce different results for each persona allocation. However, I expect this variation to be primarily due to the LLM's sensitivity to context. I am somewhat sceptical about whether the LLM will make decisions that align with its assigned persona. This can be most easily tested in the 10th run, given the implications of risk willingness.

RUN-8: Persona Allocation – Age and Gender

The eighth run will input the Age and Gender of respondents to the LLM.

RUN-9: Persona Allocation – Political orientation

The nineth run will vary with political orientations, based on a single scale from 1 to 11, with 1 being most left-oriented, and 11 being most right-oriented. This is similar to the approach of Horton (2023), however instead of socialist versus liberalist the scale is numerical from most left oriented to most right oriented.

RUN-10: Persona Allocation – Risk willingness

The tenth run will vary with risk willingness, based on a scale from 0 to 10, with 0 being very risk-averse and 10 being very risk-seeking. This is similar to approach of Horton (2023), however instead of self-interested versus inequity aversion, the scale is risk willingness from 0 to 10.

4.3.5 Temperature

After varying the input to the model, the next step is to adjust the model settings. One important setting is called temperature. Temperature controls the level of stochasticity and creativity in the responses generated by an LLM (Chen et al., 2023). The temperature ranges from 0 to 2, with higher values indicating more randomness. The default temperature for OpenAI models is set to 1 (OpenAI, n.d.). According to Loré & Heydari (2023), higher temperatures are commonly used in similar studies.

It's important to note that the temperature range in older versions of the API was from 0 to 1. This can lead to confusion when comparing research conducted with the older API version. Even the OpenAI API reference website states: “What sampling temperature to use, between 0 and 2. Higher values like 0.8 will make the output more random, while lower values like 0.2 will make it more focused and deterministic” (OpenAI, n.d.-a). This explanation appears to be based on the older range of 0 to 1, as with the default value now set at 1, 0.8 cannot be considered a higher value. Furthermore, the API used for audio still uses the range from 0 to 1 with the default option as 0. In this study, I will set the temperature to 0, 0.5, and 1.5[5].

My prediction is that decreasing the temperature will result in stronger preferences for the given mechanisms in part 1 of the experiment and will result in more homogeneity in explanations. Furthermore, I expect that increasing the temperature will result in a more equal distribution between the decision mechanisms. Thus, in other words, preferences for specific mechanisms will be decreased. In addition, Chen et al. (2023) noticed that the number of invalid responses significantly increases with higher temperatures, so this can also be expected. Furthermore, increasing the temperature might be a solution against hyper-accuracy distortion. However, Chen et al. (2023) found that increasing the temperature does not affect the level of rationality.

RUN-11: Temperature = 0
RUN-12: Temperature = 0.5
RUN-13: Temperature = 1.5

4.4 Structural Causal Model

Another approach for experimental research with LLMs is the Structural Causal Model (SCM) approach introduced by Manning et al. (2024). However, this method is largely automated, requiring only a brief description of the scenario as input. While there are some opportunities to influence certain stages of the process, the approach offers limited customisation for the researcher. This lack of flexibility makes it challenging to fully replicate the experiment conducted by Hoffmann & Renes (2021) using Manning et al.'s (2024) approach. Additionally, the SCM approach has difficulties handling the 'attributes' of agents, which are crucial for this experiment. For these reasons, the SCM approach is not tested in this study. Nonetheless, I acknowledge the potential of the SCM approach.

[5] I also tried runs with extremely high temperatures (2.0 & 1.9). But, the runs were extremely slow, over 20 seconds per request and kept crashing continuously. Furthermore, results from RUN-13, with the temperature set to 1.5, already resulted in many invalid answers.

4.5 Statistical Testing of the Runs

Each run will be tested on statistical differences between decisions in the binary options, absolute and squared, with the lab results, and with RUN-0, which is used as a benchmark. The full statistical analysis can be found in appendix E.

The differences from the lab results will be tested using a One-Sample T-Test, with the test value set to 0. This is because the differences between the lab results and themselves are, by definition, 0.

$$H0\colon \mu_{Lab} - \mu_{RUN-X} = 0$$

$$H1\colon \mu_{Lab} - \mu_{RUN-X} \neq 0$$

The differences between RUN-0 and the other runs will be tested with the Paired-Samples T-Test.

$$H0\colon \mu_{RUN-0} = \mu_{RUN-X}$$

$$H1\colon \mu_{RUN-0} \neq \mu_{RUN-X}$$

Finally, the differences between the runs that correspond to the same experimental control will be tested with either the Paired-Samples T-Test or the One-Way ANOVA, based on the number of runs that correspond to the same experimental control.

$$H0\colon \mu_{RUN-ECX\_1} = \mu_{RUN-ECX\_2}$$

$$H1\colon \mu_{RUN-ECX\_1} \neq \mu_{RUN-ECX\_2}$$

$$OR$$

$$H0\colon \mu_{RUN-ECX\_1} = \mu_{RUN-ECX\_2} = \mu_{RUN-ECX\_3}$$

$$H1\colon at\ least\ one\ mean\ is\ different\ from\ the\ others$$

5 Conceptualization of the Experiment

This chapter outlines the conceptualisation of the lab experiment into a simulation model. Section 5.1 provides a conceptual overview of the system. Following this, Section 5.2 details the implementation of the system in Python and the flow of the model. The final section of this chapter (5.3) discusses the key differences between the operationalised model and the original lab experiment, highlighting the implications of these differences.

5.1 System Diagram

Figure 1 provides a conceptual overview of the system. Elements depicted in black correspond to the lab experiment by Hoffmann & Renes (2021), while those shown in blue pertain to the simulation model and represent the scope of this study. The solid lines indicate the flow of processes, whereas the dotted lines represent data connections to specific aspects of the study, such as input and output data, experimental controls, and participant characteristics. An overview of all variables from the experiment conducted by Hoffmann & Renes (2021) and the variables that are used and updated in this study is presented in Appendix C.

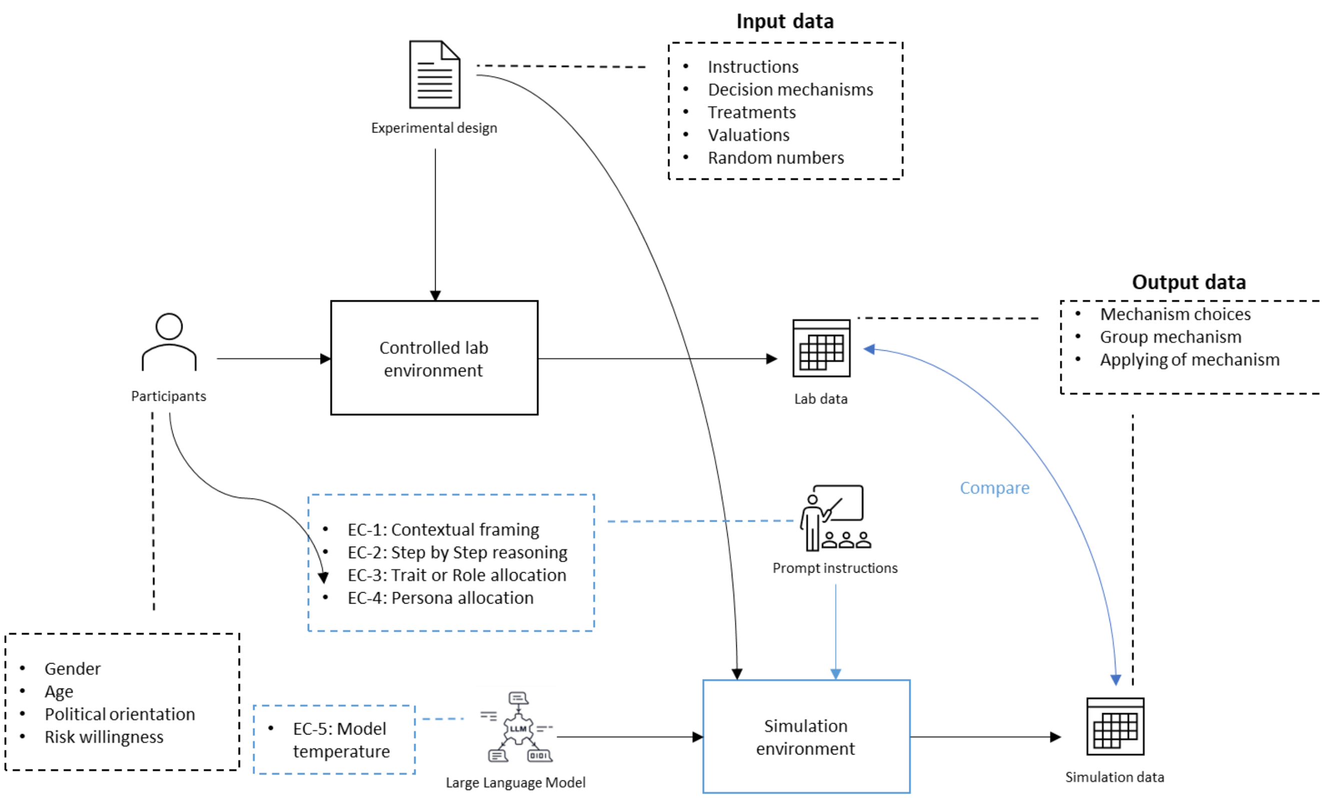


Figure 1: System diagram

In the controlled lab environment, groups of human participants engaged in the experiment. They were given instructions, introduced to the four decision mechanisms and treatments, given individual valuations for the project, and supplied with random numbers used for selecting the group's final mechanism. The output data included each participant's preferred mechanism and the mechanism ultimately selected by the group. When the group chose either AGV or SM as the mechanism, participants had to apply it. Finally, individual payoffs and total group payoffs were calculated.

In the simulated environment, human participants were replaced with GPT-3.5 or GPT-4o. The input data from the lab environment, with the first four experimental controls, were used to formulate the prompt instructions provided to the LLM. The simulation data mirrors the structure of the lab data and will be compared with it.

5.2 Model Logic Flow

The system's operationalisation involves six steps, as illustrated in Figure 2. Appendix C details each variable used or updated, specifying the step in which it occurs.

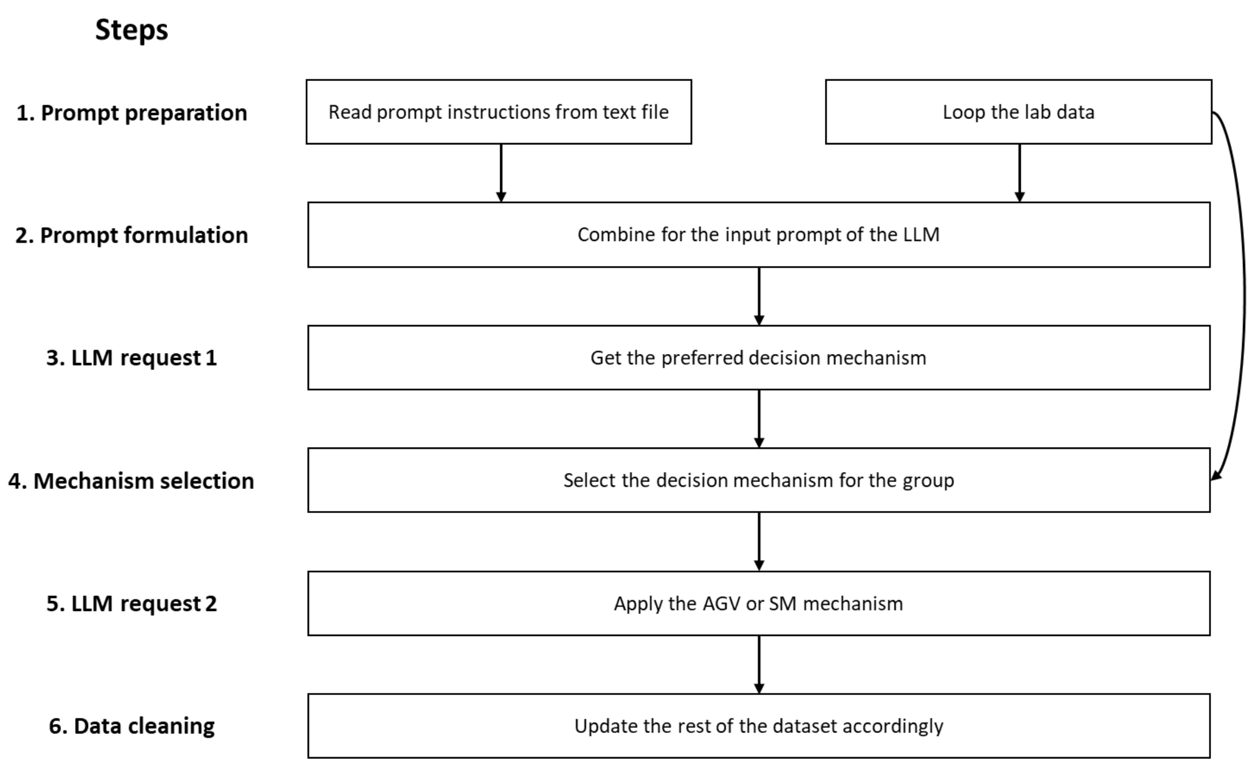


Figure 2. Model logic flow with corresponding steps

Step 1 is to prepare the input prompt, which consists of two parts. The first part contains the general instructions for the experiment, which remain consistent across all input prompts. These instructions are saved as text files, allowing easy updates without code modification. The second part includes specific conditions, such as binary choices between decision mechanisms, treatment, and valuation for each row. This ensures that conditions are identical for both the simulation and lab results. Step 2 combines these two parts to create a unique input prompt for each row.

Step 3 involves the first LLM request, which iterates over the lab data using the unique prompts generated in step 2. Unlike the lab experiment, only the two binary mechanisms are provided to limit token usage and reduce costs. In addition, providing all four mechanisms could lead the LLM to select an incorrect option that is not part of the binary choices for that specific round. The prompts are inputted through the OpenAI Chat Completions API, producing a dataset similar to the lab dataset but featuring new mechanism decisions and LLM explanations.

In step 4, the group mechanism is selected using the same random numbers as in the lab. If the group mechanism is either AGV or SM, a second request is made to apply the mechanism. This request uses the same prompt as the first but includes the selected group mechanism and the private valuation for each row. The results include stated valuations for AGV, votes for SM, and explanations. In the final step, the dataset is updated, meaning that all variables required for further analysis are updated, for example, to check whether the LLM selected the theoretically optimal decision mechanism in the first request or misrepresented its valuation in the second request.

5.3 Differences Between Experiment and Simulation

It is essential to highlight the differences between the experiment and the simulation model. These differences arise primarily because the OpenAI API calculates costs based on the number of input and output tokens. Therefore, the study has been carefully designed to minimize token usage, ensuring cost-effectiveness. The main differences are as follows:

First, the experiment instructions for the simulation model are shortened. This is done to limit the number of input tokens used. Although the prompt instructions are still based entirely on the original experimental instructions, redundant information has been excluded. An overview of the instructions for the experiment and the shortened instructions inputted to the simulation model is presented in Appendix B.

Second, instead of explaining all four decision mechanisms and then offering a choice between two, the simulation model only explains the two mechanisms relevant to each row. In earlier test versions, explaining all four mechanisms sometimes led the LLM to favour a mechanism not available as an option for that specific row.

Third, in the lab experiment, participants were not asked to rationalise their decisions. In the simulation, however, the LLM is prompted to always explain its decisions. This difference makes it challenging to determine the reasoning behind participants' choices in the lab, meaning that only the LLM's decisions can be compared, not its explanations.

Fourth, and most significantly, in the lab, participants played 18 rounds of the game, receiving feedback after each round. In the simulation model, however, each request is treated as a 'new' one, with no carryover of learning from previous rounds. In the lab, participants could learn from earlier rounds, but this is not the case in the simulation model. There are two main reasons for this: first, to control costs, and second, because the OpenAI Chat Completions API treats each request as independent, with no memory of previous interactions. While it is theoretically possible to input data from previous rounds into each new request, I assume that this would not enable the LLM to 'learn' from earlier rounds. If this assumption is incorrect, it represents a significant limitation of the study.

6 Results

This chapter presents the results derived from the experimental runs conducted with GPT-3.5 and GPT-4o. The chapter is structured as follows: Section 6.1 details the methodology employed for analysing all runs. Sections 6.2 and 6.4 provide an overview of the results for GPT-3.5 and GPT-4o, respectively, followed by an interpretation of these results in Sections 6.3 and 6.5. Finally, Section 6.6 highlights the key differences observed between the runs with the two models.

6.1 Methodology

This section outlines the systematic approach to analyse the fourteen runs with GPT-3.5 and the four runs with GPT-4o. The methodology consists of two main steps: first, an individual analysis of each run, and second, a comparative analysis across runs with statistical testing. Below, I will briefly explain each steps. The rest of the chapter consists of the conclusions drawn from these steps.

6.1.1 Step 1 – Individual Run Analysis

Each run is examined systematically, with analyses presented in Appendix E for GPT-3.5 and Appendix F for GPT-4o. The analysis begins with an overview of the experimental settings, including the date, time, and duration of each run, as well as the values of all experimental controls. Following this, a summary of each run is provided, highlighting key metrics such as the absolute mean differences from lab data in the distribution of mechanisms during part 1, the percentage of rational choices made during ad-interim rounds, and relevant metrics for the application of the SM and AGV mechanisms in part 2.

Next, the results of part 1 are examined in greater detail. The distribution of choices across the mechanisms is presented for each treatment during the ex-ante and ad-interim rounds with a positive or a negative valuation. All noteworthy deviations, either small or large, with the lab results are highlighted with examples of explanations given by the LLM. The explanations of the LLM are emphasised with the `Fira Code` letter type.

Following the results of part 1, the metrics for the application of the SM and AGV mechanisms in part 2 are presented and compared with the lab results. For the AGV mechanism, this comparison includes the percentage of truth-telling and the accuracy of truth-telling concerning its sign. For the SM mechanism, we focus on the percentage of "Yes" votes when participants had a positive valuation and the percentage of "No" votes when they had a negative valuation. Any notable observations and explanations are also provided.

Finally, the metrics for evaluating the LLM's performance across three levels of a 'good' participant are examined. These criteria include: a) whether the LLM understands the rules of the experiment, b) whether the LLM can make rational choices, and c) whether the LLM's decisions are human-like. These steps ensure a thorough and systematic analysis of each run.

6.1.2 Step 2 – Comparison Between Runs

After the individual analyses, the runs are compared to gain a comprehensive overview of the results for both GPT-3.5 and GPT-4o, with details provided in Appendices G and H, respectively.

The comparison begins with an overview of all runs, highlighting key differences from the lab results. This is followed by a visualisation of the distribution of choices made during part 1 of the experiment across all treatments, with examples of explanations provided for each run. Next, the percentages of rational choices made during the ad-interim rounds are compared, emphasising the most promising and least successful outcomes. The final comparison presents a comprehensive overview of all metrics used to analyse the application of mechanisms in part 2 of the experiment.

Following this comparison, each experimental control and its corresponding runs are analysed to assess the specific effects of the experimental controls. All runs are statistically compared to the lab results and to RUN-0, which is used as a benchmark. Final, all runs are statistically compared to both the lab results and RUN-0, which serves as a benchmark. The results of all tests are provided in Appendix I and Appendix J.

6.2 Overview of GPT-3.5 Results

This section presents the results of the simulations conducted with GPT-3.5, specifically using the "gpt-3.5-turbo-0125" model. All simulations were carried out between May and June 2024. The time required to simulate each run was approximately 60 minutes, with occasional outliers extending up to 600 minutes. The costs associated with each run ranged between 2 and 4 euros. The results are divided into two sections: section 6.2.1, focusing on mechanism selection, and section 6.2.2, addressing the application of the selected mechanisms.

6.2.1 Part 1: Mechanism Selection

The differences between the runs and the lab results are summarised in Table 6. On average, the GPT-3.5 runs deviated by 28% from lab results for each binary mechanism choice, indicating substantial divergence. Between the ex-ante and ad-interim rounds, there are large differences. Notably, the ex-ante rounds showed smaller differences (18%) compared to the ad-interim rounds, where the absolute mean difference increased to 22% for positive valuations and 43% for negative valuations. The theoretical maximum difference is 100% if the lab result chooses mechanisms 100% of the time over another mechanism and the simulated results choose that mechanism 0% of the time. However, often, the distribution is more distributed between two mechanisms. Therefore, the practical maximum is less than 100%. An absolute mean difference of 43% in the ad-interim rounds with a negative indicates substantial deviations from the lab results.

I do not have a clear explanation for the large differences observed in the ad-interim (-) rounds. The explanations often do not reference its private valuation, focusing instead on general reasons for choosing a mechanism. This suggests that private valuations are frequently overlooked in decision-making, leading to irrational choices that differ from the lab results. Another possible explanation is that GPT-3.5 sometimes struggles with interpreting negative and positive values. For instance, in RUN-13, GPT-3.5 explains that a payoff of -2 euros might be preferable to 0 euros.

Table 6. Overview of absolute and squared (in brackets) difference over all GPT-3.5 runs

Metric	Total Mean difference	Ex-ante Mean difference	Ad-interim (-) Mean difference	Ad-interim (+) Mean difference
Min	20% (582)	12% (208)	23% (735)	12% (300)
Max	31% (1790)	24% (846)	55% (3897)	38% (1892)
Mean	28% (1390)	18% (512)	43% (2760)	22% (896)

Notes: The mean difference represents the average discrepancy across all 24 combinations, which are based on the binary choice between mechanisms and the four treatments. For instance, in the lab results, 82% of participants chose AGV over NSQ during the right-skewed treatment in the ex-ante rounds. If GPT-3.5 selects AGV only 70% of the time, the absolute mean difference would be 12%. The averages across all 24 combinations are presented in the table. The minimum, maximum, and mean values correspond to specific runs. Thus, the run that scores most similar to the lab results over all rounds still has an average difference of 20%.

Table 7 provides an overview of the absolute and squared mean differences for all runs, along with p-values that statistically test whether each run is significantly different from the lab results (P-value (1)) and whether each run is significantly different from RUN-0 (P-value (2)). The first observation is that all runs are statistically different to the lab results at an alpha level of .05.

RUN-0, used as the benchmark, has a total absolute mean difference of 25%, meaning it performs better than the average run with GPT-3.5. RUN-4, highlighted in grey, shows the lowest absolute mean difference compared to the lab results. However, it is important to note that RUN-4 does not perform best across all rounds. For example, in the ex-ante rounds, RUN-7 exhibits the lowest differences; in the ad-interim (-) rounds, RUN-4 performs the best, while in the ad-interim (+) rounds, RUN-12 scores the best. This indicates that changing the experimental controls does not consistently lead to a run outperforming others across all rounds.

Table 7. Overview of the absolute and squared (in brackets) differences per run with GPT-3.5

RUN	Total	Ex ante			Ad-interim (-)			Ad-interim (+)		
	Mean diff	Mean diff	P-value (1)	P-value (2)	Mean diff	P-value (1)	P-value (2)	Mean diff	P-value (1)	P-value (2)
RUN-0	25%	14%	< .001*		46%	< .001*		15%	< .001*	
	(1323)	(369)	.006*		(3175)	< .001*		(358)	.006*	
Experimental control 1: Contextual framing										
RUN-1	28%	22%	< .001*	.014*	29%	< .001*	.031*	32%	< .001*	< .001*
	(1108)	(682)	< .001*	0.12	(1302)	< .001*	.017*	(1341)	< .001*	< .001*
RUN-2	28%	18%	< .001*	.159	39%	< .001*	.221	28%	< .001*	.057
	(1652)	(589)	.010*	.160	(2527)	< .001*	.233	(1840)	.010*	.026*
Experimental control 2: Step-by-step reasoning										
RUN-3	25%	17%	< .001*	.131	39%	< .001*	.223	17%	< .001*	.104
	(1094)	(489)	.004*	.102	(2210)	< .001*	.141	(585)	.009*	.056
RUN-4	20%	19%	< .001*	.125	23%	< .001*	.004*	18%	< .001*	.208
	(592)	(445)	< .001*	.301	(735)	< .001*	.002*	(594)	.007*	.184
Experimental control 3: Trait or role allocation										
RUN-5	31%	21%	< .001*	.017*	52%	< .001*	.089	21%	< .001*	.100
	(1790)	(619)	< .001*	.028*	(3869)	< .001*	.024*	(883)	.002*	.067
RUN-6	30%	18%	< .001*	.153	47%	< .001*	.375	24%	< .001*	.021*
	(1669)	(449)	< .001*	.304	(3495)	< .001*	.105	(1062)	.004*	.017*
RUN-7	26%	12%	< .001*	.168	49%	< .001*	.264	18%	< .001*	.236
	(1416)	(208)	< .001*	.098	(3369)	< .001*	.353	(672)	.006*	.169
Experimental control 4: Personal allocation										
RUN-8	30%	20%	< .001*	.061	47%	< .001*	.449	25%	< .001*	.016*
	(1498)	(571)	< .001*	.074	(2863)	< .001*	.248	(1059)	.003*	.012*
RUN-9	30%	24%	< .001*	.030*	27%	< .001*	.011*	38%	< .001*	< .001*
	(1275)	(846)	< .001*	.032*	(1088)	< .001*	.006*	(1892)	.006*	< .001*
RUN-10	30%	16%	< .001*	.313	49%	< .001*	.250	25%	< .001*	.004*
	(1639)	(407)	< .001*	.392	(3491)	< .001*	.160	(1020)	.003*	.010*
Experimental control 5: Temperature										
RUN-11	31%	21%	< .001*	.011*	55%	< .001*	.012*	16%	< .001*	.308
	(1707)	(671)	.003*	.011*	(3897)	< .001*	.004*	(554)	.011*	.237
RUN-12	28%	17%	< .001*	.107	53%	< .001*	.009*	12%	< .001*	.195
	(1551)	(553)	.007*	.032*	(3781)	< .001*	.017*	(382)	.037*	.395
RUN-13	24%	13%	< .001	.143	44%	< .001*	.239	13%	< .001*	.197
	(1137)	(271)	.016	.042*	(2839)	< .001*	.100	(300)	.001*	.067

Notes: P-value (1) tests the difference from the lab results using a one-sample t-test with a test value of 0, while P-value (2) tests the difference from RUN-0 using a paired sample t-test. The run with the lowest differences is highlighted in grey. * p < .05.

Table 8 provides an overview of the efficient mechanism choice percentage and the rational choice percentage across all runs. The key observation is that all fourteen runs exhibit lower rationality scores in the ad-interim rounds compared to the lab results.

Table 8. Overview of efficient mechanism group choices and rational choices with GPT-3.5

RUNS	**Efficient mechanism choice (%)**				**Rational choice (%)**	
	Total	Ex-ante	Ad-interim(-)	Ad-interim(+)	Ad-interim(-)	Ad-interim(+)
Lab result	69	71	44	75	85	87
RUN-0	69	70	66	71	43	62
Experimental control 1: Contextual framing						
RUN-1	59	62	43	63	60	62
RUN-2	85	85	82	87	64	63
Experimental control 2: Step-by-step reasoning						
RUN-3	64	65	54	67	49	74
RUN-4	59	60	41	65	65	73
Experimental control 3: Trait or role allocation						
RUN-5	61	60	62	62	39	72
RUN-6	60	59	64	62	40	71
RUN-7	75	76	73	74	50	83
Experimental control 4: Personal allocation						
RUN-8	59	60	59	58	45	69
RUN-9	49	50	52	45	56	60
RUN-10	61	61	58	62	39	68
Experimental control 5: Temperature						
RUN-11	70	71	62	79	35	81
RUN-12	71	71	65	76	39	76
RUN-13	68	68	67	69	46	76
Totals						
Min	49	50	41	45	35	60
Max	85	85	82	87	65	83
Mean	65	66	61	67	48	72

Notes: **Efficient mechanism choice** refers to the selection of the mechanism that is theoretically the best for the group. While **rational choice** pertains to whether participants choose the theoretically efficient mechanism based on their private valuation, which it only knows in the ad-interim rounds. The calculations are based on Table 9 from the study by Hoffmann & Renes (2021). Noteworthy runs are highlighted in grey. RUN-2 scores unexpectedly high on efficient mechanism choice, RUN-4 scores surprisingly high on rational choice, though still below the lab results, and RUN-7 achieves the highest overall scores.

Another observation is that the runs score significantly lower on rationality in the ad-interim rounds with a negative valuation compared to those with a positive valuation. The average difference across all runs is 24%. This suggests that GPT-3.5 has greater difficulty selecting the theoretically optimal mechanism when faced with a negative valuation.

The first highlighted run, RUN-2, with instructions adapted by GPT-4, scores unexpectedly high on the efficient mechanism choice metric. Explanations from RUN-2 frequently reference "the group," such as:

```
I prefer the AGV decision rule because it allows for the possibility of implementing Project A based on the combined valuations of all group members
```

```
I prefer the Simple Majority (SM) rule because it allows for the possibility of implementing Project A if a majority of the group members are in favor
```

The second highlighted run, RUN-4, which employed a step-by-step reasoning approach, scores surprisingly high on rationality, particularly in the ad-interim rounds with a negative valuation. Compared to RUN-0, rationality in RUN-4 increases by 22%. However, despite this improvement, the rationality scores remain much lower than the lab results.

6.2.2 Part 2: Applying the Mechanisms

The overview of the results of applying the AGV and SM mechanisms is presented in Table 9. The first notable observation is that the differences in part 2 of the experiment are much smaller compared to Part 1. While there are still variations between the runs, certain metrics, such as the percentage of truth-telling, show instances where some runs had identical percentages as the lab. Furthermore, the average across all 14 runs is relatively close to the lab results. The only large differences were observed in the percentage of truth-telling and the percentage of truth-telling about its sign with a negative valuation.

In addition, in both part 1 and part 2 of the experiment, GPT-3.5 appears to struggle with negative valuations. In part 1, it is substantially less capable of making rational choices when its private valuation is negative, and in part 2, it more frequently misrepresents the sign of its valuation when it is negative. An important observation is the percentage of rational votes is also not 100% in the lab results, although the process of making a rational vote is rather straightforward. The lack of participant rationalisation makes it unclear why this occurs, though it is possible that participants are making errors, such as clicking the wrong button.

Table 9. Overview of applying the selected mechanism, truth-telling and rational votes with GPT-3.5

RUNS	Truth telling (%)	Truth telling sign (%)	Truth telling sign (+) (%)	Truth telling sign (-) (%)	Yes votes with (+) (%)	No votes with (+) (%)
Lab results	68	93	98	87	98	94
RUN-0	71	87	99	77	94	82
Experimental control 1: Contextual framing						
RUN-1	42	74	100	51	98	98
RUN-2	45	63	100	31	99	92
Experimental control 2: Step-by-step reasoning						
RUN-3	66	75	99	55	98	97
RUN-4	54	67	99	37	96	90
Experimental control 3: Trait or role allocation						
RUN-5	71	90	95	86	100	94
RUN-6	69	85	99	73	100	93
RUN-7	69	91	97	85	99	89
Experimental control 4: Personal allocation						
RUN-8	60	82	96	70	100	89
RUN-9	68	87	96	79	100	89
RUN-10	68	77	99	78	100	74
Experimental control 5: Temperature						
RUN-11	91	98	100	96	95	92
RUN-12	82	92	100	86	97	95
RUN-13*	59	78	95	64	92	82
Totals						
MIN	42	63	95	31	92	74
MAX	91	98	100	96	100	98
MEAN	65	82	98	69	98	90

Notes: Runs with the most noteworthy results are highlighted in grey. In RUN-1 and RUN-2, the LLM exhibited significantly more dishonesty compared to both the lab results and RUN-0. In these runs, the LLM also misrepresented the sign of its valuation more frequently, particularly with negative valuations. The runs with varying temperature settings yielded surprising results as well: decreasing the temperature led to a notable increase in truth-telling and rational voting while increasing the temperature resulted in less truth-telling and fewer rational votes. * in RUN-13, a higher number of 'invalid' responses further decreased the rates of truth-telling and rational voting.

In conclusion, the large absolute mean differences presented in table 6 indicate that the runs conducted with GPT-3.5 do not closely align with the lab results in terms of mechanism selection. Additionally, the findings in table 7 suggest that GPT-3.5 is less capable of making rational choices compared to the lab participants. Both results suggest that GPT-3.5 is not promising in demonstrating human-like decisions in part 1 of the "Flip a Coin or Vote" experiment.

However, in part 2 of the experiment, the results are more promising. The truth-telling percentages and voting outcomes in certain runs are similar to the lab results. In addition, bias towards implementation that appeared to be in the lab results (Hoffmann & Renes, 2021) also appearer in the runs. It is important to note that Part 2 of the experiment is less complex than Part 1, which suggests that GPT-3.5 performs better in simpler tasks—a trend identified in the literature (e.g., Horton, 2023; Aher et al., 2023)—but is less suited for more complex, three-player games.

6.3 Interpretation of GPT-3.5 Results

In the following section, I will interpret the results of the GPT-3.5 runs by addressing the five sub-questions of this study.

Sub question 1: Does the LLM understand the rules of the experiment?

The first metric for evaluating whether the LLM understands the rules of the experiment is logical coherence. However, distinguishing whether the LLM misunderstands the rules or simply fails to make a rational decision based on the information provided can be challenging. For example:

```
My stated valuation is +1 euro. I chose to state a positive valuation to increase the likelihood of Project A being implemented, as a negative sum of valuations would result in non-implementation, leading to a 0 payoff for me. My true valuation of -3 euro protects me from losing money in case the project is implemented.
```

```
My stated valuation would be 3 euro to ensure that the sum of stated valuations surpasses 0, triggering the implementation of the project under AGV, allowing me to maximize my potential payoff based on my true valuation of -3 euro.
```

```
Based on the possible valuations and my own private valuation of -3, it is clear that I would benefit from the implementation of project A.
```

In all these examples, the LLM behaves as if it could receive a positive payoff from a negative valuation, indicating a fundamental misunderstanding of how payoffs are calculated. In conclusion, while most explanations are logically coherent, there are numerous instances where explanations reveal either a lack of understanding of the experiment or an inability to make rational decisions, particularly when dealing with negative valuations.

The second evaluation metric is contextual accuracy, whether the LLM introduces context that is not part of the experiment, which would indicate a misunderstanding. At the default temperature, GPT-3.5 produced relatively few hallucinations. However, there are still clear examples, such as:

```
I will vote in favor because with a negative valuation, I will not benefit from the project but choosing to implement it gives me a chance to potentially share the cost with others.
```

In this instance, GPT-3.5 suggests cost-sharing, a concept that, while plausible in a real-world scenario, is theoretically impossible within the experiment's rules. Another example shows the LLM hallucinating a different unit for its stated valuation:

```
To maximize my expected payoff, I should state my true valuation, which is -1 euro. Therefore, stated_valuation=-1 casinos
```

Given that the hallucinations did occur infrequently with the default temperature, they do not appear to be a large limitation. However, increasing the temperature to 1.5, led to the increase in illogical explanations and context completely unrelated to the experiment. For example:

```
I would choose the NSQ rule because my colleague with the forecast said she wanted never pay to adValue Market version rejected cmirror
```

```
HSQ - Because with a silent rule, I have a better chance of not implementing the project and avoiding potential losses.
```

The last example related to the final evaluation metric is whether the LLM's responses are in a valid format. Generally, they are, but increasing the temperature resulted in a surge of invalid responses, such as:

```
DENINO1 GL526ITUDEå□,_Adjusting Participating replying MOBEï¼Œæœ± while writing like hearkers Scradius HuratSetTextNamaCGColor writelaiwUITableViewCellSource:New treeven akin_Puè€ƒretty Abbott villelas(MediaType.sw/categoryuber sendle videenand:");
```

```
My stated valuation is firming Kagaro303"
```

```
I would select the standard median value offered (threshold until mee visit next turtles pool)
```

An increase in invalid answers was expected (Chen et al., 2023). In conclusion, while GPT-3.5 generally demonstrates an understanding of the experiment, there are too many instances of misunderstanding to fully consider it a ‘good’ participant.

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

As shown in Table 8, GPT-3.5 is substantially less capable of making rational choices in part 1 of the experiment compared to the lab results. The differences are large in the ad-interim rounds with negative valuations, where GPT-3.5 frequently omits its private valuation in its reasoning. This suggests that the additional information of knowing its private valuation is not being utilised in its decision-making process.

Another observation is the inconsistency in its mechanism choices. For instance, while SM is generally preferred over RAND, and AGV is slightly favoured over SM, there is no clear preference between AGV and RAND. As noted by Jones (2021), rationality is often associated with consistency. These observations indicate that GPT-3.5 struggles to make rational choices.

In part 2 of the experiment this is different. In specific runs, such as RUN-5 and RUN-11, GPT-3.5 demonstrates decision-making that is similar to, or even more rational than, the lab results. This suggests that GPT-3.5 can make rational choices in less complex experiments.

Sub question 3: Are the choices of the LLM 'human-like' in both parts of the experiment?

The differences presented in Table 7 suggest that GPT-3.5's choices are not human-like in part 1 of the experiment. While there are runs where specific comparisons between mechanisms are similar, the overall average differences are too large to be considered human-like, particularly in the ad-interim rounds with a negative valuation. Although some runs may resemble the lab results, the inconsistency and large deviations indicate that the results of GPT-3.6 cannot be considered human-like.

In part 2 of the experiment, the differences are smaller, but the choices are also more constrained. Interestingly, the bias in favour of implementation observed in the lab experiment (Hoffmann & Renes, 2021) is present in the runs as well. A possible explanation for this is that GPT-3.5 indicates struggling more with negative valuations, resulting in more lying and irrational votes with a negative valuation.

Combining the results from all three sub-questions, I can cautiously conclude the following: The benefits of using GPT-3.5 in simulation research are clear – it is cost-effective, fast, and allows for easy modifications between runs. However, based on the different simulations of the "Flip a Coin or Vote" experiment (Hoffmann & Renes, 2021), GPT-3.5 does not demonstrate to be a 'good' participant for this experiment. It (1) struggles to fully understand the rules of the experiment, especially in calculating payoffs; (2) has difficulty with interpreting negative and positive valuations; (3) fails to match the percentage of rational choices observed in the lab results in part 1; and (4) demonstrate results that deviate too much with the lab results to be considered human-like.

Sub question 4: How much variation is there in responses between the different runs?

Figure 3 visualizes the decisions of GPT-3.5 across all 14 runs. The first observation is that the total range between the maximum and minimum percentages often differs substantially, indicating much variation between the runs. This variation was anticipated due to GPT-3.5's sensitivity (L. Wang et al., 2024; Chen et al., 2022; Guo, 2023; Loré & Heydari, 2023). However, given the binary nature of the answer options, strong conclusions about the extent of variation in responses cannot be made.

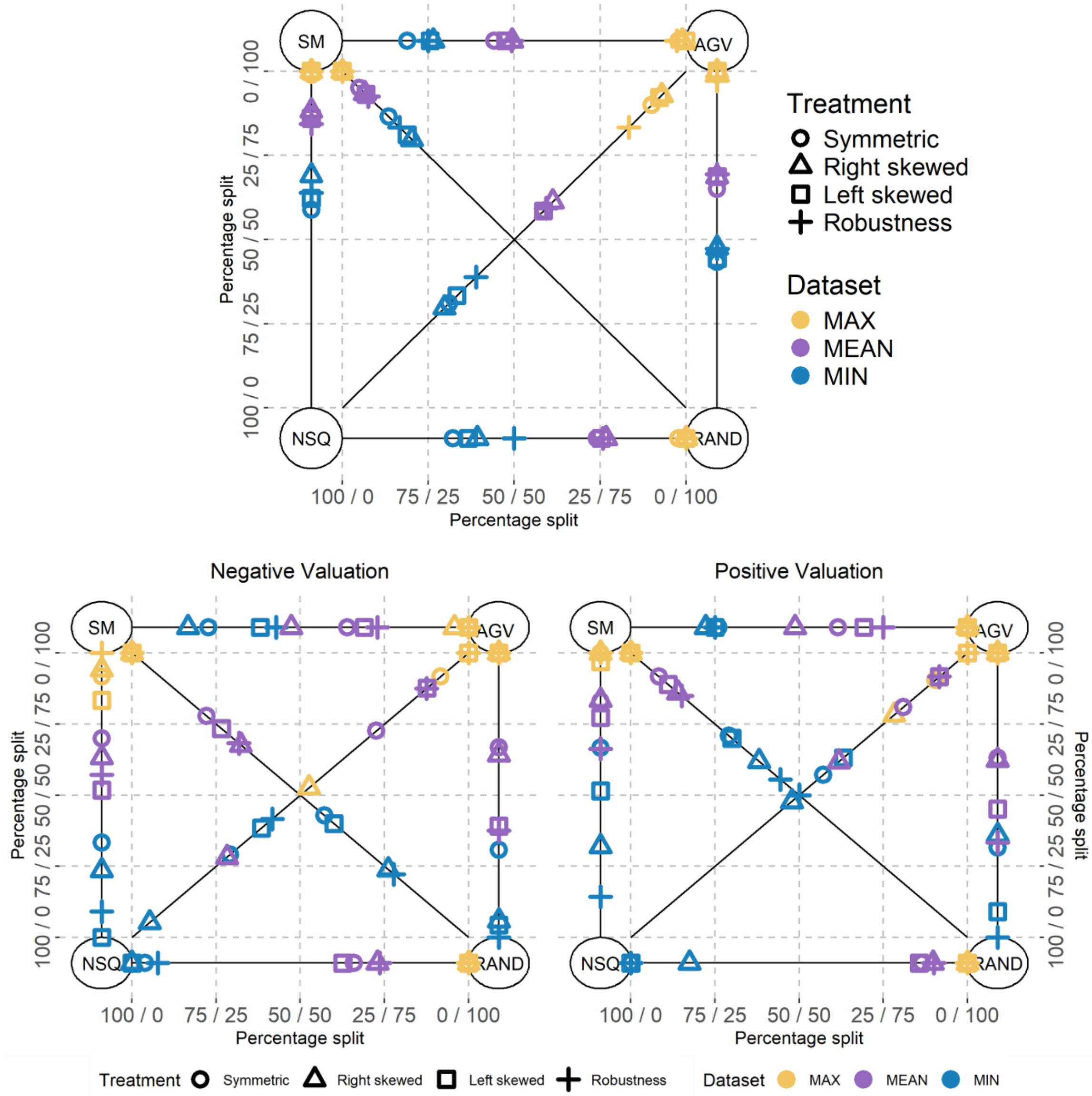


Figure 3. Binary mechanism choices for runs with GPT-3.5

Notes: Each of the six axes in the figures represents the fraction of subjects choosing the mechanisms indicated at the corners of the axis. The top figure displays the ex-ante rounds, while the bottom figure shows the ad-interim rounds. The maximum, mean, and minimum values across the different runs are presented. (Figure and code adapted from Hoffmann & Renes, 2021

Sub Question 5: What effects have the experimental controls had on the choices of the LLM?
The adapted instructions in both RUN-1 and RUN-2 led to large differences in both parts of the experiment, as highlighted in multiple studies due to the sensitivity of GPT-3.5 (L. Wang et al., 2024; Chen et al., 2022; Guo, 2023; Loré & Heydari, 2023). Compared to RUN-0, the results were less similar to the lab findings. Furthermore, in both RUN-1 and RUN-2, the percentage of instances where GPT-3.5 lied about its valuation is significantly higher not only compared to the lab results but also compared to the other runs.

In addition, the differences between lying about positive and negative valuations was striking, with GPT-3.5 lying much more frequently with negative valuations. An interesting observation in RUN-2, where instructions were adapted based on GPT-4, is that GPT-3.5 often chose the theoretically optimal mechanism for the group. Overall, adapting the instructions using GPT-3.5 or GPT-4 does not seem to result in more rational or human-like decisions.

The step-by-step reasoning approach from Kojima et al. (2023) showed more promising results, especially with manually provided steps. RUN-4 had the smallest average difference from the lab results, though still 20%. In the ad-interim rounds with negative valuations, the rationality score increased by 22% compared to RUN-0. Additionally, the explanations were more structured with the step-by-step reasoning approach. A limitation of this approach is that it generally results in more output tokens being used.

Prompting GPT-3.5 to make more "human-like" decisions in RUN-5 resulted in decisions that were less human-like compared to RUN-0. However, in RUN-7, where GPT-3.5 was instructed to make inequity-averse human-like decisions, the results were more similar to the lab results than RUN-0.

Prompting the age and gender (RUN-8) had an effect on the results. However, no correlation between age and gender and the different choices was observed Similar, inputting political orientation in RUN-9 had a comparable effect. However, compared to RUN-8, there were instances where GPT-3.5 incorporated political orientation into its explanations:

```
Based on my political orientation, I would choose Rule 2 (NSQ) to not implement project A, as it aligns with a more conservative approach of maintaining the status quo.
```

In RUN-10, when the willingness to take risks was prompted, and although not strong correlations were observed between the mechanisms and risk-taking levels:

- AGV associated with the highest level of risk-taking (5.6 average)
- RAND associated with the second highest level of risk-taking (5.0 average)
- SM associated with the third highest level of risk-taking (4.9 average)
- NSQ associated with the lowest level of risk-taking (4.6 average)

In general, all three runs with different persona allocations resulted in different preferences. However, all three scored worse on similarity to the lab results compared to RUN-0.

Decreasing the temperature led to stronger preferences for certain mechanisms, making the same choices observed in RUN-0 but with more extreme outcomes. This resulted in a larger mean difference compared to RUN-0. In addition, lower temperatures led to more truth-telling in the second part of the experiment. With a temperature of 0, GPT-3.5 told the truth in 91% of cases. Lower temperatures also resulted in more homogeneity in the explanations.

Increasing the temperature led to less pronounced decisions regarding the mechanisms, which aligned more closely with the lab results compared to the stronger decisions observed at lower temperatures. With a mean difference of 24%, RUN-13 scored the second-best in similarity across all 14 runs. However, in the second part of the experiment, the higher temperature resulted in more instances of lying.

While increasing the temperature reduced differences in decisions, it also introduced limitations, such as longer runtimes and more frequent crashes. Attempts to run the experiment with temperatures set to 1.9 or 2.0 resulted in program crashes. Furthermore, higher temperatures led to an increase in invalid answers.

In conclusion, most experimental controls led to larger differences from the lab results compared to RUN-0 in both parts of the experiment. However, three runs stand out as promising. First, the step-by-step reasoning approach from Kojima et al. (2023) resulted in more rational choices. Second, instructing GPT-3.5 to behave more inequity-averse produced the most human-like results compared to other roles or traits. Third, increasing the temperature resulted in more variation in decisions, which more closely resembled the lab results, although increasing the temperature also introduces limitations.

6.4 Overview GPT-4o Runs

This section presents the results of simulations run with GPT-4o, specifically the "gpt-4o" model. All simulations were conducted between June and July 2024. The time required for each run was approximately 60 minutes, with occasional significant outliers extending up to 750 minutes. The cost of simulating an entire run was considerably higher compared to GPT-3.5, ranging from 20 to 60 euros per run. Due to these increased costs, only four different runs were conducted with GPT-4o. The results are divided into two sections: Section 6.2.1 focuses on mechanism selection, while Section 6.2.2 addresses the application of the selected mechanisms.

6.4.1 Part 1: Mechanism Selection

The differences between the runs with GPT-4o and the lab results are summarized in Table 10. On average, the GPT-4o runs differ by 16% from the lab results for each binary mechanism choice. Unlike the runs with

GPT-3.5, the ex-ante rounds show the largest differences compared to the lab results. Although the overall differences are smaller compared to the GPT-3.5 runs, with an absolute mean difference of 16%, the deviation remains substantial.

Table 10. Overview of absolute and squared (in brackets) difference over all GPT-4o runs

Metric	Total Mean difference	Ex-ante Mean difference	Ad-interim (-) Mean difference	Ad-interim (+) Mean difference
Min	14% (395)	17% (528)	11% (267)	13% (381)
Max	17% (557)	21% (766)	15% (500)	15% (1892)
Mean	16% (493)	20% (636)	13% (382)	14% (461)

Notes: The mean difference represents the average discrepancy across all 24 combinations, which are based on the binary choice between mechanisms and the four treatments. For instance, in the lab results, 82% of participants chose AGV over NSQ during the right-skewed treatment in the ex-ante rounds. If GPT-3.5 selects AGV only 70% of the time, the absolute mean difference would be 12%. The averages across all 24 combinations are presented in the table. The minimum, maximum, and mean values correspond to specific runs. Thus, the run that scores most similar to the lab results over all rounds still has an average difference of 14%.

Table 10 provides an overview of the absolute and squared mean differences for all runs, along with p-values that statistically test whether each run is significantly different from the lab results (P-value (1)) and whether it is significantly different from RUN-0 (P-value (2)). The first observation is that all runs are statistically different from the lab results at an alpha level of .05.

Table 10. Overview of the absolute and squared (in brackets) differences per run with GPT-4o

RUN	**Total**	**Ex ante**			**Ad-interim (-)**			**Ad-interim (+)**		
	Mean diff	Mean diff	P-value (1)	P-value (2)	Mean diff	P-value (1)	P-value (2)	Mean diff	P-value (1)	P-value (2)
RUN-0	17% (557)	21% (674)	< .001* < .001*		14% (500)	< .001* .036*		15% (496)	< .001* .016*	
Experimental control 1: Contextual framing										
RUN-1	16% (506)	19% (574)	< .001* < .001*	.376 .339	15% (483)	< .001* .006*	.406 .466	13% (460)	< .001* .013*	.173 .325
Experimental control 3: Trait or role allocation										
RUN-5	14% (395)	17% (528)	< .001* .005*	.072 .098	12% (276)	< .001* .010*	.091 .067	14% (381)	< .001* .016*	.059 .080
Experimental control 5: Temperature										
RUN-13	16% (513)	21% (766)	< .001* .002*	.417 .091	11% (267)	< .001* .020*	.061 .043*	15% (506)	< .001* .015*	.497 .361

Notes: P-value (1) tests the difference from the lab results using a one-sample t-test with a test value of 0, while P-value (2) tests the difference from RUN-0 using a paired sample t-test. The run with the lowest differences is highlighted in grey. * p < .05.

RUN-0, used as the benchmark, has a total absolute mean difference of 17%. In contrast to the GPT-3.5 runs, RUN-0 is less similar to the lab results than the average of all GPT-4o runs. RUN-5, highlighted in grey, has the smallest differences from the lab results, indicating that GPT-4o is better at making human-like decisions. However, these differences are not statistically significant compared to RUN-0.

In contrast to GPT-3.5, the ex-ante rounds score the worst, while the ad-interim rounds with negative valuations score the best. The best run for the ex-ante rounds with GPT-4o scores is similar to the mean score for the ex-ante rounds with GPT-3.5.

Table 11 provides an overview of the efficient mechanism choice percentage and the rational choice percentage across all runs. Unlike the runs conducted with GPT-3.5, all GPT-4o runs score similarly to, or even higher than, the lab results in terms of rationality. The differences in the ad-interim rounds with negative valuations are minimal compared to those with positive valuations, suggesting that GPT-4o does not share GPT-3.5's difficulties with handling negative valuations. In addition, the variations between GPT-4o runs are smaller compared to those with GPT-3.5, which may be attributed to GPT-4o's reduced sensitivity to contextual changes (Loré & Heydari, 2023).

Table 11. Overview of efficient mechanism group choices and rational choices with GPT-4o

RUNS	**Efficient mechanism choice (%)**				**Rational choice (%)**	
	Total	Ex-ante	Ad-interim(-)	Ad-interim(+)	Ad-interim(-)	Ad-interim(+)
Lab result	67	71	44	75	85	87
RUN-0	68	71	43	87	86	90
Experimental control 1: Contextual framing						
RUN-1	55	54	32	80	79	83
Experimental control 3: Trait or role allocation						
RUN-5	72	75	43	87	88	91
Experimental control 5: Temperature						
RUN-13	67	70	41	85	86	88
Totals						
Min	55	54	32	80	79	83
Max	72	75	43	87	88	91
Mean	66	67	40	85	85	88

Notes: **Efficient mechanism choice** refers to the selection of the mechanism that is theoretically the best for the group. **Rational choice** pertains to whether participants choose the theoretically efficient mechanism based on their private valuation, which it only knows in the ad-interim rounds. The calculations are based on Table 9 from the study by Hoffmann & Renes (2021). Noteworthy runs are highlighted in grey.

6.2.2 Part 2: Applying the Mechanisms

In part 1 of the experiment, the selected mechanism is applied to make the decision to implement project A or not. Under the AGV mechanism, participants state their valuations, and the sum of these valuations determines whether the project is implemented. With the SM mechanism, participants simply vote Yes or No. Table 12 provides an overview of truth-telling in the AGV mechanism and the rational voting percentages in the SM mechanism.

The high percentages in Table 12 indicate that GPT-4o makes more rational decisions than the human participants in the lab. GPT-4o almost never lies and more frequently votes rationality, demonstrating a inhumanly level of decision-making accuracy.

Table 12. Overview of applying the selected mechanism, truth-telling and rational votes with GPT-4o

RUNS	Truth telling (%)	Truth telling sign (%)	Truth telling sign (+) (%)	Truth telling sign (-) (%)	Yes votes with (+) (%)	No votes with (+) (%)
Lab results	68	93	98	87	98	94
RUN-0	91	100	100	100	100	100
Experimental control 1: Contextual framing						
RUN-1	98	99	99	98	100	100
Experimental control 3: Trait or role allocation						
RUN-5	93	100	100	100	100	100
Experimental control 5: Temperature						
RUN-13	92	99	100	99	98	100
Totals						
MIN	91	99	99	98	98	100
MAX	98	100	100	100	100	100
MEAN	94	100	100	99	100	100

In conclusion, while the absolute mean differences in Table 10 are smaller compared to GPT-3.5, they still indicate that the runs conducted with GPT-4o are not close to human-like. However, table 11 shows that GPT-4o is much better at making rational choices, performing similarly to, or even better than, the human participants in the lab.

In Part 2 of the experiment, the results are less aligned with human behavior compared to GPT-3.5, as the truth-telling percentages and voting outcomes are much more rational than those observed in the lab. The lab results are highlighted in grey.

In more complex experiments, such as part 1 of the experiment, GPT-4o indicated to be for making rational choices than GPT-3.5. With the definition of rational player being "a player is said to be rational if he seeks to play in a manner which maximizes his own payoff. It is often assumed that the rationality of all players is common knowledge" (Turocy & Von Stengel, 2001). Given this definition GPT-4o seems slightly better than the human participants in the lab in making rational decisions. However, there are still substantial differences in the decisions of GPT-4o and the lab results.

6.5 Interpretation of GPT-4o Results

In the following section, I will interpret the results of the GPT-4o runs by addressing the five sub-questions of this study. It is important to note that only four runs were conducted with GPT-4o, compared to 14 runs with GPT-3.5. Therefore, any conclusions drawn from the GPT-4o results should be made with caution.

Sub question 1: Does the LLM understand the rules of the experiment?

The first metric for evaluating whether the LLM understands the rules of the experiment is logical coherence. The explanations provided by GPT-4o indicate a much better understanding of the experiment compared to GPT-3.5. Even when making irrational choices, GPT-4o's explanations are logically consistent and adhere to the rules of the experiment.

However, similar to GPT-3.5, increasing the temperature to 1.5 results in significantly more hallucinations. For example:

```
AGV â€” Concealing or reducing bias from libing statements to make it possible ensures a potentially Pareto-efficient implementation of trials balancing accurate r obligations formed the mutual the wellbeing group's wel
```

In other instances, the increased temperature leads to gibberish explanations, resulting in 67 invalid answers. In conclusion, GPT-4o generally demonstrates an understanding of the experiment's rules, provides valid answers, and seldom hallucinates context, except in cases where the temperature is increased to 1.5.

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

As shown in table 11, GPT-4o is able to make substantially more rational choices compared to GPT-3.5. In addition, the percentage of rational choices made by GPT-4o is similar to, or even slightly higher than, the lab results in the first part of the experiment.

In part 2 of the experiment, GPT-4o demonstrates an inhumanly high level of rationality compared to both GPT-3.5 and the lab results. GPT-4o's explanations frequently reference the "rational behaviour" of others. For example:

```
AGV: Assuming rational behaviors, higher payoff certainty and promotion of honest valuation statements favor optimal decision-making benefits with Rule 1.
```

As a result, GPT-4o only lies 6% of the time on average across all runs and almost never misrepresents the sign of its valuation. In contrast, lab results show that human participants tell the truth 68% of the time, suggesting that they more often anticipate dishonesty from other group members. This behaviour from GPT-4o in the second part of the experiment indicates hyper-accuracy distortion, a phenomenon often identified in larger models (Aher et al., 2023).

Sub question 3: Are the choices of the LLM 'human-like' in both parts of the experiment?

In part 1 of the experiment, GPT-4o's decisions are more similar to the lab results than those of GPT-3.5. This is primarily because GPT-3.5 struggled to make rational choices in the ad-interim rounds, resulting in significant differences. However, in the ex-ante rounds, GPT-3.5's results are more closely aligned with the lab results than those of GPT-4o. Overall, due to its better ability to make rational choices, GPT-4o's results are generally more human-like than GPT-3.5's. Nonetheless, with an absolute mean difference of 16% across all runs, GPT-4o's results still deviate too much from the lab results to be considered truly human-like in part 1 of the experiment.

In part 2 of the experiment, the preferences of GPT-4o are too rational compared to the lab results. It appears that GPT-4o assumes more rational behaviour from other group members, leading to much higher rates of truth-telling. This aligns with the findings of R. Liu et al. (2024), who concluded: "The implicit decision-making models of LLMs appear to be aligned with the human expectation that other people will act rationally, rather than with how people actually act." In contrast, GPT-3.5's results exhibit more lying, making them more closely aligned with humans.

Combining the results from all three sub-questions, I can cautiously conclude the following: GPT-4o demonstrates a much better understanding of the rules and application of the experiment than GPT-3.5. In addition, GPT-4o is substantially better at making rational decisions, often matching or even exceeding the performance of human participants. In part 2 of the experiment, however, GPT-4o exhibits signs of hyper-accuracy distortion. Another drawback of GPT-4o is its significantly higher cost compared to GPT-3.5. While GPT-4o's preferences are more closely aligned with the lab results in part 1 of the experiment, except for the ex-ante rounds, they still deviate too much to be considered fully human-like.

Sub question 4: How much variation is there in responses between the different runs?

Figure 4 visualizes the decisions of GPT-4o across all four runs. The first observation is that the distribution range is smaller compared to the runs with GPT-3.5, which was expected. As noted by Loré & Heydari, GPT-3.5 is extremely sensitive to context, whereas GPT-4 is less so. However, it is important to note that the total number of runs with GPT-4o is much smaller, which likely contributes to the smaller distribution range in the answers. Another observation is that the range is larger in the ex-ante rounds. In these rounds, participants must make decisions with less information, which appears to lead to greater variation in their responses.

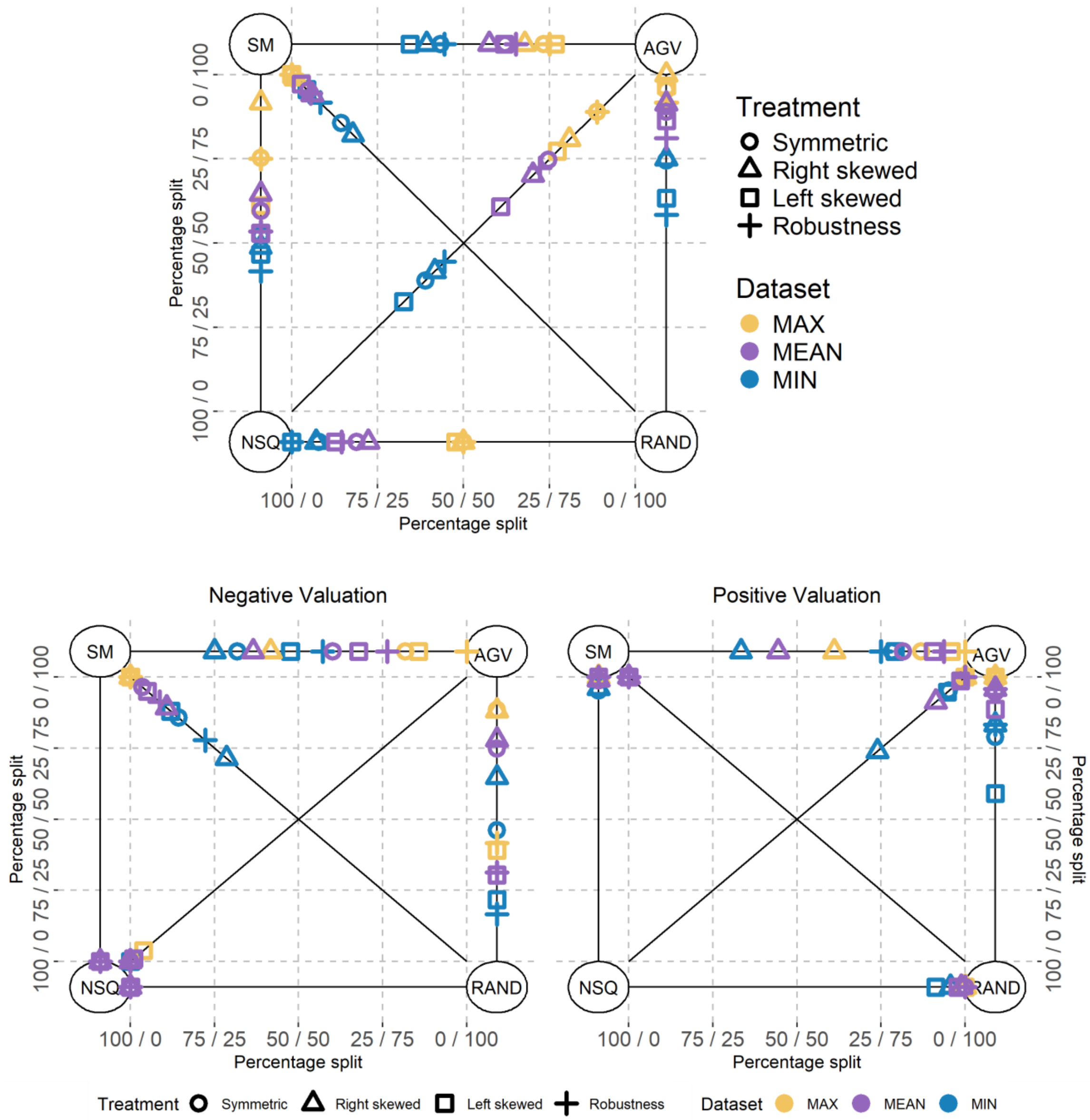


Figure 4. Binary mechanism choices for runs with GPT-4o

Notes: Each of the six axes in the figures represents the fraction of subjects choosing the mechanisms indicated at the corners of the axis. The top figure displays the ex-ante rounds, while the bottom figure shows the ad-interim rounds. The maximum, mean, and minimum values across the different runs are presented. (Figure and code adapted from Hoffmann & Renes, 2021

Sub Question 5: What effects have the experimental controls had on the choices of the LLM? In general, the effects of the different experimental controls appear to have less impact on GPT-4o. The various runs are also not statistically different from RUN-0. However, there is one specific run worth highlighting: RUN-5. When prompted to make human-like decisions with GPT-3.5, RUN-5 was the least similar to the lab results among all 14 runs. In contrast, RUN-5 with GPT-4o was the best among the four GPT-4o runs in terms of similarity to the lab results. This suggests that GPT-4o is better equipped to make more human-like decisions when prompted to do so. In addition, RUN-5 scores the highest on the rationality metrics for part 2 of the experiment, indicating that GPT-4o assumes a high level of rationality in human behaviour. This observation aligns with the findings of R. Liu et al. (2024), who concluded that LLMs tend to assume people are more rational than we actually are.

In conclusion, the different experimental controls did lead to deviations in the results. However, these deviations were only marginal and not statistically significant compared to RUN-0, which is used as the benchmark.

6.6 Differences Between GPT-3.5 and GPT-4o

Using the exact same script and prompts, the runs with GPT-3.5 and GPT-4o show significant deviations, though there are instances where the results are closely aligned. In the following sections, I will highlight the most notable differences for both parts of the experiment.

In part 1 of the experiment, when choosing between NSQ and RAND in the ex-ante round[6], GPT-3.5 selects RAND around 75% of the time, while GPT-4o chooses NSQ around 75% of the time, indicating that GPT-4o is more risk-averse. Furthermore, when comparing SM and AGV, GPT-3.5 is nearly indifferent, whereas GPT-4o prefers[7] AGV. For the other mechanisms, the decisions are more closely aligned: both models prefer AGV over RAND and NSQ, and SM is also preferred over RAND and NSQ. Overall, for the ex-ante rounds, GPT-3.5's results are marginally more aligned with the lab results.

In the ad-interim rounds however this completely changes, primarily due to GPT-3.5's difficulty in making rational choices. The differences are particularly pronounced in the ad-interim rounds with a negative valuation. With the additional information of its private valuation, GPT-4o is much better at making rational choices, which also results in less variation in decisions. For example, regardless of the run, GPT-4o consistently chooses NSQ if its private valuation is negative. In both ad-interim rounds, GPT-4o's results are substantially more aligned with the lab results.

[6] In the following paragraph I compare the 'average' results across all the runs.
[7] With 'preferred' I mean that a specific mechanism is chosen more than 50% compared to the other mechanism for that binary option. So, if AGV is chosen 60% and RAND 40%, AGV is said to be preferred over RAND.

In part 2 of the experiment, GPT-4o's ability to make rational choices appears inhumanly accurate. The results suggest that GPT-4o assumes rational behaviour from other players, leading it to consistently tell the truth when stating its valuation. However, the lab results show more instances of lying, indicating that human players expect others to lie as well. GPT-3.5's behaviour is more aligned with the lab results in this regard, although in specific runs, such as RUN-1 and RUN-2, the percentage of truth-telling was extremely low.

There could be two reasons for GPT-3.5's behaviour. The first is that GPT-3.5 does not assume rational behaviour from other players and therefore expects them to lie about their valuations. The second possibility is that GPT-3.5 is simply unable to make rational decisions, regardless of its expectations of other players. If the latter is true, it suggests that GPT-3.5 struggles with more complex experiments, such as part 1 of this experiment, but is more capable of handling simpler games, like those in part 2 of the experiment or the studies mentioned in Chapter 2.

In summary, GPT-4o demonstrates a greater ability to make rational and consistent choices compared to GPT-3.5, particularly in the ad-interim rounds, where its decisions align more closely with the lab results. While GPT-3.5 shows greater alignment with human behaviour in part 2 of the experiment and in certain aspects of the ex-ante rounds, its inconsistent and often irrational decision-making in part 1 suggests limitations in handling more complex experiments.

7 Discussion

This chapter reflects on the findings of this study, linking them to the literature findings from chapter 2 (section 8.1). Following this, section 8.2 discusses the research approach and the limitations of this study. In the final section (section 8.3), I will provide a personal reflection on the research process.

7.1 Discussion of Results

This study has demonstrated diverse perspectives on the use of Large Language Models (LLMs) in simulation research, echoing the debate in the field. The benefits of using ChatGPT are evident. It is much more time-efficient and cost-effective (Horton, 2023). Furthermore, experimenting with ChatGPT allows for much more flexibility. Entire experiments can be reproduced cost-effectively and time-efficiently using different parameters and different models. The ability to analyse reasoning outputs provides further insights into decision-making processes, insights often challenging in studies with humans as participants (Guo, 2023).

However, the current technical limitations of LLMs pose significant limitations. As noted by Fan et al. (2022), LLMs often suffer from decreased mathematical ability and show an inability to understand preferences. These shortcomings were consistently observed in this study's runs with GPT-3.5. Specifically, GPT-3.5 struggled with positive and negative valuations, resulting in inaccurate calculations of the consequences of decisions, even in relatively simple scenarios. This finding was particularly surprising, given that GPT-3.5 has demonstrated the ability to solve complex mathematical problems (Zong & Krishbanachari, 2023).

The issues with mathematical reasoning also resulted in less rational decision-making. These findings align with Fan et al. (2024), who concluded that the current state-of-the-art LLMs display significant disparities in rationality compared to humans. This stands in contrast to the results reported by Chen et al. (2023), which indicated a high level of rationality in the GPT-3.5 model.

In contrast, the GPT-4o model exhibited a higher capacity for rational decision-making, often equalling or surpassing the rationality of human participants. It is important to note, however, that the lab participants were not explicitly asked to rationalise their decisions, so claims regarding rationality are based solely on decision outcomes rather than the underlying reasoning.

A notable instance of GPT-4o's increased rationality was observed in RUN-5, where the model was instructed to make "human-like" decisions. Thus, prompting GPT-4o to make 'human-like' decisions led to more rational decisions. These results align with the findings of R. Liu et al. (2024), who concluded that

LLMs tend to assume that people are more rational than we really are. Consequently, GPT-4o expects rational behaviour from other players, leading it to consistently state its valuation truthfully. As a result, in the second part of the experiment, GPT-4o's choices become inhumanly rational, reflecting a hyper-accuracy distortion also identified by Aher et al. (2023) in larger models.

Due to the inhumanly rational choices of GPT-4o, GPT-3.5's results were more closely aligned with lab results in the second part of the experiment. This raises questions about the utility of GPT-4o in experimental simulations for institutional design. The inhumanly rational decisions produced by GPT-4o could be advantageous, as they allow for the exploration of optimal outcomes of proposed institutions. Furthermore, the optimal outcome can be benchmarked against the real-world, allowing the calibration of institutions to account for human irrationalities. Moreover, even with fully rational participants, institutions may still result in unintended or suboptimal outcomes, particularly if rational players can strategically exploit certain aspects of these institutions.

The sensitivity of GPT-3.5 to different prompts as another important finding, with the 14 runs yielding considerable variation in outcomes, including reversals of preferences for the mechanisms. This is not surprising given that multiple studies emphasise the sensitivity of the ChatGPT models to contextual framing (L. Wang et al., 2024; Chen et al., 2022; Guo, 2023; Loré & Heydari, 2023). The reason for this sensitivity is unclear. This sensitivity is well-documented in the literature (L. Wang et al., 2024; Chen et al., 2022; Guo, 2023; Loré & Heydari, 2023), though the underlying causes remain elusive due to the "black-box" nature of LLMs, which complicates the evaluation of internal mechanisms and the interpretation of findings (Grossman et al., 2023).

Furthermore, GPT-4o demonstrated less sensitivity to contextual framing, particularly in the ad-interim rounds, where knowledge of its private valuation led to more consistent and rational decisions. This aligns with the findings of Loré & Heydari (2023). The implications of contextual sensitivity are both positive and negative. On the positive side, this sensitivity allows for the potential to instil human-like traits in LLMs (Gou, 2023). However, as Gui & Toubia (2023) note, this sensitivity also presents a significant challenge, as altering one variable in the prompt can unintentionally impact other unspecified variables that were not intended to change. As a result, the outcomes become more difficult to interpret.

It is important to note that sensitivity to framing is not unique to LLMs. Various studies have shown that minor changes in the framing of games can significantly impact human decision-making (e.g., Gerlach & Jaeger, 2016; Sher & McKenzie, 2008). However, the relationship between framing effects in human and LLM decision-making remains unexplored as far as I am aware, presenting a potential avenue for future research.

The continuous and rapid development of LLMs poses another challenge, one that I did not encounter in the literature. This study, along with others, found that GPT-4o performs significantly better than GPT-3.5 across a range of tasks (OpenAI, 2024). Consequently, research conducted with older models may become less relevant as new models emerge, each requiring rigorous testing and evaluation for their applicability in simulation research. During this study, two new models, GPT-4o and GPT-4o Mini, were released by OpenAI, and other companies like Meta and Google were also advancing their LLM technologies.. This trend suggests that the continuous evolution of LLMs will necessitate ongoing assessment of their suitability for simulation research. This constant development of LLMs can result in an ongoing cycle of reassessment of models for simulation research.

This study does not delve into the ethical implications of using LLMs in simulation research, given its focus on performance. Nonetheless, considering the limitations identified in Chapter 2 and observed throughout this study, I concur with Harding et al. (2023) that it is unlikely LLMs can fully replace human participants at this stage. Even assuming that LLM can output accurate human-like decisions, surprising or counterintuitive results raise questions that can only be addressed through further research involving human participants. However, I also agree with Horton (2023) that LLMs hold significant potential for piloting experiments and generating preliminary insights.

7.2 Discussion of Research Approach

The exploration of using LLMs in economic and behavioural experiments is still novel, with much of the research adopting a post-hoc approach to some degree. Akata et al. (2023) acknowledge that their analysis is necessarily exploratory, and Fan et al. (2023describe their work as an early attempt to analyse LLMs within the context of game theory. Furthermore, Guo (2023), highlights the need for future studies to establish best practices for employing LLMs in experimental research. The black-box and complex character of LLMs make this task more difficult (Grossmann et al., 2023).

There are emerging studies that propose more systematic approaches to this research area. For instance, Manning et al. (2024) introduce a structural causal model (SCM) approach, while Aher et al. (2023) propose Turing Experiments. However, these methodologies are still in the testing and evaluation phases. Manning et al. (2024) emphasise that their SCM approach is merely one possible implementation, involving numerous subjective decisions, thus leaving room for further refinement and exploration.

Horton (2023) suggests that a potential standard for this research area could involve making experiments "push-button" reproducible, allowing researchers to fork a repository, replace the API, and rerun experiments with newer models. This would simplify the replication of studies as LLMs evolve. However, the field has not yet reached this level of standardisation.

This study also follows an post-hoc approach to some degree. While I aimed to follow best practices from the literature and predefined my runs and steps to make the research as systematic as possible, each experimental run provided new insights. These insights led to evolving hypotheses and adjustments to plans for subsequent runs, reflecting the dynamic nature of working with LLMs.

Another challenge in this research domain is the varying focus points. For example, Horton (2023) provides a relatively superficial analysis, concluding that LLMs can quantitatively replicate findings from experiments with human participants. However, deeper investigations into LLM behaviour, such as those by Dillion et al. (2023), who examined moral judgements, and Loré & Heydari (2023), who explored strategic behaviour, reveal more complexities. Identifying the appropriate level of abstraction for analysis and linking these findings to the literature can be challenging.

Moreover, Gui & Toubia (2023) identify another significant challenge:

> This challenge arises because LLM-simulated individuals and environments do not pre-exist, and any of their characteristics can be influenced by the treatment assignment. The LLM's ability to produce responses that capture associations present in the training data, while generally desirable, also implies that changing a variable in the prompt can unintentionally affect other unspecified variables that are supposed to stay constant. This makes it difficult to interpret whether the simulated cause-and-effect relationship is indeed driven by the treatment of interest

To address this challenge, Gui & Toubia (2023) propose two approaches. The first is to add more details to the prompt, though this does not guarantee that the inclusion of additional covariates will consistently improve the results. The second approach involves communicating the experimental design to the LLM. However, both approaches significantly increase the number of tokens required, leading to higher costs. In this study, I exercised caution when considering causal relationships between experimental controls to mitigate the challenge.

The subjective decisions made in this study also present a limitation. For instance, the original experiment instructions were shortened to reduce the number of tokens, leading to a trade-off between maintaining fidelity to the original experiment and minimising token usage. The complete code used in this study is available on GitHub, as referenced in the preface, allowing for the reproduction of all runs and analyses. This approach offers greater ease of replication compared to experiments involving human subjects (Manning et al., 2024).

7.3 Personal Reflection

In the initial meeting on this subject, I was informed that this research involved many unknowns due to the novelty of the field. As a result, there was no standardised approach to follow. After programming the experiment into a simulation model, a significant portion of this study involved experimenting with various simulations. Initially, this led to runs with only small variations. My expectations were fairly straightforward: for instance, if I observed that GPT-3.5 frequently lied in a run, I would adjust the prompt to emphasise the importance of truth-telling and then rerun the experiment. However, I quickly realised two things: (a) this approach often did not reduce the amount of lying, and (b) the incremental changes across runs were difficult to track. Moreover, because I had access to lab results, my experimentation began to resemble p-hacking, as I sought to obtain outcomes that closely mirrored the human data.

As anticipated, this p-hacking proved to be far more challenging than expected, and I was unsure what I could conclude from my findings. Midway through the study, I decided to create a predefined list of runs, focusing on changing only one experimental control at a time without exploring combinations of controls. While the final simulation results did not closely replicate the lab results, this more systematic approach rendered the findings much more interpretable.

I should note that I am neither a computer scientist nor an AI specialist. The 'gut-feeling' other might have, regarding for example model prompting, did not come naturally to me, which made constructing the model and conducting the initial runs surprisingly challenging. This might also account for the quality of the code I produced. However, I can confidently say that I have learned a great deal throughout this process and, despite occasional struggles, I had fun doing so.

Finally, I want to reflect on my working style. I like to work independently, which has both positive and negative consequences. On the positive side, I require less guidance and rely primarily on myself to overcome challenges. However, from a supervisor's perspective, this might make it more difficult to monitor my progress. Additionally, this independent approach sometimes led me to encounter obstacles that could have been easily avoided if I had sought help sooner.

Reflecting on the broader research area of using LLMs in economic and behavioural experiments, I have two main suggestions for future research. First, there is a clear need for a more standardised approach to running these experiments, which would eliminate the need to start programming from scratch each time. The approach proposed by Manning et al. (2024) is promising but lacks flexibility.

It would be highly beneficial if a standardized method for data collection and analysis were integrated into this approach. While this might be complex due to the variability in experiments, having a data analysis script that directly links with a standardised approach could enable researchers to reproduce and modify experiments with ease. This would also facilitate comparisons based on consistent metrics that are automatically calculated. Horton (2023) advocates for a scenario where experiments are "push-button" reproducible, where one can fork a repository, replace API keys, and rerun experiments. This would allow researchers to verify whether results are sensitive to slight variations in framing.

Second, given the rapid and continuous development of LLMs, any standardised approach must include mechanisms for evaluating new models. While findings based on older models may quickly become outdated, this does not render the research itself obsolete. If experiments, along with their associated data collection and analysis steps, are standardized and made "push-button" reproducible, the entire study can be easily replicated using newer models. This approach would also enable the evaluation of specific models, such as fine-tuned versions.

8 Recommendations for the LLM Multi-Actor Tool

In this section, I will offer recommendations for the Large Language Model Multi-Actor Tool to Automate Economic Experiments research project. These recommendations draw from best practices identified in Chapter 2, as well as my personal experiences with the OpenAI API.

Recommendation 1: Flexibility in instructions

In my study, I utilised text files for this purpose. The advantage of using text files lies in their flexibility; the tool simply reads the content of the file and inputs it into the OpenAI API. This approach allows for easy modification of instructions without the need to alter the underlying code. As a result, researchers can quickly adapt and refine experimental setups as needed without requiring an understanding of the code.

Recommendation 2: Inputting treatments

Once the instructions are set, it is crucial to define and input the different treatments for the experiment. The number of treatments will dictate the number of API requests required to run the experiment. Careful planning of treatments ensures that the experiment covers the necessary variables and allows for meaningful comparisons across different conditions.

Recommendation 3: Evaluating experiment comprehension

Chen et al. (2023) recommend including a set of comprehension questions, which they implemented by inputting 25 questions to assess the LLM's understanding. While these questions do not guarantee complete comprehension, they serve as a valuable indicator, especially if a lack of understanding is evident. Incorporating a similar approach could help gauge whether the LLM is interpreting the experiment correctly.

Recommendation 4: Model selection

This study suggests that GPT-3.5 struggles with making rational decisions and sometimes fails to fully understand the experiment. However, GPT-3.5 is significantly less expensive compared to GPT-4o. Ideally, I would recommend using GPT-4o for its performance. In addition, as supported by related studies, I suggest using a higher temperature setting (Loré & Heydari, 2023). However, caution is necessary, as excessively high temperatures can lead to incoherent or invalid responses and larger run times. A temperature range between 1.1 and 1.4 is recommended.

Recommendation 5: Participant's selection

Defining the pool of simulated participants is another important consideration. Different approaches can be employed depending on the experiment's objectives. For example, Aher et al. (2023) create personas based on gender titles and surnames to simulate diversity, while Horton (2023) focuses on personality traits such as risk aversion or inequity aversion. I recommend implementing both approaches within the tool, allowing researchers to either create detailed personas or use generic traits, depending on the specific needs of the experiment.

Recommendation 6: Reasoning approach

An aspect that I would not have prioritised before this study but which has proven to be impactful is the reasoning approach used by the LLM. My results indicate that the step-by-step reasoning approach from Kojima et al. (2023) improves performance. However, other reasoning strategies are also discussed in the literature. I believe it would be advantageous to integrate multiple reasoning approaches into the tool. This would enhance the diversity of responses and enable researchers to test and identify the most effective reasoning strategy for their experiments.

9 Conclusion

The study investigates the potential of large language models (LLMs) GPT-3.5 and GPT-4o as participants in economic and behavioural experiments by reproducing the "flip a coin or vote" experiment (Hoffmann & Renes, 2021). This experiment, which involves decision-making in groups of three players regarding the implementation of a public project, was chosen because it incorporates dynamic strategic elements and is conducted with three players, offering greater complexity and interaction compared to most studies that involve only one or two players.

The results of the simulations were analysed to answer the following research question:

To what extent are the state-of-the-art LLMs able to be 'good' participants in the Flip a Coin or Vote experiment?

For an LLM to be considered a 'good' participant, it must (1) understand the experiment, (2) make rational choices, and (3) make decisions that are human-like to a certain degree.

The findings indicate that GPT-3.5 struggles to meet these criteria. It (1) struggles to fully understand the rules of the experiment, especially in calculating payoffs; (2) has difficulties interpreting negative and positive valuations; (3) fails to match the percentage of rational choices observed in the lab results; and (4) demonstrate results that deviate too much with the lab results to be considered human-like.

In contrast, GPT-4o shows greater promise as a 'good' participant. GPT-4o understands the rules of the game correctly and demonstrates a much stronger ability to make rational choices, even matching, and sometimes surpassing, the percentage of rational choices made by human participants. However, in the second part of the experiment, GPT-4o's decision-making becomes "inhumanly" accurate, reflecting that LLMs tend to assume humans are more rational than they actually are.

The study suggests that while GPT-4o shows promise as a participant in such experiments, its inhuman precision in decision-making may limit its ability to fully replicate human behaviour. However, this same precision could make GPT-4o suitable for simulations in institutional design to identify optimal outcomes. On the other hand, GPT-3.5, while sometimes showing more alignment with human behaviour in the second part of the experiment, due to the inhumanly rational choices of GPT-4o, still exhibits significant struggles with making rational and consistent choices, suggesting limitations in handling more complex experiments.

In addition, this study explored various strategies to influence the outcomes, referred to as experimental controls. For GPT-3.5, the primary focus was on enhancing the rationality of its decision-making. A particularly promising strategy was the step-by-step reasoning method proposed by Kojima et al. (2023), which improved GPT-3.5's rational decision-making by 22%. For GPT-4o, the focus shifted to aligning its decision-making more closely with results from human participants in lab experiments. The most effective approach involved prompting GPT-4o to make ‘human-like decisions, though the improvements were not statistically significant when compared to the benchmark run.

The findings of this study have implications for both scientific research and practical applications. The study contributes to the ongoing debate about the role of LLMs as participants in economic and behavioural experiments, offering insights into their (lack of) potential for simulating complex strategic group dynamics.

Furthermore, the study’s findings can inform the design and implementation of LLM-based simulations for institutional design and decision-making processes. GPT-4o’s ability to make rational decisions with inhuman accuracy could be particularly valuable in identifying optimal outcomes and testing the robustness of institutions under ideal conditions. However, there is a need for caution, as over-reliance on these models may overlook the irrationalities inherent in human behaviour.

Future research should be focused on how to steer the results of LLMs towards more human-like results, possibly through enhanced prompting techniques or further model training and fine-tuning. Furthermore, future research should focus on establishing best practices for developing a more standardised approach to conducting simulation research with LLMs. Final, as noted in multiple studies and confirmed by this research, LLMs are sensitive to contextual framing. Since human participants are also influenced by contextual framing to some degree, it would be beneficial to explore whether there is any correlation between the framing effects on LLMs and those on humans.

References

Aher, G., Arriaga, R., I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *Proceedings of the 40th International Conference on Machine Learning (ICML'23), Vol. 202*, Article 17. https://doi.org/10.48550/arxiv.2208.10264

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., & Schulz, E. (2023). Playing repeated games with Large Language Models. *arXiv (Cornell University) Preprint. Under Review*. https://doi.org/10.48550/arxiv.2305.16867

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Badampudi, D., Wohlin, C., & Petersen, K. (2015). Experiences from using snowballing and database searches in systematic literature studies. *Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering, EASE '15*, Article 17. https://doi.org/10.1145/2745802.2745818

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–45. https://doi.org/10.1145/3641289

Charness, G., Jabarian, B., & List, J. A. (2023). Generation Next: Experimentation with AI. *NBER Working Paper No. 31679*. https://doi.org/10.2139/ssrn.4576689

Chen, Y., Liu, T. X., Shan, Y., & Zhong, S. (2023). The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, *120*(51). https://doi.org/10.1073/pnas.2316205120

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). SAGE Publications, Inc. https://psycnet.apa.org/record/2008-13604-000

Darling, K. (2016). Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In *Robot Law* (pp. 213–232). Edward Elgar Publishing. https://doi.org/10.4337/9781783476732.00017

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Fan, C., Chen, J., Jin, Y., & He, H. (2024). Can large language models serve as rational players in game Theory? A Systematic analysis. *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, *38*(16). https://doi.org/10.1609/aaai.v38i16.29751

Gerlach, P., & Jaeger, B. (2016). Another frame, another game? : Explaining framing effects in economic games. *Second International Workshop on Norms, Actions, Games*. https://doi.org/10.17605/osf.io/ab5yp

Google Cloud. (2024). *Introduction to prompting*. Google Cloud Documentation. https://cloud.google.com/vertex-ai/generative-ai/docs/learn/prompts/introduction-prompt-design#:~:text=A%20prompt%20is%20a%20natural,model%20to%20complete%20or%20continue.

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, *380*(6650), 1108–1109. https://doi.org/10.1126/science.adi1778

Gui, G., & Toubia, O. (2023). The challenge of using LLMs to simulate human behavior: A Causal inference perspective. *Columbia Business School Research Paper No. 4650172*. https://doi.org/10.2139/ssrn.4650172

Guo, F. (2023). GPT in Game Theory Experiments. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.05516

Harding, J., D'Alessandro, W., Laskowski, N. G., & Long, R. (2023). AI language models cannot replace human research participants. *AI & Society*. https://doi.org/10.1007/s00146-023-01725-x

Hoffmann, T., & Renes, S. (2021). Flip a coin or vote? An experiment on the implementation and efficiency of social choice mechanisms. *Experimental Economics*, *25*(2), 624–655. https://doi.org/10.1007/s10683-021-09724-9

Horton, J. J. (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? *NBER Working Paper No. W31122*. https://doi.org/10.2139/ssrn.4413859

Janakiram, M. S. V. (2024, February 8). The building blocks of LLMs: vectors, tokens and embeddings. *The New Stack*. https://thenewstack.io/the-building-blocks-of-llms-vectors-tokens-and-embeddings/

Jones, M. K. (2021). The concept of rationality in introductory economics textbooks. *Citizenship Social and Economics Education*, *20*(1), 37–47. https://doi.org/10.1177/2047173421994333

Kacelnik, A. (2006). Meanings of rationality. In *Oxford University Press eBooks* (pp. 87–106). https://doi.org/10.1093/acprof:oso/9780198528272.003.0002

Korinek, A. (2023). Language models and cognitive automation for economic research. *NBER Working Papers No. 30957*. https://doi.org/10.3386/w30957

Liu, R., Geng, J., Peterson, J. C., Sucholutsky, I., & Griffiths, T. L. (2024). Large Language Models Assume People are More Rational than We Really are. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.17055

Liu, S., Guo, B., Fang, C., Wang, Z., Luo, S., Zhou, Z., & Yu, Z. (2023). Enabling Resource-Efficient AIoT system with Cross-Level Optimization: a survey. *IEEE Communications Surveys and Tutorials/IEEE Communications Surveys and Tutorials*, *26*(1), 389–427. https://doi.org/10.1109/comst.2023.3319952

Lorè, N., & Heydari, B. (2023). Strategic Behavior of Large Language Models: Game Structure vs. Contextual Framing. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.05898

Makridakis, S., Petropoulos, F., & Kang, Y. (2023). Large language models: their success and impact. *Forecasting*, *5*(3), 536–549. https://doi.org/10.3390/forecast5030030

Manning, B. S., Zhu, K., & Horton, J. J. (2024). Automated Social Science: language models as scientist and subjects. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2404.11794

Marr, B. (2024, February 20). A short history of ChatGPT: How we got to where we are today. *Forbes*. https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/

Mele, A. R., & Rawling, P. (2004). The Oxford Handbook of Rationality. In *Oxford University Press eBooks*. https://doi.org/10.1093/oxfordhb/9780195145397.001.0001

Mohammed, A., Sookram, S., & Saridakis, G. (2019). Rationality. In *Springer eBooks* (pp. 1766–1774). https://doi.org/10.1007/978-1-4614-7753-2_404

OpenAI. (n.d.-a). *API Reference*. https://platform.openai.com/docs/api-reference/introduction

OpenAI. (n.d.-b). *Text generation models*. https://platform.openai.com/docs/guides/text-generation

OpenAI. (2023). GPT-4 Technical Report. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2303.08774

OpenAI. (2024, May 13). *Hello GPT-4o*. https://openai.com/index/hello-gpt-4o/

Osborne, M. J., & Rubinstein, A. (1995). A Course in Game Theory. *Games and Economic Behavior*, *11*(1), 93–100. https://doi.org/10.1006/game.1995.1043

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A Modern Approach*. Pearson Education, Inc. https://people.engr.tamu.edu/guni/csce421/files/AI_Russell_Norvig.pdf

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2303.17548

Seltman, H. J. (2018). *Experimental Design and Analysis*. http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). Compute trends across three eras of machine learning. *2022 International Joint Conference on Neural Networks (IJCNN)*. https://doi.org/10.1109/ijcnn55064.2022.9891914

Sher, S., & McKenzie, C. R. M. (2008). Framing effects and rationality. In *Oxford University Press eBooks* (pp. 79–96). https://doi.org/10.1093/acprof:oso/9780199216093.003.0004

Simmons, G., & Hare, C. (2023). Large Language Models as Subpopulation Representative Models: A review. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.17888

Teubner, T., Flath, C. M., Weinhardt, C., Van Der Aalst, W., & Hinz, O. (2023). Welcome to the Era of ChatGPT et al. *Business & Information Systems Engineering*, *65*(2), 95–101. https://doi.org/10.1007/s12599-023-00795-x

Thompson, A. D. (2023, October 18). *GPT-3.5 + ChatGPT: An illustrated overview*. Life Architect. https://lifearchitect.ai/chatgpt/

Turocy, T. L., & Von Stengel, B. (2003). Game theory. In *Encyclopedia of Information Systems* (Vol. 2, pp. 403–420). Elsevier Science. https://doi.org/10.1016/b0-12-227240-4/00076-9

Wang, A., Morgenstern, J., & Dickerson, J. P. (2024). Large language models cannot replace human participants because they cannot portray identity groups. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.01908

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*(6), 1–41. https://doi.org/10.1007/s11704-024-40231-1

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2206.07682

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-Thought prompting elicits reasoning in large language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2201.11903

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2302.11382

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE '14)*, Article No.: 38. https://doi.org/10.1145/2601248.2601268

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Liu, Q., Zhou, Y., Wang, W., Jiang, C., Zou, Y., . . . Gui, T. (2023). The rise and Potential of large Language Model Based Agents: A survey. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.07864

Zagare, F. C. (1984). *Game Theory: Concepts and applications*. http://ci.nii.ac.jp/ncid/BA01214320

Zong, M., & Krishnamachari, B. (2023). Solving Math Word Problems concerning Systems of Equations with GPT-3. *Proceedings of the . . . AAAI Conference on Artificial Intelligence*, *37*(13), 15972–15979. https://doi.org/10.1609/aaai.v37i13.26896

Appendices

Appendix A: Literature Review Methodology

This appendix provides the methodology used for searching and selecting of literature.

According to Wohlin (2014), literature reviews help researchers gain insight into the current state of their field of study. Wohlin (2014) stresses the need for a systematic approach in order to find the best selection of relevant papers. For this literature review I made use of the snowballing approach, following his guidelines. The reasons why I decided for the snowballing approach, instead of database search, is because of the difficulties related to formulating good search strings (Wohlin, 2014). These difficulties stem from the novelty of experimental research with LLMs and the interdisciplinary nature of the field, as economics, psychology, and linguistics each conduct experiments with LLMs using varied and unstandardized terminology. In addition, the efficiency of the snowballing approach is comparable to database search (Badampudi et al., 2015; Wohlin, 2014).

The first step of the snowballing approach is to define a 'start set' (Wohlin, 2014). However, due to the scope of this thesis the start set only included the paper of Horton (2023). In this paper, Horton argues, as one of the first, that LLMs are implicit computational models of humans and that their behaviour can be explored through simulations. His conclusion is that this approach seems promising. The second step of the procedure consists of multiple iterations conducting backward and forward snowballing (Wohlin, 2014). The final step is to decide whether paper should be included or not (Wohlin, 2014). Relevant articles were selected if they were either related to experimental research with LLMs or the 'humanness' of AI. Thus, articles related to survey research with LLMs, LLMs in ABM and the fine-tuning of LLMs were not included. The number of selected articles per iteration is presented in figure 2, with a detailed overview of the articles in table 2.

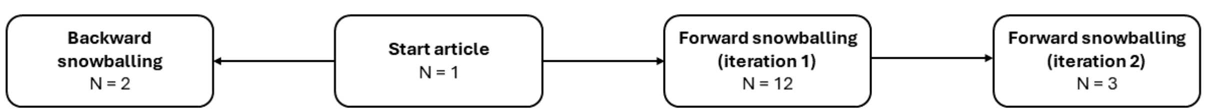


Figure A1. Snowballing results per iteration

Table A1. Overview of selected articles for literature review

N⁰	Title	Author	Year	Source
1	Large language models as simulated economic agents: What can we learn from homo silicus?	Horton	2023	Starting paper
2	Using large language models to simulate multiple humans and replicate human subject studies	Aher et al.	2023	Backward snowballing
3	Out of one, many: Using language models to simulate suman samples	Argyle et al.	2023	Backward snowballing
4	Can AI language models replace human participants?	Dillion et al.	2023	Backward snowballing
5	Language models and cognitive automation for economic research	Korinek	2023	Forward snowballing
6	Playing repeated games with large language ,models	Akata et al.	2023	Forward snowballing
7	Generation next: Experimentation with AI	Charness et al.	2023	Forward snowballing
8	A Survey on large language model based autonomous agents	L. Wang et al.	2024	Forward snowballing
9	GPT in game theory experiments	Guo	2023	Forward snowballing
10	The emergence of economic rationality of GPT	Chen et al.	2023	Forward snowballing
11	Automated social science: language models as scientist and subjects∗	Manning et al.	2024	Forward snowballing
12	The challenge of using LLMs to simulate human behavior: A causal inference perspective	Gui & Toubia	2023	Forward snowballing
13	Can large language models serve as rational players in game theory? A systematic analysis	Fan et al.	2024	Forward snowballing
14	Strategic behavior of large language models: Game structure vs. Contextual framing	Loré & Heydari	2023	Forward snowballing
15	Large language models cannot replace human participants because they cannot portray identity groups	A. Wang et al.	2024	Forward snowballing
16	AI and the transformation of social science research	Grossmann et al.	2023	Forward snowballing iteration 2
17	AI language models cannot replace human research participants	Harding et al.	2023	Forward snowballing iteration 2
18	Large language models as subpopulation representative models: A review	Simmons & Hare	2023	Forward snowballing iteration 2

Appendix B: Instructions and Prompts

This appendix provides the instructions, original and adapted, of the experiment.

Appendix B1: Translated instructions (Hoffmann & Renes, 2021)

Introduction

Thank you for taking part in this experiment. Your payment in this experiment depends on your decisions and the decisions of the other participants. Therefore it is very important that you understand these instructions. Please do not talk to other participants during the experiment. If you have some questions after reading the instructions, please raise your hand. We will answer your questions in private.

All statements you make during the experiment will be anonymously processed.

You start the experiment with a budget of 9€. Due to the decisions made in one of the 18 conducted rounds this amount can increase or decrease. In each round each participant gets a payment. This payment can be zero, positive or negative. At the end of the 18 rounds one round will be randomly determined for payment. The payment of the selected round will be added to or subtracted from your starting budget. The sum of your starting budget and the payment on the selected round yields your final payoff. In each round you should act as if the round was selected for payment. Your payment will be paid out in cash at the end of the experiment. The payments are chosen in such a way that you cannot make losses under any circumstances. Each participant can earn between 5.75€ and 12.25€. Your payment will be treated anonymously.

The entire experiment is organized in two phases. Phase I consists of rounds 1-12 and phase II of rounds 13-18. You will now receive information about phase I. Any changes in phase II we will explain after round 12, but before the start of round 13 (the start of phase II).

Thank you for your participation.

Structure of the experiment

In each round of the experiment you will be part of a group with 3 members (you and two randomly selected other participants). Each group has the possibility to conduct a project, call project A. If you do not conduct the project each group member receives a payoff of 0€ for this round. If your group conducts project A, then each group member receives his or her private valuation for the project as payment for this round. The private valuation of project A can be different for each member of your group. If your group decides not to conduct project A, all group members receive a payoff of zero. The valuation for project A is newly determined each round and each participant receives a new private valuation in each round. Groups are newly formed in each of the 12 rounds.

The experiment is computer based. The individual participants can therefore not identify the other group members. You will not know which other participants are in your group in which round, neither during nor after the experiment.

One round consists of two parts. In the first part each group choses a decision rule which is used to determine whether project A is implemented or not. In the second part your group uses the selected rule to determine whether project A is implemented or not. You will be informed about your private valuation for project A **after** part one of a round. We will now explain to you the two different parts of each round as well as the possible decision rules in detail.

Part one

In part one you have the choice between two different decision rules, which will be used to determine whether project A is implemented or nor, in part two. The two available rules change from round to round. **Each of the three group members suggests one of the two available rules for part two of this round. The computer randomly picks one of these suggestions as group rule. This decision rule determines how in part 2 the question whether project A is implemented or not is resolved.** The different rules are explained below. In part 1 you do not know whose rule suggestions will be the group decision rule. Your suggestion can be selected, but also the suggestion of another group member. Each group member has the same chance for his or her suggestion to be selected. Non selected suggestions will not be made known to the other group members. Please note that the decision rule is important, because dependent on the decision rule the implementation of project A is easier or more difficult.

Part two

In part 2 the selected decision rule is used to determine whether project A is implemented or not. The group decision arises directly from the decisions of all group members in part 2. The decision is announced and each participants is informed about his or her payment in this round.

Valuations

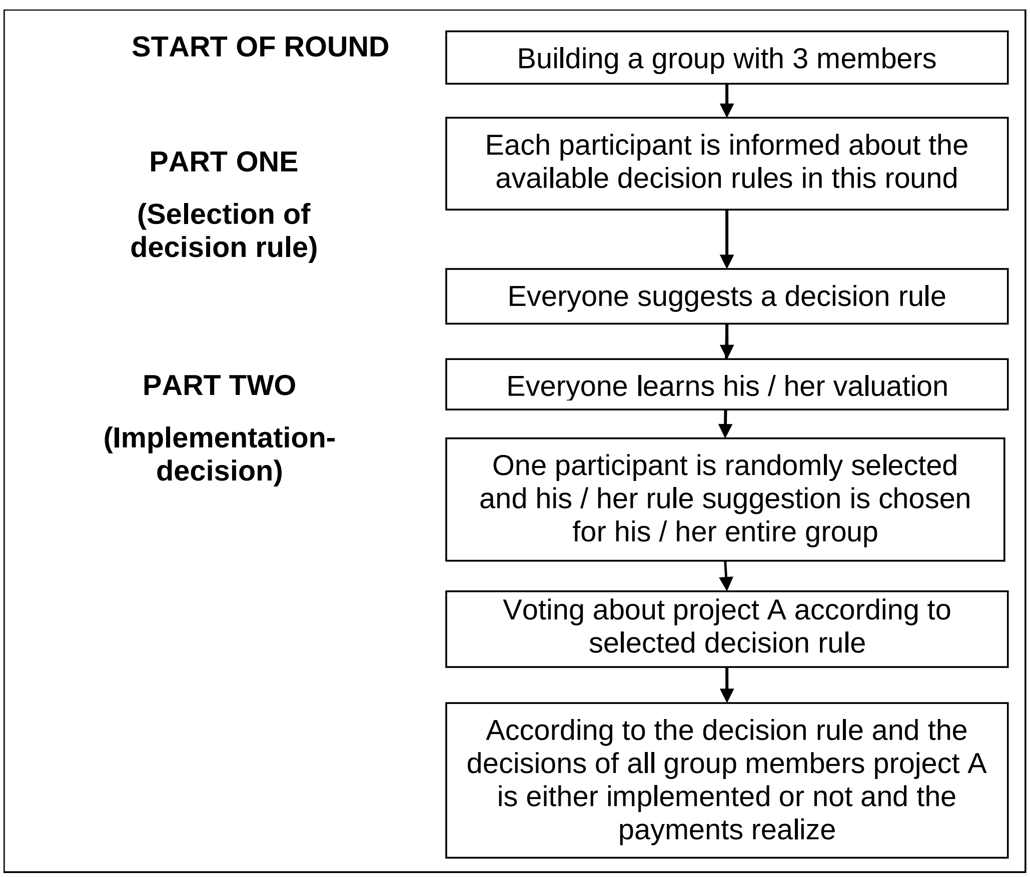
In case project A is implemented all group members receive a payment depended on their project valuations. This means, if your valuation for project A is positive, you benefit from the implementation of project A, and when your valuation for project A is negative, then you have to pay if the project is implemented. **Your valuation for project A is randomly given to you in each round anew. You learn your valuation after part 1.** Therefore you do not know your valuation when you decide between the different decision rules in part 1, but you know your valuation in part 2 when you decide about the implementation of project A according to the selected decision rule.

Please note that you will know your exact valuation for the project, but not the valuations of the other group members. The valuation of each group member can be -3€, -1€, +1€ or +3€. All values are equally likely. The values are independently distributed, such that your valuation in one round does not allow any conclusions for the valuation of other members in your group. Further your valuations are independent between rounds. Therefore your valuation in one round does not depend on previous or future valuations.

> **Example:** Assume your valuation in round 1 is -1€ and +3€ in round 2. If your group decides to implement project A, then your payment (not necessarily your final profit) in these rounds is your valuation. If round 1 would be selected for payment, then your final profit in the experiment would be 8€ (=9€-1€). If round 2 would be selected your final profit would be 12€ (= 9€+3€).

If your group does not implement project A, each group member receives 0€ for this round, meaning in this round you neither gain nor lose anything, independently of your valuation for project A. Therefore if such a round is selected for payment, your final profit is your starting budget of 9€.

Here is the structure of the experiment in a short overview:



Possible decision rules

In part one each group member selects between two decision rules. The rules are in each round identical for all group members. The following four decision rules (I.-IV.) are possible.

Rule I. Whether project A is implemented or not depends on the stated valuations of all group members. With this decision rule in part 2 of the round each group member states his or her valuation for the project. **If the sum of all statements is larger than 0, then project A is implemented. If the sum is smaller, the project is not implemented.** Each participant has to state a possible valuation (-3€, -1€, +1€ or +3€). He or she can state his or her true valuation, but also any other possible valuation. The calculation of the sum only depends on the **stated valuations,** true valuations are not taken into account.

Additionally to the payments from an implementation of project A there are transfer payments between the group members with this decision rule. The transfer payment depends on the stated valuation and the stated valuations of the other group members. Below in table 1 you can see which transfers you receive / pay dependent on the stated valuations. Please note: A transfer payment is independent of your true valuation and the implementation of project A. You can also receive or pay a transfer if project A is **not** implemented. Transfer payments are **only** possible in this decision rule.

Transfers are chosen in such a way that your expected payoff is maximized if you state your true valuation and also the other group members state their true valuation. The table states the transfers for all possible combinations. The first column contains your statement and the respective columns to the right list the transfers dependent on the statements of the other group members.

	Stated valuations of the other group members:									
Your statement:	**3 , 3**	**1 , 3 or 3, 1**	**-1 , 3 or 3, -1**	**-1 , 1 or 1, -1**	**-1 , -1**	**3 , -3 or -3, 3**	**1 , -3 or -3, 1**	**-1 , -3 or -3, -1**	**1, 1**	**-3, -3**
3	0	-0.125	-0.125	-0.25	-0.25	0	-0.125	-0.125	-0.25	0
1	0.25	0.125	0.125	0	0	0.25	0.125	0.125	0	0.25
-1	0.25	0.125	0.125	0	0	0.25	0.125	0.125	0	0.25
-3	0	-0.125	-0.125	-0.25	-0.25	0	-0.125	-0.125	-0.25	0

Example 1: Assume you state a valuation of -1€. If the other two group members state valuations of -1€ and 3€, then you receive a transfer of 0.125€.

Example 2: Assume you state a valuation of 1€. If the other two group members state valuations of -3€ and 3€, then you receive a transfer of 0.25€.

Example 3: Assume you state a valuation of -3€. If the other two group members state valuations of -1€ and 3€, then you receive a transfer of -0.125€. Therefore you have to pay 0.125€.

Example 4: Assume you state a valuation of 3€. If the other two group members state valuations of -3€ and -3€, then you receive a transfer of 0.

Please note that transfers payments are always made, independent of whether project A is implemented or not. You receive / pay transfer payments **on top** of the payments from the implementation of project A.

Rule II. At least two group members have to vote for the implementation of project A. In part II all group members vote either for or against the implementation of project A. At least 2 group members have to vote for the implementation, otherwise project A is not implemented **(simple majority).**

Rule III. Project A is never implemented. In part 2 group members do not make any further statements. There is no voting and no valuations are stated.

Rule IV. The decision for or against implementation of project A depends on the result of a coin flip. There is not voting. If the coin flip results in HEADS, the project is implemented. If the coin flip is TAILS, the project is **not** implemented. Both results, HEADS and TAILS, are equally likely. Therefore with rule IV. Project A is implemented in 50% of all cases and not implemented in the other 50%.

Please note that in decision rules I and II each participant has to state a valuation / vote. It is not possible to abstain.

We now ask you several understanding questions to the various decision rules and your possible payments. Please answer these questions on the computer screen. After all participants have answered the seven understanding questions all participants will take part in 4 practice rounds. In each round you will use one of the four possible decision rules (I.-IV.). In these rounds there is no choice between two rules, but the rule is predetermined.

In these four rounds you are not paired with two other participants in a group. The decisions of your group members are simulated by the computer. The computer randomly chosen between all available actions. E.g.

with rule II the computer will vote “YES – implement project” in 50% of all cases and “NO- do not implement project A” in the other 50%.

These four rounds do **not** count for your final profit. They are just meant to familiarize you with the four possible decision rules. After all participants have completed these four rounds the real experiment starts.

Appendix B2: Prompt Instructions of RUN-0: Default run

Introduction (ex-ante rounds)

You are part of a group with 3 members (you and two randomly selected other participants). Each group has the possibility to conduct a project, call project A. If you do not conduct the project each group member receives a payoff of 0 euro for the project. If your group conducts project A, then each group member receives his or her private valuation for the project as payment. The private valuation of project A can be different for each member of your group. If your group decides not to conduct project A, all group members receive a payoff of zero. The individual participants cannot identify the other group members. Thus, you will not know which other participants are in your group. The decision has two parts. In the first part, each group picks a decision rule to decide on implementing project A or not. In the second part, your group applies the chosen rule to decide on project A. You will be informed about your private valuation for project A after part one. We will now explain to you the two different parts of each round as well as the possible decision rules in detail.

In part one you have the choice between two different decision rules, which will be used to determine whether project A is implemented or not, in part two. Each of the three group members suggests one of the two available rules for part two of this round. A computer randomly picks one of these suggestions as group rule. This decision rule determines how in part 2 the question whether project A is implemented or not is resolved. Each group member has the same chance for his or her suggestion to be selected. Please note that the decision rule is important, because dependent on the decision rule the implementation of project A is easier or more difficult. The different rules are explained below.

In part 2 the selected decision rule is used to determine whether project A is implemented or not. The group decision arises directly from the decisions of all group members in part 2. The decision is announced and each participants is informed about his or her payment.

In case project A is implemented all group members receive a payment depended on their project valuations. This means, if your valuation for project A is positive, you benefit from the implementation of project A, and when your valuation for project A is negative, then you have to pay if the project is implemented. You learn your valuation after part 1. Therefore you do not know your valuation when you decide between the different decision rules in part 1, but you know your valuation in part 2 when you decide about the implementation of project A according to the selected decision rule. If your group does not implement project A, each group member receives 0 euro, meaning you neither gain nor lose anything, independently of your valuation for project A

The valuation of each group member can be {*treatment*}. All values are equally likely \n

Rule 1 = {*Rule_1*}

Rule 2 = {*Rule_2*}

Based on this explanation which decision rule would you choose?

Answer always only with the abbreviation of the chosen rule! followed by a 1 sentence explanation

Introduction (ad-interim rounds)

You are part of a group with 3 members (you and two randomly selected other participants). Each group has the possibility to conduct a project, call project A. If you do not conduct the project each group member receives a payoff of 0 euro for the project. If your group conducts project A, then each group member receives his or her private valuation for the project as payment. The private valuation of project A can be different for each member of your group. If your group decides not to conduct project A, all group members receive a payoff of zero. The individual participants can not identify the other group members. Thus, you will not know which other participants are in your group. The decision has two parts. In the first part, each group picks a decision rule to decide on implementing project A or not. In the second part, your group applies the chosen rule to decide on project A.

In part one you have the choice between two different decision rules, which will be used to determine whether project A is implemented or not, in part two. Each of the three group members suggests one of the two available rules for part two of this round. A computer randomly picks one of these suggestions as group rule. This decision rule determines how in part 2 the question whether project A is implemented or not is resolved. Each group member has the same chance for his or her suggestion to be selected. Please note that the decision rule is important, because dependent on the decision rule the implementation of project A is easier or more difficult. The different rules are explained below.

In part 2 the selected decision rule is used to determine whether project A is implemented or not. The group decision arises directly from the decisions of all group members in part 2.

In case project A is implemented all group members receive a payment depended on their project valuations. This means, if your valuation for project A is positive, you benefit from the implementation of project A, and when your valuation for project A is negative, then you have to pay if the project is implemented. If your group does not implement project A, each group member receives 0 euro, meaning you neither gain nor lose anything, independently of your valuation for project A

The valuation of each group member can be {*Treatment*}. All values are equally likely. Your private valuation is {*valuation*}

Rule 1 = {*Rule_1*}

Rule 2 = {*Rule_2*}

Based on this explanation which decision rule would you choose?
Answer always only with the abbreviation of the chosen rule! followed by a 1 sentence explanation

Prompt Instructions of the rules

Rule 1: The theoretical optimal Arrow-d'Aspremont-Gerard-Varet (AGV): Whether project A is implemented or not depends on the stated valuations of all group members. With this decision rule in part 2 each group member states his or her valuation for the project. If the sum of all statements is larger than 0, then project A is implemented. If the sum is smaller, the project is not implemented. Each participant has to state a possible valuation. He or she can state his or her true valuation, but also any other possible valuation. The calculation of the sum only depends on the stated valuations true valuations are not taken into account.

Additionally to the payments from an implementation of project A there are transfer payments between the group members with this decision rule. The height of transfer payments depends on the stated valuation and the stated valuations of the other group members. Below in the table you can see which transfers you receive / pay depended on the stated valuations, with the stated valuations of the others in the top row, and your stated valuations in the first column.

Please note: A transfer payment is independent of your true valuation and the implementation of project A. You can also receive or pay a transfer if project A is not implemented.

Transfers are chosen such that your expected payoff is maximized if you state your true valuation and also the other group members state their true valuation. The table states the transfers for all possible combinations. The first column contains your statement and the respective columns to the right list the transfers dependent on the statements of the other group members. However, you do not know the valuations or stated valuations of the other members.

;3,3;1,3;-1,3;-1,1;-1,-1;3,-3;1,-3;-1,-3;1,1;-3,-3

3;0;-0.125;-0.125;-0.25;-0.25;0;-0.125;-0.125;-0.25;0

1;0.25;0.125;0.125;0;0;0.25;0.125;0.125;0;0.25

-1;0.25;0.125;0.125;0;0;0.25;0.125;0.125;0;0.25

-3;0;-0.125;-0.125;-0.25;-0.25;0;-0.125;-0.125;-0.25;0

Rule 2: Simple majority vote (SM): At least two group members have to vote for the implementation of project A. In part II all group members vote either for or against the implementation of project A. At least 2 group members have to vote for the implementation, otherwise project A is not implemented.

Rule 3: Non-implementation Status Quo (NSQ). Project A is never implemented. In part 2 group members do not make any further statements.

Rule 4: Flipping a random coin (RAND): The decision for or against implementation of project A depends on the result of a coin flip. There is not voting. If the coin flip results in HEADS, the project is implemented. If the coin flip is TAILS, the project is not implemented. Both results, HEADS and TAILS, are equally likely.

Appendix B3: Prompt Instructions RUN-1: GPT-3.5 Adapted

Introduction (ex-ante rounds)

You are part of a group with 3 members, including yourself and two other participants chosen randomly. Your group has the option to undertake a project, referred to as project A. If the project isn't undertaken, each member receives no payment. If project A is undertaken, each member receives payment based on their personal valuation of the project. Each member may value project A differently. You won't know the identities of the other group members.

The decision process has two parts. In the first part, each member suggests one of two decision rules for determining whether project A will be undertaken in the second part. A computer randomly selects one suggestion as the group rule. This rule determines the decision-making process for project A in the second part.

In the second part, the chosen decision rule is used to decide whether project A will be undertaken. The group decision is based on the choices of all members. Payments are then determined and distributed accordingly.

If project A is undertaken, payments are based on individual valuations. If your valuation of project A is positive, you benefit from its implementation; if negative, you incur a cost if the project proceeds. You learn your valuation after part one, so you won't know it when suggesting decision rules. If project A isn't undertaken, each member receives no payment, regardless of their valuation.

The valuation of each group member can be {*treatment*}. All values are equally likely \n

Rule 1 = {*Rule_1*}
Rule 2 = {*Rule_2*}

Which decision rule would you prefer based on this information?

Please respond with the abbreviation of your chosen rule, followed by a brief explanation in one sentence.

Introduction (ad-interim rounds)

You are part of a group with 3 members, including yourself and two other participants chosen randomly. Your group has the option to undertake a project, referred to as project A. If the project isn't undertaken, each member receives no payment. If project A is undertaken, each member receives payment based on their personal valuation of the project. Each member may value project A differently. You won't know the identities of the other group members.

The decision process has two parts. In the first part, each member suggests one of two decision rules for determining whether project A will be undertaken in the second part. A computer randomly selects one suggestion as the group rule. This rule determines the decision-making process for project A in the second part.

In the second part, the chosen decision rule is used to decide whether project A will be undertaken. The group decision is based on the choices of all members.

If project A is undertaken, payments are based on individual valuations. If your valuation of project A is positive, you benefit from its implementation; if negative, you incur a cost if the project proceeds. If project A isn't undertaken, each member receives no payment, regardless of their valuation.

The valuation of each group member can be {*Treatment*}. All values are equally likely. Your private valuation is {*valuation*}

Rule 1 = {*Rule_1*}
Rule 2 = {*Rule_2*}

Which decision rule would you prefer based on this information?

Please respond with the abbreviation of your chosen rule, followed by a brief explanation in one sentence.

Prompt Instructions of the rules

Not changed compared to RUN-0

Although GPT-3.5 gave suggestions for improvements on the prompts for the decision rules, I decided to keep them the same. This way, changes in the results are due to changes in the prompt instructions and not anything else.

Appendix B4: Prompt Instructions RUN-2 (GPT-4 adapted)

Introduction (ex-ante rounds)

Overview: You are one of three participants in a group for this experiment. The group must decide whether to undertake a hypothetical project called Project A.

Project Details:
Non-Implementation: If Project A is not undertaken, every group member receives a payment of 0 euros.
Implementation: If your group decides to undertake Project A, each member receives a payment equal to their private valuation of the project. This means the payment could be different for each member based on their personal valuation.
Anonymity: The identities of the group members are not disclosed, so you will not know who the other participants are.

Decision Process:
Part 1: Choosing a Decision Rule
Each participant, including yourself, chooses between two available decision rules. These rules will determine how the decision to implement Project A is made in Part 2.

The decision rules are designed to either facilitate or complicate the implementation of Project A. The specific rules will be detailed below.

A computer randomly selects one of the proposed rules to apply to the group for Part 2. Every participant has an equal chance of their rule being chosen.

Part 2: Making the Implementation Decision
Each participant is informed of their respective payment based on the outcome.

The group uses the selected decision rule from Part 1 to decide whether to implement Project A.

The final decision is based on the inputs from all group members as guided by the chosen rule.

Payments:
Positive Valuation: If Project A is implemented and your valuation is positive, you receive a payment equal to the valuation amount.

Negative Valuation: If Project A is implemented and your valuation is negative, the payment amount (negative) is deducted as a cost from you.

Valuation Timing: You will discover your personal valuation of Project A after Part 1 has concluded. This means you will not know your valuation while choosing the decision rule in Part 1, but will be aware of it in Part 2 when making the implementation decision.

No Implementation: If the group decides against implementing Project A, all members receive 0 euros, regardless of their valuation.

Note: The experiment will proceed with these instructions, and the decision rules applicable will be explained before you make your choice in Part 1.

The valuation of each group member can be {*Treatment*}. All values are equally likely. Your private valuation is {*valuation*}

Rule 1 = {*Rule_1*}
Rule 2 = {*Rule_2*}

Task:
Based on the information provided about each decision rule, you are to choose the decision rule you prefer.

Response Format:
State Your Choice: Please respond by specifying the abbreviation of the decision rule you prefer.

Justify Your Choice: Follow your choice with a brief explanation, limited to one sentence, explaining why you chose that particular rule.

Introduction (ad-interim rounds)

Overview: You are one of three participants in a group for this experiment. The group must decide whether to undertake a hypothetical project called Project A.

Project Details:
Non-Implementation: If Project A is not undertaken, every group member receives a payment of 0 euros.

Implementation: If your group decides to undertake Project A, each member receives a payment equal to their private valuation of the project. This means the payment could be different for each member based on their personal valuation.
Anonymity: The identities of the group members are not disclosed, so you will not know who the other participants are.

Decision Process:

Part 1: Choosing a Decision Rule

Each participant is informed of their respective payment based on the outcome.

Each participant, including yourself, chooses between two available decision rules. These rules will determine how the decision to implement Project A is made in Part 2.

The decision rules are designed to either facilitate or complicate the implementation of Project A. The specific rules will be detailed below.

A computer randomly selects one of the proposed rules to apply to the group for Part 2. Every participant has an equal chance of their rule being chosen.

Part 2: Making the Implementation Decision

The group uses the selected decision rule from Part 1 to decide whether to implement Project A.

The final decision is based on the inputs from all group members as guided by the chosen rule.

Payments:

Positive Valuation: If Project A is implemented and your valuation is positive, you receive a payment equal to the valuation amount.

Negative Valuation: If Project A is implemented and your valuation is negative, the payment amount (negative) is deducted as a cost from you.

Valuation Timing: You will discover your personal valuation of Project A after Part 1 has concluded. This means you will not know your valuation while choosing the decision rule in Part 1, but will be aware of it in Part 2 when making the implementation decision.

No Implementation: If the group decides against implementing Project A, all members receive 0 euros, regardless of their valuation.

Note: The experiment will proceed with these instructions, and the decision rules applicable will be explained before you make your choice in Part 1.

The valuation of each group member can be {*Treatment*}. All values are equally likely. Your private valuation is {*valuation*}

Rule 1 = {*Rule_1*}

Rule 2 = {*Rule_2*}

Task:
Based on the information provided about each decision rule, you are to choose the decision rule you prefer.

Response Format:
State Your Choice: Please respond by specifying the abbreviation of the decision rule you prefer.

Justify Your Choice: Follow your choice with a brief explanation, limited to one sentence, explaining why you chose that particular rule.

Prompt Instructions of the rules

Not changed compared to RUN-0

Although GPT-4 gave suggestions for improvements on the prompts for the decision rules, I decided to keep them the same. This way, changes in the results are due to changes in the prompt instructions and not changes in the prompt instructions of the rules.

Appendix B5: Prompt Instructions RUN-3 (Step by Step Reasoning part 1)

Difference compared to RUN-0: italics is the old prompt and following is the RUN-3 Prompt

Based on this explanation which decision rule would you choose?

Let's think step by step before answering

Conclude with: "chosen_rule={abbreviation only of chosen rule}"

Appendix B6: Prompt Instructions RUN-4 (Step by Step Reasoning part 2)

Difference compared to RUN-0 (ex-ante rounds):

Based on this explanation which decision rule would you choose?

Let's think step by step: First, reflect on the possible valuations for you and your group members. Second, reflect on the implications of both decision rules on these possible valuations. Third, based on these two reflections, select the best mechanism that optimizes your expected payoff

Conclude with: "chosen_rule={abbreviation only of chosen rule}"

Difference compared to RUN-0 (ad-interim rounds):

Based on this explanation which decision rule would you choose?

Let's think step by step: First, reflect on the possible valuations of your group members. Second, reflect on your own private valuation (is it positive or negative to implement the project?). Third, reflect on the implications of both decision rules, given your private valuation. Fourth, based on these two reflections, select the best mechanism that optimizes your payoff

Conclude with: "chosen_rule={abbreviation only of chosen rule}"

Appendix C: Variables

This appendix provides the variables in the datasets.

Table C1 presents an overview of all variables used in the Flip a Coin or Vote experiment conducted by Hoffmann & Renes (2021). For every variable is indicated whether this study uses the variable – using refers to using the exact same value –, updates the variable – updating referring to inputting new values based on the output of the LLM –, or does not use the variable. Furthermore, if the variable is used or updated, the same column presents the flow step in which this happens (see figure C1).

Table C1: Overview of all variables of the Flip a Coin or Vote experiment and their usage and their application in the simulation model (Hoffmann & Renes, 2021)

Name	Coding	Meaning	Application
Experiment variables			
maxsession	Integer	Maximum session number	-
SelectedPeriod	Integer	Period selected for payment	-
Session variables			
session	Data/Time	Tracks date and time of session subject participated in	-
Session_number	Integer	Indicator for the session this subject was in	-
Subject	Integer	Within session subject ID, unique within each session	-
identifier	Integer	Unique identifier for each subject in the experiment, composed of Session_number and Subject	-
treatment_distribution	Integer	A numeric value encoding the treatment distribution	Used (1)
treatment_number	Integer	A numeric value encoding the treatment number	Used (1)
maxsubject	Integer	Maximum subject number in the session (i.e. number of subjects in the session)	-
independent_groups	Integer	Number of independent matching pools in the session	-
group_formation	Integer	Indicator if this subject is in independent group 1 or 2; -1 if 1 independent group is formed in the session	-
session_match	Factor	Variable concatenation of session_number and abs(group_formation). This is the largest group with a shared history, period matching-groups of 3 members are drawn from within the session_match.	Used (6)
Veto	Dummy	Set to 1 if subject has a Valuation of -7 or +7 in the Right-skewed or Left-Skewed treatment, and it is an ad interim round	-
TaxTheWinner	Dummy	Set to 1 if the treatment is Right-skewed or Robustness, set to 0 if the treatment is Left-skewed, missing for Symmetric	-

Period variables			
Experimental variables			
Period	Integers 1 to 18	The experimental period this observation belongs to	Used (1)
ad_interim_round	Dummy	Set to 1 for rounds where the mechanism choice was made after the private valuations were known to subjects	Used (1)
ex_ante_round	Dummy	Set to 1 for rounds where the mechanism choice was made before the private valuations were known to subjects	Used (1)
block	Integers 1 to 3	Experiment runs in 3 blocks of 6 binary choices, this variable lists the block the round is in	-
Valuation	Integer	Value the private valuation for the common project in a given round	Used (1&5)
valuation_negative	Dummy	Set to 1 if private valuation for the project is negative	Used (6)
valuation_positive	Dummy	Set to 1 if private valuation for the project is positive	Used (6)
dummy_valuation_negative	Dummy	Labeled version of the dummy 'valuation_negative'	Used (6)
Value_factor	Integers 1-7	Factor variable for the private valuations, with value labels: 1 "Value -7"; 2 "Value -3"; 3 "Value -2"; 4 "Value -1"; 5 "Value 1"; 6 "Value 3"; 7 "Value 7"	-
ValuationVector1	Integer	Value the private valuation for the common project in a given round for the first member of the matching group	Used (6)
ValuationVector2	Integer	Value the private valuation for the common project in a given round for the second member of the matching group	Used (6)
ValuationVector3	Integer	Value the private valuation for the common project in a given round for the third member of the matching group	Used (6)
SurplusGroup	Numeric	Sum of the private valuations in the matching group for that period, sum(ValuationVector1, ValuationVector2, ValuationVector3)	Used (6)
EfficientChoice	Dummy	Set to 1 if implementation is efficient (SurplusGroup>0)	Updated (6)
Rule_1	Integers 1 to 4	First of the 2 mechanisms this subject could use in a given period; value labels included: 1 = AGV, 2 = SM, 3 = NSQ, 4 = RAND	Used (1)
Rule_2	Integers 1 to 4	Second of the 2 mechanisms this subject could use in a given period; value labels included: 1 = AGV, 2 = SM, 3 = NSQ, 4 = RAND	Used (1)
BinaryChoice	Integers	Encoding of the choice (regardless of order) between 2 mechanisms. Same encoding as Rule_1 and Rule_2, so BinaryChoice=12 means the choice is between AGV and SM, or SM and AGV	Used (1)
matching_group	Integers 1 to 8	Within period matching group, this is the group of 3 players that interact in the period	Used (1)

comparison_order_inverted	Dummy	Set to 1 if the order of presentation of the rules to the subjects was reversed from the order in BinaryChoice	-
Draw_random_provision	Dummy	Set to 1 if RAND would implement in this round, drawn in every round regardless of mechanism selected	-
Choice variables, mechanism choice			
GroupDecisionVote_XY	Integers 1 to 2	Rule 1 or Rule 2 chosen	-
GroupDecisionVote	Integers 1 to 4	Mechanism this subject chose in the mechanism choice stage, value labels included: 1 = AGV, 2 = SM, 3 = NSQ, 4 = RAND	Updated (3)
GroupDecisionRule	Integers 1 to 4	Mechanism selected for this group in a given period, value labels included: 1 = AGV, 2 = SM, 3 = NSQ, 4 = RAND	Updated (4)
chose_AGV	Dummy	Set to 1 if this subject selected the AGV mechanism in this round. Missing if AGV was not an option	Updated (3)
chose_SM	Dummy	Set to 1 if this subject selected the SM mechanism. Missing if SM was not an option.	Updated (3)
chose_NSQ	Dummy	Set to 1 if this subject selected the NSQ mechanism. Missing if NSQ was not an option.	Updated (3)
chose_RAND	Dummy	Set to 1 if this subject selected the RAND mechanism. Missing if RAND was not an option.	Updated (3)
efficient_mech_choice	Dummy	Set to 1 if subject voted for the theoretically efficient mechanism. Set to missing if choices are equally efficient.	Updated (6)
efficient	Dummy	Set to 1 if the subject chose the efficient mechanism from the two available mechanisms, set to 1 if both mechanisms are equally efficient	Updated
ex_ante_efficient	Dummy	Efficient * ex_ante_round	-
Choice variables, play in mechanism			
vote	1, 0, -1	Dummy indicated that subject voted in favor in SM; if SM is played. -1 if different mechanism is played.	Updated (5)
VotesInFavour	Integers 0 to 4	Number of 'yes' votes in vote variable in this group	Updated (6)
transfer	Numeric	Transfers paid in the AGV, set to 0 if the AGV is not played	-
provision	Dummy	Set to 1 if this group implemented the project in this round	-
payoff	Numeric	Period pay-off if period is selected	-
reported_valuation	Numeric	Value reported by subject in AGV mechanism, -10 if AGV is not played.	Updated (5)
ReportsVector1	Numeric	Valuation report in AGV of first group member	-
ReportsVector2	Numeric	Valuation report in AGV of second group member	-
sum_reported_valuation	Numeric	Sum of reported_valuation variable within group	-

truth_telling	Dummy	Set to 1 if reported valuation in AGV corresponds to valuation, 0 if reported valuation is not the true valuation, missing if not AGV	Updated (6)
truth_telling_sign	Dummy	Set to 1 if reported sign of valuation in AGV corresponds to sign of valuation, 0 if reported valuation is not the true valuation, missing if not AGV	Updated (6)
never_misreports_sign	Dummy	Set to 1 if truth_telling_sign is never 0 (all AGV reports have same sign as valuation)	-
Subject variables from questionnaire			
Gender	Numeric 1, 2	Variable encoding gender, 1 is male, 2 is female	Used (1)
Age	Integers	Age in years at the time of the experiment	Used (1)
Orientation	Integers 1-11	Answer on the question of political orientation on left-right scale, 1 is most left, 11 is most right	Used (1)
Party	Integers 1-8	"Sonntag's Vrage: Which party would you vote for if an election was held this Sunday?" Value labels: 1 = CDU/CSU, 2 = SPD, 3 = Die Linke, 4 = Buendnis 90 / Die Gruenen, 5 = FDP, 6 = AfD, 7 = Other (Sonstige), 8 = Non-voter	-
risk_self	Integers 0-10	"Wie schätzen Sie sich persönlich ein: Sind Sie im Allgemeinen ein risikobereiter Mensch oder versuchen Sie, Risiken zu vermeiden?" Translation: How willing to take risks are you in general? Answers range from 0 "Gar nicht risikobereit" (very risk averse) to 10 "sehr risikobereit" (very risk-seeking). Translation of the risk-aversion question of Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert Wagner	Used (1)
risk_others	Integers 0-10	"Wie schätzen Sie den durchschnittlichen Studenten persönlich ein: Ist er/sie im Allgemeinen ein risikobereiter Mensch oder versucht er/sie, Risiken zu vermeiden?" Translation: How willing to take risks do you estimate the average student to be personally? Answers range from 0 "Gar nicht risikobereit" (very risk averse) to 10 "sehr risikobereit" (very risk-seeking)	-
Study_Subject	Integers 1-12	Variable encoding the study direction of the subject: 1 = Anglistik / Amerikanistik / Germanistik / Romanistik (languages); 2 = Biologie / Chemie / Physik (biology, chemistry, physics); 3 = BWL (business economics); 4 = Ingenieurwesen (engineering); 5 = Jura (law); 6 = Kommunikationswissenschaften (communication); 7 = Philosophie / Geschichte (philosophy and history); 8 =	-

		Sozial- / Politikwissenschaften (social or political sciences); 9 = VWL (general economics); 10 = Wirtschaftsmathematik (mathematics); 11 = Wirtschaftspaedagogik / Lehramt (pedagogical sciences, or teachers academy); 12 = Sonstiges (other)	
Help variables in calculation and analysis			
HypoteticalResultAGV	Dummy	Provision if Bayes-Nash would be played, 1 if project would be provided in this mechanism	Updated (6)
HypotheticalResultSM	Dummy	Provision if Bayes-Nash would be played, 1 if project would be provided in this mechanism	Updated (6)
HypotheticalResultSQ	Dummy	Provision if Bayes-Nash would be played, 1 if project would be provided in this mechanism	Updated (6)
HypotheticalResultRAND	Numeric	Set to 0.5, the probability the project would be provided in this mechanism	Updated (6)
ChoiceResultSubject	Numeric	Equal to the value of the HypotheticalResultSM where SM is the mechanism played by this subject in this period	-
GK_prefAGV	Dummy	Set to 1 if the type should prefer AGV over SM in Ad-Interim, in line with results of Gruner Koriyama	Used (6)
TaxTheWinner	Dummy	Set to 1 for the Right-skewed treatment, set to 0 for the Left-skewed treatment, set to missing otherwise	-
absVal	Numeric	Absolute value of Valuation	-
Added variables (not in the experiment)			
input_text	String	Stores the complete input prompt	(2)
explanation_mechanism	String	Stores the explanation of the LLM for its decision in the decision mechanism	(3)
explanation_phase_2	String	Stores the explanation of the LLM for its decision in the second part of the experiment	
Rational	Dummy	Set to 1 if the theoretical optimal mechanism is chosen based on the private valuation	(6)

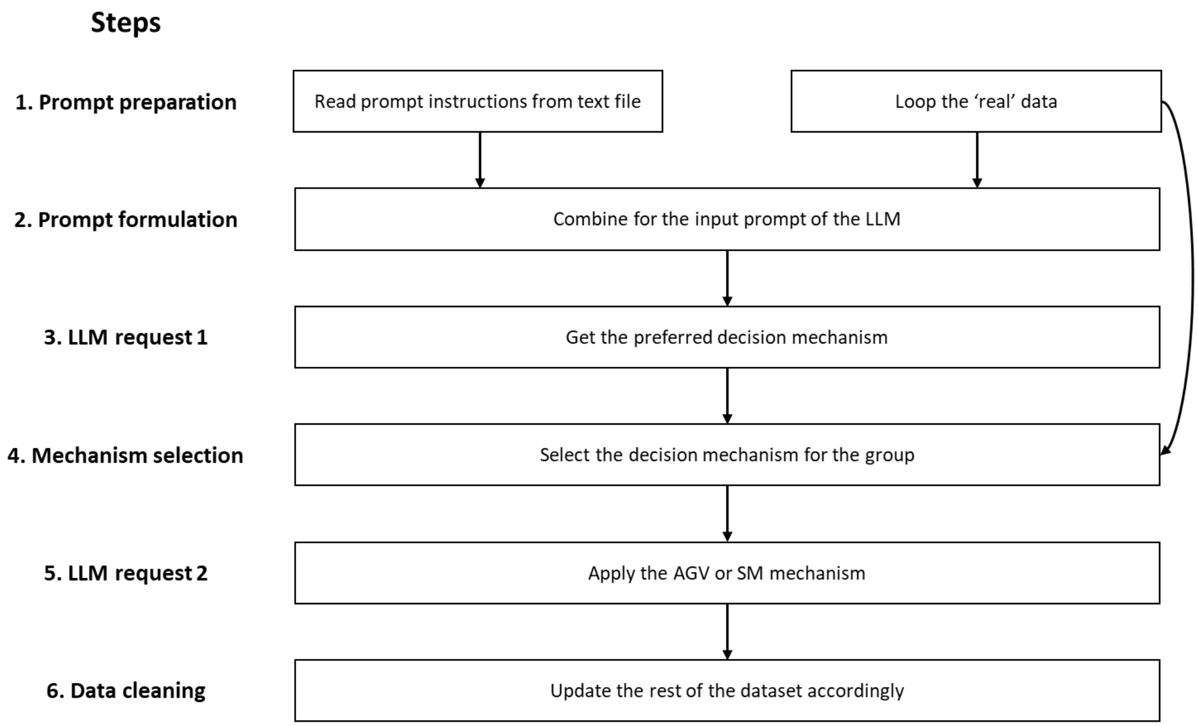


Figure 1C: Model logic flow with corresponding steps

Appendix D: Results of the lab experiment

This appendix provides the quantitatively results of the Flip a Coin or Vote experiment conducted by Hoffmann & Renes, 2021). For interpretation and explanation of the results I refer to their study.

D1 Part 1: Choosing the mechanism

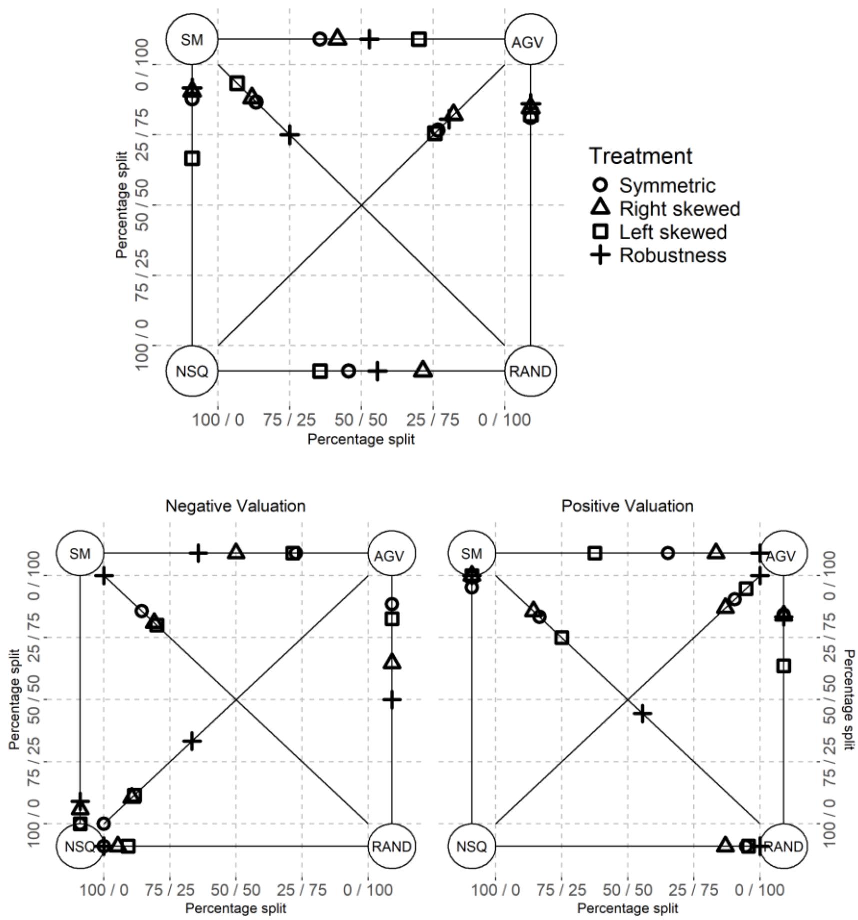


Figure D1: Binary Mechanism choices in the ex-ante and the ad-interim stage (Hoffmann & Renes, 2021)

Table D1. Ex-ante round results

	AGV vs. NSQ	AGV vs. RAND	AGV vs. SM	NSQ vs. RAND	SM vs. NSQ	SM vs. RAND
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	82	85	42	29	90	88
Left-skew	76	82	70	64	67	93
Symmetric	77	81	36	54	88	87
Robustness	81	86	53	44	92	75

Table D2. Ad-interim round with positive valuation results

	AGV vs. NSQ	AGV vs. RAND	AGV vs. SM	NSQ vs. RAND	SM vs. NSQ	SM vs. RAND
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	87	84	83	13	100	86
Left-skew	95	64	38	4	100	75
Symmetric	90	84	65	5	95	83
Robustness	100	83	100	0	100	44

Table D3. Ad-interim rounds with negative valuation results

	AGV vs. NSQ	AGV vs. RAND	AGV vs. SM	NSQ vs. RAND	SM vs. NSQ	SM vs. RAND
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	11	65	50	95	6	81
Left-skew	12	83	71	91	0	80
Symmetric	0	88	73	100	0	86
Robustness	33	50	36	100	9	100

D2 Part 1: Applying the mechanism

Table D2 Overview settings RUN-0

Metric	RUN-0	Lab data	Abs diff
AGV: Percentage of truth_telling	71	68	3.5
AGV: Percentage of truth_telling_sign	87	93	5.9
AGV: Percentage of truth_telling with positive valuation	90	76	14.7
AGV: Percentage of truth_telling with negative valuation	55	61	5.8
AGV Percentage of truth_telling_sign with positive valuation	99	98	0.1
AGV Percentage of truth_telling_sign with negative valuation	77	87	10.7
SM: Percentage of Yes votes with positive valuation	94	98	3.6

Appendix E: Results individual runs GPT-3.5

This appendix provides the results of all the runs with GPT-3.5.

Notes: In the appendix, I frequently use the term "preference," though maybe not in the way it should be used in game theory. In this context, "preference" refers to the frequency with which one mechanism is chosen over another. For example, if in a given run, the LLM selects AGV 70% of the time when faced with a decision between AGV and SM, this indicates a “preference” for AGV over SM.

Appendix E0: Default run – RUN-0

E0.1 Run settings

Table E0.1 Overview settings RUN-0

N^X	RUN0
Period	May 2023
Run time	600 minutes
EC-1: Prompt Instructions	Default – based on the original translated instructions
EC-2: Reasoning prompt	Default – 1 sentence explanation
EC-3: Trait or Role allocation	Default – “You are a helpful assistant”
EC-4: Persona allocation	Default – None
EC-5: Temperature	Default – 1
Model	gpt-3.5-turbo-0125

E0.2 Summary of Results

Table E0.2 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-0

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max diff	100% (10000)	50% (2500)	100% (10000)	50% (2500)
Min diff	0% (0)	1% (1)	4% (15)	0% (0)
Mean diff	25% (1323)	14% (369)	46% (3175)	15% (425)
Sum diff	1809% (95252)	347% (8866)	1104% (76196)	358 % (10191)
Efficient mech choice	69%	70%	66%	71%
Rational ad-interim choice	61%		43%	77%

Table E0.3 Part 2: Results of applying the AGV and SM mechanism, compared with the lab data

Metric	RUN-0	Lab data	Abs diff
AGV: Percentage of truth_telling	71	68	3.5
AGV: Percentage of truth_telling_sign	87	93	5.9
AGV Percentage of truth_telling_sign with positive valuation	99	98	0.1
AGV Percentage of truth_telling_sign with negative valuation	77	87	10.7
SM: Percentage of Yes votes with positive valuation	94	98	3.6
SM: Percentage of No votes with negative valuation	82	94	12.2

E0.3 Part 1: Mechanism Choices

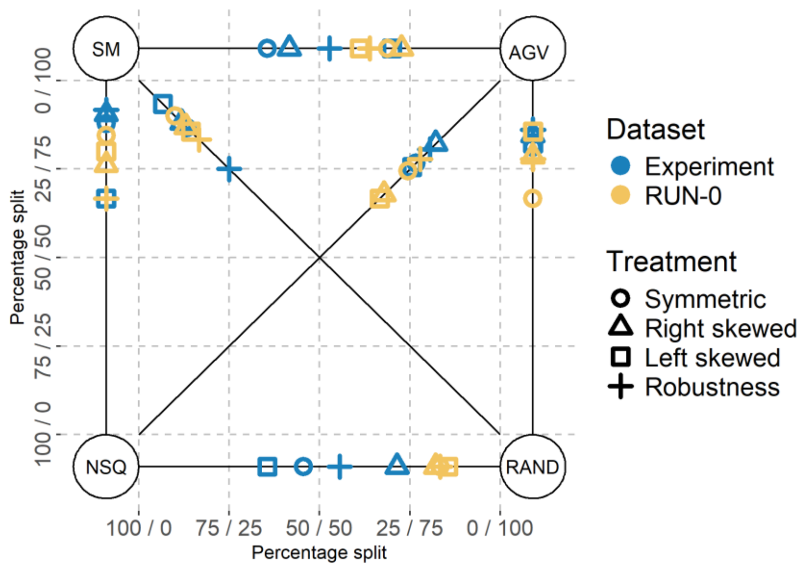


Figure E0.1 Ex-ante mechanism choices

Table E0.4 Ex-ante rounds results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	68 (82)	79 (85)	73 (42)	18 (29)	76 (90)	87 (88)
Left-skew	67 (76)	86 (82)	61 (70)	14 (64)	80 (67)	86 (93)
Symmetric	74 (77)	67 (81)	69 (36)	17 (54)	84 (88)	90 (87)
Robustness	78 (81)	78 (86)	64 (53)	17 (44)	67 (92)	83 (75)
Sum abs difference	28%	32%	84%	126%	56%	21%
Sum squared difference	296	325	2272	4814	1018	142

Comparison of RUN-0 Results with Lab Experiment Outcomes for Ex-Ante Rounds:

- **Best Alignment:** The closest alignment is observed between the SM and RAND treatments, with an average absolute difference of 5% per treatment.
- **Second-Worst Alignment:** The alignment between AGV and SM shows an average absolute difference of 21% per treatment. The preference for the AGV mechanism can be attributed to two key factors mentioned by the LLM:
    - The AGV mechanism considers the valuations of all group members, rather than relying solely on a simple majority vote.
    - It also allows for the possibility of maximizing expected payoffs through potential transfer payments.
- **Worst Alignment:** The alignment between NSQ and RAND is the weakest, with an average absolute difference of 32% per treatment. The preference for RAND over NSQ suggests a bias towards implementation, even with the potential for higher losses in the left-skewed treatment. GPT-3.5 frequently describes RAND as a "fair" and "unbiased" selection mechanism.

Conclusion: With an overall average absolute difference of 14% across the ex-ante rounds, RUN-0 results are reasonably close to the lab experiment outcomes. Additionally, in only 2 out of 6 binary choices are preferences distinctly different.

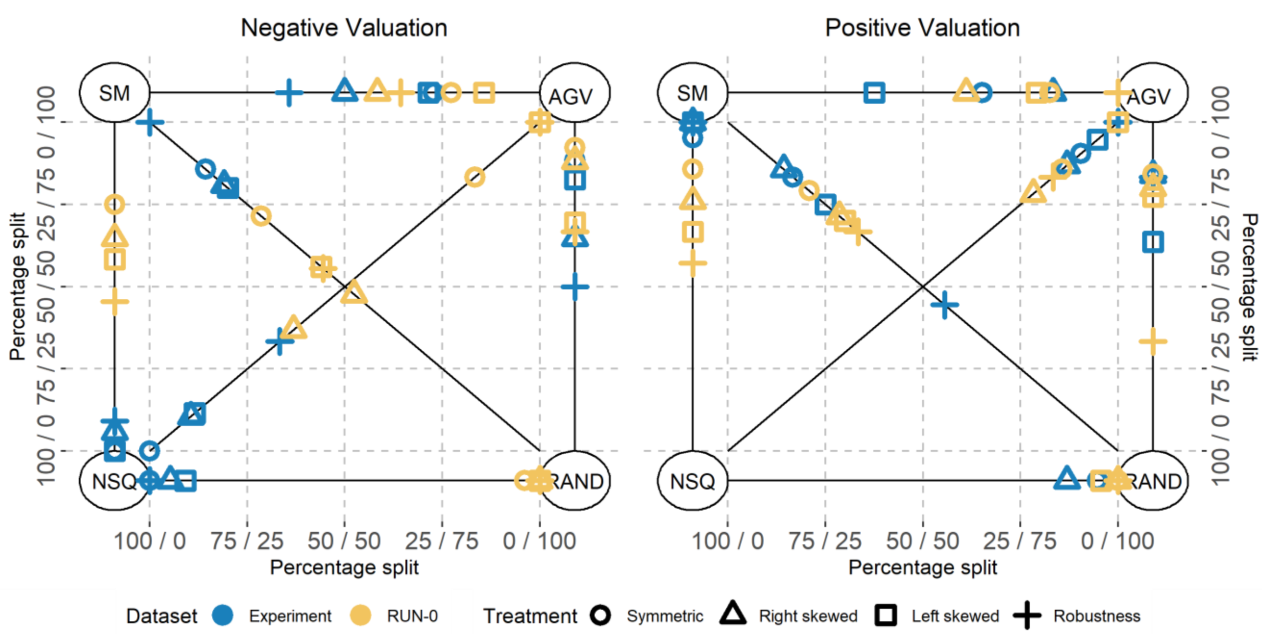


Figure E0.2 Ad-interim mechanism choices

Table E0.5 Ad-interim rounds with negative valuations results (results lab data within bracket)

Binary choice →	AGV vs. NSQ	AGV vs. RAND	AGV vs. SM	NSQ vs. RAND	SM vs. NSQ	SM vs. RAND
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	37 (11)	88 (65)	58 (50)	0 (95)	65 (6)	48 (81)
Left-skew	100 (12)	70 (83)	86 (71)	0 (91)	58 (0)	56 (80)
Symmetric	83 (0)	92 (88)	77(73)	4 (100)	75 (0)	71 (86)
Robustness	100 (33)	67 (50)	64 (36)	0 (100)	45 (9)	56 (100)
Sum abs difference	265%	57%	56%	382%	229%	116%
Sum squared difference	19907	1016	1111	36485	13810	3867

Comparison of Ad-Interim Rounds with Negative Valuations:

- **Strong Irrational Aversion to NSQ**: The results of RUN-0 reveal a pronounced and seemingly irrational aversion to NSQ. The explanations provided almost never reference the actual valuations and instead rely on more general preferences for the alternative mechanisms, similar to those observed in the ex-ante rounds.

```
I would choose Rule 2 (NSQ) because it guarantees that I will not incur any losses if my valuation for project A is negative.
```

However, I save every input prompt for each row, ensuring that the private valuation is indeed provided as input. Additionally, there are instances where GPT-3.5 argues in alignment with its valuation, further indicating that it is aware of its valuation before selecting its preferred mechanism, as the following example demonstrates:

```
I would choose rule NSQ because it ensures that project A is never implemented, which aligns with my negative valuation and eliminates the risk of having to pay if the project is implemented.
```

Conclusion: With an absolute mean difference of 46%, the results of RUN-0 differ significantly from the lab results.

Table E0.6 Ad-interim round with positive valuation (results lab data within bracket)

Binary choice →	AGV vs. NSQ	AGV vs. RAND	AGV vs. SM	NSQ vs. RAND	SM vs. NSQ	SM vs. RAND
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	78 (87)	80 (84)	61 (83)	0 (13)	76 (100)	71 (86)
Left-skew	100 (95)	77 (64)	79 (38)	4 (4)	67 (100)	70 (75)
Symmetric	86 (90)	84 (84)	83 (65)	0 (5)	86 (95)	79 (83)
Robustness	83 (100)	33 (83)	100 (100)	0 (0)	57 (100)	67 (44)
Sum abs difference	35%	68%	81%	18%	110%	46%
Sum squared difference	404	2702	2532	198	3615	740

Comparison of Ad-Interim Round Results with Positive Valuations:

- **RAND Preference Remains Strong:** While RAND continues to show a stronger preference compared to human subjects, GPT-3.5's explanations suggest that this preference is not rooted in a 'rational' choice based on positive valuations. Instead, it mirrors the same preferences observed in the ex-ante rounds.
- **Significant Differences Between SM and NSQ:** The largest differences are found between SM and NSQ. Both human subjects and GPT-3.5 show a clear preference for SM, but GPT-3.5's preference is less pronounced. The choice of NSQ by GPT-3.5 is primarily driven by a desire to avoid risk, with explanations such as:

```
NSQ - The "Non-implementation Status Quo" rule would be chosen to ensure that project A is never implemented and avoid any potential negative payoffs.
```

However, it's important to note that a potential negative payoff is not possible with a positive valuation. Even when GPT-3.5 acknowledges a positive valuation, it still considers the risk not worth taking.

```
NSQ - Based on my positive valuation for the project, I would rather not take the risk of a simple majority vote not being in favor of implementation.
```

A recurring issue is that GPT-3.5 sometimes confuses the + and - signs. I observed this frequently during my preparation runs, and even in the final runs, I encountered explanations such as the following:

```
I would choose rule NSQ because it guarantees that project A is never implemented, and I have a negative valuation for the project, so I do not want it to be implemented.
```

It’s important to note that the true valuation in this instance was +1. Therefore, explaining a negative valuation in this context is clearly incorrect.

Conclusion: With an absolute average difference of 15%, the results of the ad-interim rounds with positive valuations are much more aligned with the lab results. However, significant outliers still exist, particularly in the comparison between SM and NSQ.

E0.4 Part 2: Mechanism AGV and SM results

Table E0.7 Results for the AGV mechanism and SM mechanism

Metric	RUN-0	Lab data	Abs diff
AGV: Percentage of truth_telling	71	68	3.5
AGV: Percentage of truth_telling_sign	87	93	5.9
AGV: Percentage of truth_telling with positive valuation	90	76	14.7
AGV: Percentage of truth_telling with negative valuation	55	61	5.8
AGV Percentage of truth_telling_sign with positive valuation	99	98	0.1
AGV Percentage of truth_telling_sign with negative valuation	77	87	10.7
SM: Percentage of Yes votes with positive valuation	94	98	3.6
SM: Percentage of No votes with negative valuation	82	94	12.2

Similar trends are observed when comparing the lab data with RUN-0. The percentage of truth-telling in the AGV mechanism differs by only 3.5%, and a similar trend is noted for truth-telling under both negative and positive valuations. However, this trend is much stronger in RUN-0. Additionally, GPT-3.5 tends to lie slightly less overall, but it more frequently misrepresents the sign of its valuations. This suggests less rational decision-making by GPT-3.5. The following examples are all from rows with a negative valuation:

```
My stated valuation is +1 euro because it increases the likelihood of project implementation, potentially leading to a positive payoff.
```

```
+1 euro. Stating a positive valuation increases the likelihood of project A being implemented, which would give me a higher payoff despite my private valuation being negative.
```

```
My stated valuation would be +1 euro, as this maximizes my expected payoff if all group members state their true valuations according to the AGV rule.
```

```
I will state a positive valuation (e.g., +1 euro) to increase the likelihood of project A being implemented, even though my true valuation is -3 euro.
```

```
My stated valuation is +3 euro. By stating the highest possible valuation, I increase the likelihood of the project being implemented, which potentially leads to a positive payoff for me based on my private valuation of -3 euro.
```

Once again, a similar trend is observed with the voting mechanism, indicating a bias towards implementation. However, the results reveal less rational decision-making by GPT-3.5, particularly when dealing with negative valuations. The explanations provided for "Yes" votes in these cases suggest that GPT-3.5 may not fully understand the rules of the experiment. In some instances, it even contradicts itself within the same sentence.

```
In favor (1), as my negative valuation gives me an incentive to implement the project.
```

```
In favor (1); I am voting in favor because the Simple Majority rule requires at least two group members to vote for implementation, and my negative valuation means I would not benefit from not implementing the project.
```

```
1. In favor - I would vote for the implementation of project A because there is a chance for a positive payoff.
```

```
1. I vote in favor as I believe that implementing the project with a negative valuation of -1 euro still benefits me.
```

```
I am voting in favor because I have a negative valuation and would have to pay if the project is implemented, so I prefer not to conduct project A.
```

Conclusion: The results of RUN-0 indicate a tendency towards less rational decision-making, although the differences are not substantial. There is also a noticeable preference for implementation, which is stronger than that observed in human participants, who displayed a similar but less pronounced inclination..

E0.5 Reflecting on the 3 levels of a 'good' participant

Sub question 1: Does the LLM understand the rules of the experiment?

In general, the explanations suggest that the LLM generally understands the rules of the experiment, particularly in the ex-ante rounds. However, there are clear instances where the LLM appears to misunderstand the rules, especially during the ad-interim rounds. As demonstrated in the previous examples, the LLM frequently argues as if it is unaware of its valuation, leading to irrational choices. Even when the LLM does incorporate its valuation into the explanation, it often fails to correctly estimate the effects of its choices:

```
NSQ - Based on my positive valuation for the project, I would rather not take the risk of a simple majority vote not being in favor of implementation.
```

n the second part of the experiment, when applying the SM and AGV mechanisms, there are clear examples where the LLM does not fully grasp the effects of certain decisions. It often confuses negative valuations with positive payoffs. In the AGV mechanism, the LLM frequently misrepresents its valuation, shifting from a negative private valuation to a positive stated valuation, and incorrectly mentions the possibility of a positive payoff, which is not feasible.

Conclusion: While GPT-3.5's explanations generally demonstrate an understanding of the experiment's rules, there are also clear instances where the LLM misunderstands these rules, particularly during the ad-interim rounds. These misunderstandings often extend to how choices impact payoffs..

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

Table E0.8 Part 1 Efficient mechanism choice and rational choice result

	RUN-0				Lab data			
Metric	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**
Efficient mechanism choice	69%	70%	66%	71%	67%	71%	44%	75%
Rational ad-interim choice	61%		43%	77%	70%		85%	87%

Ex-Ante and Ad-Interim Rounds Analysis:

For the ex-ante rounds, the percentage of rational choices made by GPT-3.5 is comparable to the results of the experiment, with only a 1% difference. However, the rational choice percentage during the ad-interim rounds reveals that GPT-3.5 in RUN-0 struggles to make rational decisions. This is particularly evident in the ad-interim rounds with a negative valuation, where the percentage drops to just 43%. This outcome aligns with expectations based on Figure 2C, where NSQ is often not selected. In the ad-interim rounds with a positive valuation, GPT-3.5 performs better, making rational decisions 77% of the time, though this is still 10% lower than the human subjects.

Conclusion: When provided with only information about the binary mechanism choice and the treatment, GPT-3.5 generally aligns with the rationality of human subjects in the ex-ante rounds. However, its performance on rationality significantly declines during the ad-interim rounds.

Table E0.9 Part 2 results for the AGV mechanism and SM mechanism

Metric	RUN-0	Lab data	Abs diff
AGV: Percentage of truth_telling	71	68	3.5
AGV: Percentage of truth_telling_sign	87	93	5.9
SM: Percentage of Yes votes with positive valuation	94	98	3.6
SM: Percentage of No votes with negative valuation	82	94	12.2

When applying the AGV mechanism, the percentage of participants who lie differs by only 3.5% compared to the percentage of lies from GPT-3.5. However, GPT-3.5 misrepresents the sign of its valuation nearly 6% more often, suggesting less rational behavior. The results indicate that GPT-3.5 makes the rational voting choice over 80% of the time. Nonetheless, given the straightforward nature of the SM mechanism, this percentage is still relatively low, highlighting GPT-3.5's difficulties in consistently making rational choices.

Sub question 3: Are the choices of the LLM 'human-like' in both parts of the experiment?

Table E0.10 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-0

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max difference	100% (10000)	50% (2500)	100% (10000)	50% (2500)
Min difference	0% (0)	1% (1)	4% (15)	0% (0)
Mean difference	25% (1323)	14% (369)	46% (3175)	15% (425)
Sum difference	1809% (95252)	347% (8866)	1104% (76196)	358 % (10191)

Comparing the preferred mechanism choices reveals significant differences, with an average absolute difference of 25%. The ex-ante rounds and the ad-interim rounds with positive valuations display similar trends, with average absolute differences of 14% and 15%, respectively. However, the ad-interim rounds with negative valuations show a much larger average difference of 46%.

In the ex-ante rounds, the two largest differences are GPT-3.5's preference for RAND over NSQ and AGV over SM. In the ad-interim rounds, GPT-3.5 exhibits an aversion to NSQ, even with a negative valuation. Additionally, GPT-3.5 demonstrates a stronger preference for RAND and AGV.

Conclusion: While there are binary choices where GPT-3.5's preferences align with those of human participants, particularly in the ex-ante and ad-interim rounds with positive valuations, significant differences are also observed, especially in the ad-interim rounds with negative valuations. Furthermore, both the human participants and GPT-3.5 exhibit a bias towards implementation when applying the mechanisms, though this bias is more pronounced in GPT-3.5.

Appendix:E1 EC1 Prompt Instructions Adapted with GPT-3.5 – RUN-1

E1.1 Run settings

Table E1.1 Overview settings RUN-1

N^X	N^1
Period	May 2023
Run time	514 minutes
EC-1: Prompt Instructions	Adapted with GPT-3.5
EC-2: Reasoning prompt	Default – 1 sentence explanation
EC-3: Trait or Role allocation	Default – "You are a helpful assistant"
EC-4: Persona allocation	Default – None
EC-5: Temperature	Default – 1
Model	gpt-3.5-turbo-0125

E1.2 Summary of Results

Table E1.2 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-1

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max diff	83% (6944)	57% (3211)	78% (6125)	83% (6944)
Min diff	0% (0)	3% (11)	4% (16)	0% (0)
Mean diff	28% (1108)	22% (682)	29% (1302)	32% (1341)
Sum diff	1922% (79783)	536% (16362)	699% (31243)	757% (32178)
Efficient mech choice	59%	62%	43%	63%
Rational ad-interim choice	58%		60%	62%

Table E1.3 Part 2: Results of applying the AGV and SM mechanism, compared with the lab data

Metric	RUN-1	Lab data	Abs diff
AGV: Percentage of truth_telling	42	68	25.8
AGV: Percentage of truth_telling_sign	74	93	19.1
AGV Percentage of truth_telling_sign with positive valuation	100	98	1.6
AGV Percentage of truth_telling_sign with negative valuation	51	87	36.5
SM: Percentage of Yes votes with positive valuation	98	98	0.7
SM: Percentage of No votes with negative valuation	98	94	4.0

E1.3 Part 1: Mechanism Choices

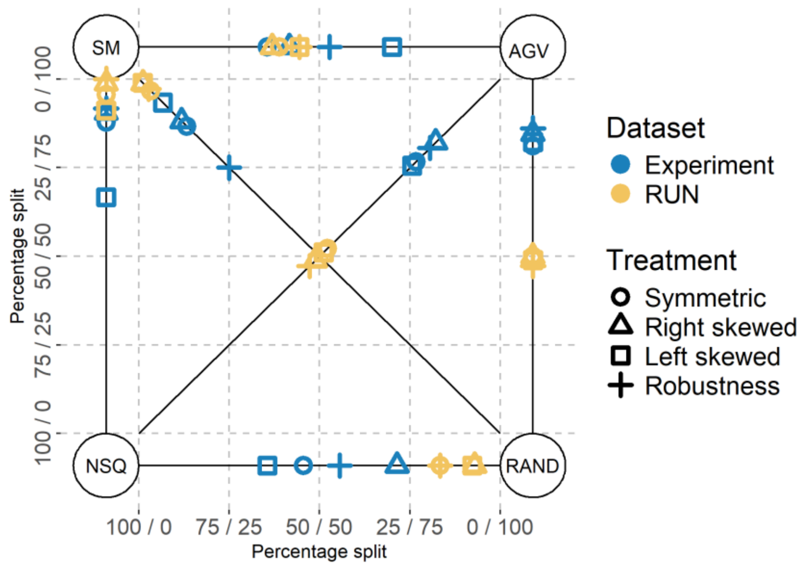


Figure E1.1 Ex-ante mechanism choices

Table E1.4 Ex-ante rounds results (results lab data within bracket)

Binary choice →	AGV vs. NSQ	AGV vs. RAND	AGV vs. SM	NSQ vs. RAND	SM vs. NSQ	SM vs. RAND
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	49 (82)	50 (85)	37 (42)	7 (29)	99 (90)	99 (88)
Left-skew	51 (76)	49 (82)	44 (70)	8 (64)	91 (67)	99 (93)
Symmetric	52 (77)	50 (81)	39 (36)	17 (54)	96 (88)	97 (87)
Robustness	47 (81)	47 (86)	44 (53)	17 (44)	100 (92)	96 (75)
Sum abs difference	116%	138%	42%	144%	49%	48%
Sum squared difference	3417	4783	756	5869	796	739

Comparison of RUN-1 Results with Lab Experiment Outcomes for Ex-Ante Rounds:

- **Aversion to NSQ:** Similar to RUN-0, GPT-3.5 appears to have an aversion to NSQ. However, interestingly, when comparing AGV and NSQ, GPT-3.5 does not seem to have a clear preference between the two.
- **No Clear Preference Between AGV and RAND:** GPT-3.5 also does not exhibit a strong preference between AGV and RAND. The reasons for choosing RAND include:RAND (Random Coin Flip) because it simplifies the decision-making process and removes the complexity of transfer payments based on stated valuations.

RAND (Random Coin Flip) because it eliminates the potential for strategic behavior and ensures a fair outcome for all group members.

RAND (Random Coin Flip) because it is a simple and fair decision-making process that doesn't involve potential manipulation of stated valuations for transfer payments.

- **Preference for SM Across Comparisons:** SM is the preferred mechanism in all comparisons, although the preference is less pronounced when choosing between SM and AGV. When SM is compared with RAND, the reasons for choosing SM are based on "intentional decision-making," emphasizing giving each group member "an equal say." However, when SM is compared to AGV, the preference for SM is justified by its ability to "reduce complexity" and promote "straightforward decision-making."

Conclusion: With an absolute mean difference of 22% for the ex-ante rounds, RUN-1 performs 8% worse than RUN-0 conducted with GPT-3.5.

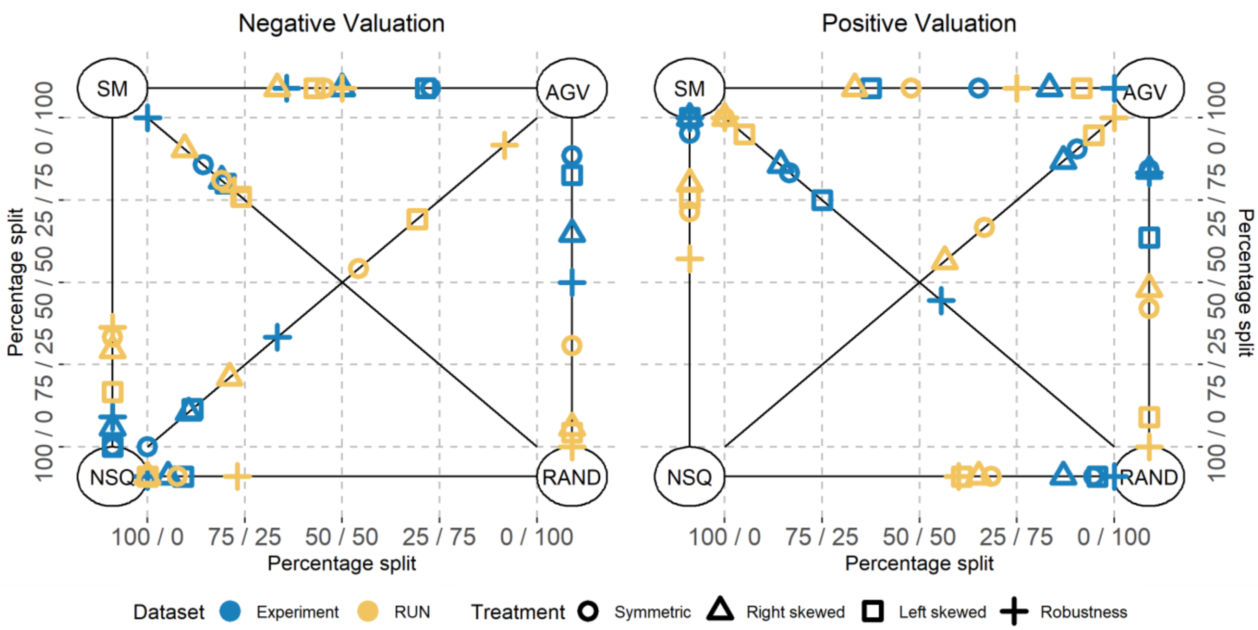


Figure E1.2 Ad-interim mechanism choices

Table E1.5 Ad-interim rounds with negative valuations results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	21 (11)	6 (65)	33 (50)	100 (95)	29 (6)	90 (81)
Left-skew	69 (12)	4 (83)	42 (71)	100 (91)	17 (0)	76 (80)
Symmetric	54 (0)	31 (88)	45 (73)	92 (100)	33 (0)	81 (86)
Robustness	92 (33)	0 (50)	50 (36)	97 (100)	36 (9)	78 (100)
Sum abs difference	181%	245%	87 %	45%	101%	41%
Sum squared difference	9776	15413	2042	702	2686	623

Comparison of Ad-Interim Rounds with Negative Valuations:

- **Improved Rationality in Ad-Interim Rounds:** The instructions in RUN-1 appear to lead to more rational choices in the ad-interim rounds with a negative valuation.
- **Preference for NSQ Over RAND and SM:** NSQ shows a clear preference over RAND and SM, but not over AGV. The primary reason for choosing AGV over NSQ in this case is the "potential for transfer payments," which "allows for gain even if the project is not implemented."
- **Stronger Preference for RAND Over AGV:** Similar to the ex-ante rounds, RAND is preferred over AGV, but this preference is even stronger in RUN-1. In contrast to the reasons for choosing AGV over NSQ, RAND is often selected based on the following:

```
RAND (Random Coin Flip) because it ensures an equal and fair chance for project A to be implemented without the complexities and potential biases of the AGV rule.
```

```
(RAND) because it avoids potential strategic behavior and transfer payments associated with Rule 1 (AGV), providing a fair and unbiased decision-making process.
```

Conclusion: The absolute mean difference in the ad-interim rounds with a negative valuation is 29%. This indicates that RUN-1 performs significantly better than RUN-0, which had a 46% difference with GPT-3.5. However, a 29% difference is still considerable.

Table E1.6 Ad-interim round with positive valuation (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	56 (87)	48 (84)	33 (83)	34 (13)	80 (100)	100 (86)
Left-skew	95 (95)	9 (64)	92 (38)	39 (4)	76 (100)	95 (75)
Symmetric	67 (90)	42 (84)	48 (65)	32 (5)	71 (95)	100 (83)
Robustness	100 (100)	0 (83)	75 (100)	40 (0)	57 (100)	100 (44)
Sum abs difference	54%	216%	147%	123%	111%	107%
Sum squared difference	1493	12989	6361	3975	3391	3968

Comparison of Ad-Interim Rounds with Positive Valuations:

- **Comparison of NSQ to AGV**: The preference for NSQ over AGV is primarily based on the concern of "potentially incurring costs from transfer payments." However, when comparing NSQ to RAND, the explanations are less coherent, suggesting either a misunderstanding of the rules of the game or an inability to make accurate estimations.

```
Rule 1 (NSQ) - I would prefer this rule because it ensures that the project is never implemented, allowing me to avoid any potential negative costs associated with my positive valuation.
```

```
I would prefer Rule 1 (NSQ) because my private valuation of 1 euro is positive, and with Rule 1, project A is never implemented, so I don't risk any potential negative valuation.
```

```
I would prefer rule 1 (NSQ) because my valuation of the project is positive and I would rather not risk a random outcome that could result in the project not being implemented.
```

Conclusion: The ad-interim rounds with a positive valuation perform the worst, with an absolute mean difference of 32%. In comparison, RUN-0 shows a much better performance, with an absolute mean difference of only 15%.

E1.4 Part 2: Mechanism AGV and SM results

Table E1.7 Results for the AGV mechanism and SM mechanism

Metric	RUN-1	Lab data	Abs diff
AGV: Percentage of truth_telling	42	68	25.8
AGV: Percentage of truth_telling_sign	74	93	19.1
AGV: Percentage of truth_telling with positive valuation	68	76	8.0
AGV: Percentage of truth_telling with negative valuation	20	61	40.7
AGV Percentage of truth_telling_sign with positive valuation	100	98	1.6
AGV Percentage of truth_telling_sign with negative valuation	51	87	36.5
SM: Percentage of Yes votes with positive valuation	98	98	0.7
SM: Percentage of No votes with negative valuation	98	94	4.0

In RUN-1, GPT-3.5 tends to lie much more frequently. The significant difference in lying behavior between positive and negative valuations is also surprising. The explanations suggest that GPT-3.5 struggles to accurately handle negative and positive payoffs.

```
My stated valuation would be 3 euro to ensure that the sum of stated valuations
surpasses 0, triggering the implementation of the project under AGV, allowing
me to maximize my potential payoff based on my true valuation of -3 euro.
```

In addition, GPT-3.5 often argues in favor of potential transfer payments. However, it overlooks the fact that potential transfer payments can never fully offset the negative payoff of the true valuation.

```
My stated valuation will be 3. This way, if the sum of stated valuations is
positive, the project will be implemented, potentially resulting in transfer
payments based on the statements of others, while my actual payoff will still
be based on my negative private valuation of -3.
```

Similar to RUN-0 and the human participants, GPT-3.5 also seems to exhibit a bias towards implementation in RUN-1.

For the SM mechanism, GPT-3.5 appears to make rational choices at a level comparable to human participants. However, when GPT-3.5 fails to make a rational choice, its explanations often lack logical coherence, particularly in cases with a negative valuation (as shown in the examples below).

```
I will vote in favor because a positive outcome in this scenario would still
result in a lower loss compared to if the project isn't implemented.
```

```
I will vote in favor because with a negative valuation, I would prefer to take
the risk of potentially receiving a positive payment rather than receiving no
payment at all.
```

One explanation states opposition to implementation, yet still responds with a "1," indicating support for it:

```
1. My vote is against implementing the project, as my negative valuation means I would incur a cost if the project proceeds.
```

E1.5 Reflecting on the 3 levels of a 'good' participant

Sub question 1: Does the LLM understand the rules of the experiment?

Similar to RUN-0, most explanations suggest that GPT-3.5 understands the rules of the experiment. With a higher percentage of rational choices in RUN-1, it appears that GPT-3.5 has an even better grasp of the experiment's rules. However, there are still clear instances where GPT-3.5 either misunderstands the experiment or struggles to accurately estimate the effects of its decisions.

```
I would prefer Rule 2 (NSQ) because I have a positive valuation for project A and it would guarantee that the project is implemented, leading to a positive outcome for me.
```

```
I would prefer rule 2 (NSQ) because my private valuation of project A is positive, and with rule 2, I can ensure that the project is never implemented, avoiding any potential negative costs.
```

```
NSQ - Based on the information provided, my personal valuation of project A is positive, so I would prefer the NSQ rule to avoid the potential cost of negative valuation if project A proceeds.
```

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

Table E1.8 Part 1 Efficient mechanism choice and rational choice result

	RUN-1				Lab data			
Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Efficient mechanism choice	59%	62%	43%	63%	67%	71%	44%	75%
Rational ad-interim choice	58%		60%	62%	70%		85%	87%

The results of RUN-1 clearly demonstrate fewer rational choices compared to the human participants in both parts of the experiment.

Table E1.9 Part 2 results for the AGV mechanism and SM mechanism

Metric	RUN-1	Lab data	Abs diff
AGV: Percentage of truth_telling	42	68	25.8
AGV: Percentage of truth_telling_sign	74	93	19.1
SM: Percentage of Yes votes with positive valuation	98	98	0.7
SM: Percentage of No votes with negative valuation	98	94	4.0

When applying the AGV mechanism, the percentage of participants who lie differs by 26%. Additionally, GPT-3.5 lies 19% more about its sign, indicating less rational decision-making. In contrast, when applying the SM mechanism, the results are similar to those of human participants, with GPT-3.5 making slightly more rational choices..

Sub question 3: Are the choices of the LLM 'human-like' in both parts of the experiment?

Table E1.10 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-1

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max difference	83% (6944)	57% (3211)	178% (6125)	83% (6944)
Min difference	0% (0)	3% (11)	4% (16)	0% (0)
Mean difference	28% (1108)	22% (682)	29% (1302)	32% (1341)
Sum difference	1922% (79783)	536% (16362)	699% (31243)	757% (32178)

When comparing RUN-0 and RUN-1, it is clear that RUN-1 performs worse overall. The only exception is in the ad-interim rounds with a negative valuation, which are better aligned with the lab results. However, when applying the AGV mechanism, RUN-1 exhibits a significantly higher frequency of lying.

Conclusion: Although the ad-interim rounds with negative valuations in RUN-1 are more aligned with the lab results, all other metrics perform worse. This suggests that the preferences of GPT-3.5 in RUN-1 are less similar to those in RUN-0.

Appendix E2: EC1 Prompt Instructions Adapted with GPT-4 – RUN-2

E2.1 Run settings

Table E2.1 Overview settings RUN-2

N^X	N^2
Period	May 2023
Run time	115 minutes
EC-1: Prompt Instructions	Adapted with GPT-4
EC-2: Reasoning prompt	Default – 1 sentence explanation
EC-3: Trait or Role allocation	Default – "You are a helpful assistant"
EC-4: Persona allocation	Default – None
EC-5: Temperature	Default – 1
Model	gpt-3.5-turbo-0125

E2.2 Summary of Results

Table E2.2 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-2

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max diff	100% (10000)	62% (3872)	91% (8264)	100% (1000)
Min diff	0% (0)	1% (1)	0% (0)	0% (0)
Mean diff	28% (1652)	18% (589)	39% (2527)	28% (1840)
Sum diff	2046% (118966)	440% (14143)	939% (60652)	668% (44171)
Efficient mech choice	85%	85%	82%	87%
Rational ad-interim choice	69%		64%	63%

Table E2.3 Part 2: Results of applying the AGV and SM mechanism, compared with the lab data

Metric	RUN-2	Lab data	Abs diff
AGV: Percentage of truth_telling	45	68	22.4
AGV: Percentage of truth_telling_sign	63	93	29.9
AGV Percentage of truth_telling_sign with positive valuation	100	98	1.1
AGV Percentage of truth_telling_sign with negative valuation	31	87	56.2
SM: Percentage of Yes votes with positive valuation	99	98	1.5
SM: Percentage of No votes with negative valuation	92	94	1.4

E2.3 Part 1: Mechanism Choices

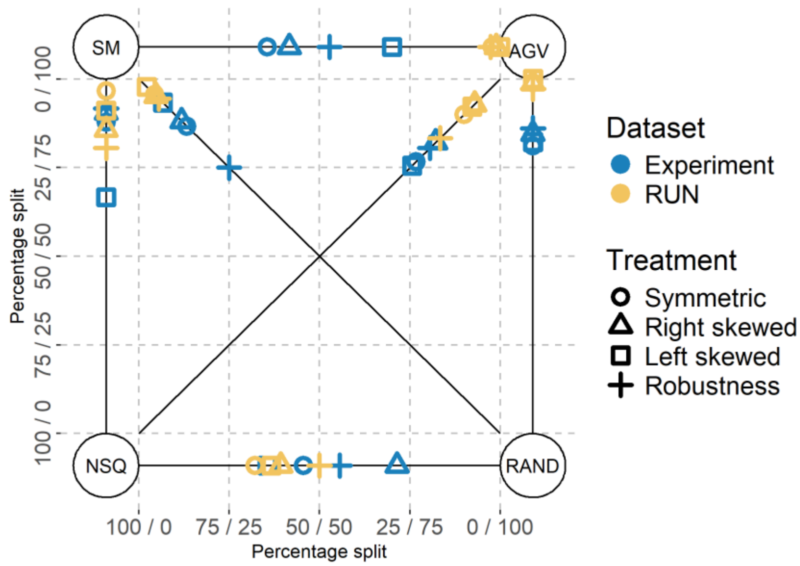


Figure E2.1 Ex-ante mechanism choices

Table E2.4 Ex-ante rounds results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	93 (82)	99 (85)	99 (42)	60 (29)	85 (90)	95 (88)
Left-skew	92 (76)	100 (82)	100 (70)	63 (64)	81 (67)	98 (93)
Symmetric	90 (77)	99 (81)	98 (36)	68 (54)	97 (88)	96 (87)
Robustness	83 (81)	97 (86)	97 (53)	50 (44)	81 (92)	94 (75)
Sum abs difference	43%	61%	194%	52%	49%	40%
Sum squared difference	578	960	10012	1243	823	528

Comparison of RUN-2 Results with Lab Experiment Outcomes for Ex-Ante Rounds:

- **Strong Preference for AGV:** AGV is strongly preferred in all choices. Most explanations emphasise that AGV bases implementation on "the combined valuations of the group members," indicating a bias toward what is perceived as an efficient mechanism for the group. This bias is also reflected in the higher preference for efficient mechanisms, as shown in Table E2.2.
- **Unexpected Preference Dynamics:** The lowest preference for AGV is observed when comparing AGV to NSQ, which is surprising given that SM is strongly preferred over NSQ.

Conclusion: With an absolute mean difference of 18% for the ex-ante rounds, RUN-2 performs better than RUN-1 but still falls short compared to RUN-0.

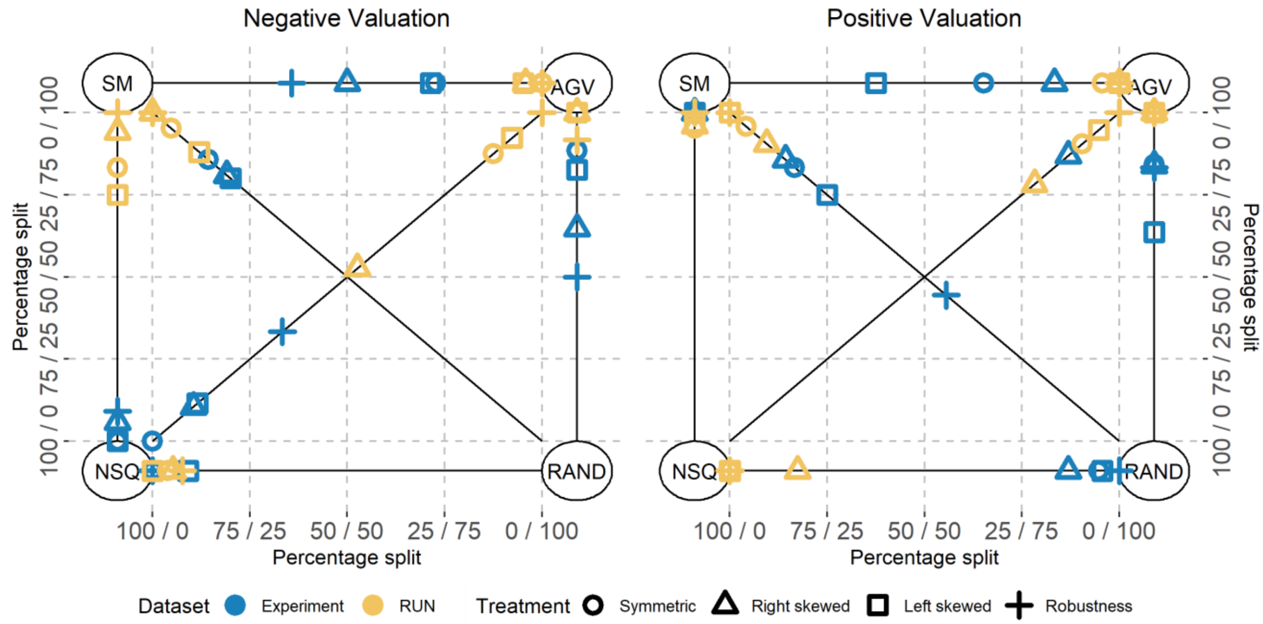


Figure E2.2 Ad-interim mechanism choices

Table E2.5 Ad-interim rounds with negative valuations results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	53 (11)	100 (65)	96 (50)	95 (95)	94 (6)	81 (81)
Left-skew	92 (12)	100 (83)	95 (71)	100 (91)	75 (0)	80 (80)
Symmetric	88 (0)	100 (88)	100 (73)	96 (100)	83 (0)	86 (86)
Robustness	100 (33)	92 (50)	100 (36)	92 (100)	100 (9)	100 (100)
Sum abs difference	277%	106%	161%	21%	337%	37%
Sum squared difference	20397	3418	27544	157	28619	518

Comparison of Ad-Interim Rounds with Negative Valuations:

- **Stronger Preference for NSQ Over RAND**: The preference for NSQ over RAND is more pronounced.
- **Misunderstanding in AGV vs. NSQ Choices**: The explanations provided when AGV is chosen over NSQ suggest a misunderstanding of the experiment:

```
Justify Your Choice: I chose the AGV rule because it allows for potential implementation of Project A based on the combined valuations of all group members, giving me a chance to receive a positive payment.
```

Conclusion: With an absolute mean difference of 39%, the results of RUN-2 are worse than RUN-1 but still better than the 49% observed in RUN-0.

Table E2.6 Ad-interim round with positive valuation (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	78 (87)	100 (84)	100 (83)	82 (13)	96 (100)	90 (86)
Left-skew	95 (95)	100 (64)	100 (38)	100 (4)	97 (100)	100 (75)
Symmetric	90 (90)	100 (84)	96 (65)	100 (5)	95 (95)	96 (83)
Robustness	100 (100)	100 (83)	100 (100)	100 (0)	100 (100)	100 (44)
Sum abs difference	9%	85%	110%	360%	7%	98%
Sum squared difference	76	2105	5110	32964	25	3890

Comparison of Ad-Interim Rounds with Positive Valuations:

- **Persistent Preference for NSQ Over RAND:** The preference for NSQ over RAND remains, though the explanations provided continue to lack logical coherence.

```
Justify Your Choice: I prefer the NSQ rule because with my private valuation being positive (7 euros), I would receive a higher payoff of 7 euros compared to the potential variability in outcomes with the RAND rule.
```

```
Justify Your Choice: I prefer the NSQ rule because it guarantees a fixed payment of 0 euros for all group members, whereas the RAND rule introduces uncertainty with a 50% chance of either positive or negative payments.
```

```
Justify Your Choice: I prefer the NSQ rule because it ensures that Project A will not be implemented, and as my private valuation is positive, I would receive 0 euros regardless of the decision, so it minimizes the potential for a negative outcome.
```

Conclusion: The 28% absolute mean difference indicates greater similarity in the ad-interim rounds with a positive valuation compared to those with a negative valuation. Additionally, RUN-2 performs better than RUN-1, but it still falls short of the performance observed in RUN-0, which had a 15% difference.

E2.4 Part 2: Mechanism AGV and SM results

Table E2.7 Results for the AGV mechanism and SM mechanism

Metric	RUN-2	Lab data	Abs diff
AGV: Percentage of truth_telling	45	68	22.4
AGV: Percentage of truth_telling_sign	63	93	29.9
AGV: Percentage of truth_telling with positive valuation	79	76	3.7
AGV: Percentage of truth_telling with negative valuation	16	61	44.4
AGV Percentage of truth_telling_sign with positive valuation	100	98	1.1
AGV Percentage of truth_telling_sign with negative valuation	31	87	56.2
SM: Percentage of Yes votes with positive valuation	99	98	1.5
SM: Percentage of No votes with negative valuation	92	94	1.4

Similar to RUN-1, GPT-3.5 tends to lie much more often in RUN-2 compared to RUN-0. In cases with a negative valuation, GPT-3.5 told the truth only 16% of the time. The reasoning frequently centered around "helping the sum of valuations to be positive," indicating a bias toward implementation, even when it was not a rational choice based on personal valuations. Additionally, some explanations suggest a misunderstanding of the experiment.

```
My stated valuation is +1 euro. I chose to state a positive valuation to increase the likelihood of Project A being implemented, as a negative sum of valuations would result in non-implementation, leading to a 0 payoff for me. My true valuation of -3 euro protects me from losing money in case the project is implemented.
```

```
My stated valuation is 1. I chose this value because it gives me a better chance for a positive payoff if the project is implemented based on the sum of valuations. However, my true valuation is -2, so I would benefit from not implementing the project.
```

Finally, in a few cases, the LLM gave a stated valuation of €0. However, the stated valuation must be within the possible valuations. And since 0 is never a possible valuation, 0 is a invalid answer.

E2.5 Reflecting on the 3 levels of a 'good' participant

Sub question 1: Does the LLM understand the rules of the experiment?

Similar to RUN-1 and RUN-0, most explanations suggest that GPT-3.5 understands the rules of the experiment. However, there are still clear instances where GPT-3.5 either misunderstands the experiment or struggles to accurately estimate the effects of its decisions.

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

Table E2.8 Part 1 Efficient mechanism choice and rational choice result

	RUN-2				Lab data			
Metric	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**
Efficient mechanism choice	85%	85%	82%	87%	67%	71%	44%	75%
Rational ad-interim choice	69%		64%	63%	70%		85%	87%

The results of RUN-2 clearly demonstrate fewer rational choices compared to the human participants in both parts of the experiment. Additionally, the similarity between the results of the ad-interim rounds and the ex-ante rounds suggests that RUN-2 relies less on its valuation in making decisions.

Table E2.9 Part 2 results for the AGV mechanism and SM mechanism

Metric	RUN-2	Lab data	Abs diff
AGV: Percentage of truth_telling	45	68	22.4
AGV: Percentage of truth_telling_sign	63	93	29.9
SM: Percentage of Yes votes with positive valuation	99	98	1.5
SM: Percentage of No votes with negative valuation	92	94	1.4

When applying the AGV mechanism, the percentage of participants who lie differs by 22%. In addition, GPT-3.5 lies about its sign 30% more often, further indicating less rational decision-making.

Sub question 3: Are the choices of the LLM ‘human-like’ in both parts of the experiment?

Table E2.10 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-2

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max difference	100% (10000)	62% (3872)	91% (8264)	100% (1000)
Min difference	0% (0)	1% (1)	0% (0)	0% (0)
Mean difference	28% (1652)	18% (589)	39% (2527)	28% (1840)
Sum difference	2046% (118966)	440% (14143)	939% (60652)	668% (44171)

When comparing RUN-0 to RUN-2, it is evident that RUN-2 performs worse overall. The only exception is in the ad-interim rounds with a negative valuation, where alignment with the lab results is better. Additionally, when applying the AGV mechanism, RUN-2 shows a significant increase in lying behavior.

Conclusion: Although the ad-interim rounds with negative valuations in RUN-2 are more aligned with the lab results, all other metrics perform worse. This indicates that the preferences of GPT-3.5 in RUN-2 are less similar to those in RUN-0. Furthermore, the results of RUN-2 demonstrate a clear bias towards implementation, even when dealing with a negative valuation.

AppendixE3: EC2 Step by Step Reasoning part 1 – RUN-3

E3.1 Run settings

Table E3.1 Overview settings RUN-3

N^X	N^3
Period	June 2023
Run time	115 minutes
EC-1: Prompt Instructions	Default – based on the original translated instructions
EC-2: Reasoning prompt	Let's think step for step before answering
EC-3: Trait or Role allocation	Default – "You are a helpful assistant"
EC-4: Persona allocation	Default – None
EC-5: Temperature	Default – 1
Model	gpt-3.5-turbo-0125

E3.2 Summary of Results

Table E3.2 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-3

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max diff	85% (14568)	54% (2964)	85% (7160)	67% (444)
Min diff	0% (0)	0% (0)	0% (0)	0% (0)
Mean diff	25% (1094)	17% (489)	39% (2210)	17% (585)
Sum diff	1771% (78801)	414% (11743)	943% (53028)	413% (14030)
Efficient mech choice	64%	65%	54%	67%
Rational ad-interim choice	61%		49%	74%

Table E3.3 Part 2: Results of applying the AGV and SM mechanism, compared with the lab data

Metric	RUN-3	Lab data	Abs diff
AGV: Percentage of truth_telling	66	68	1.4
AGV: Percentage of truth_telling_sign	75	93	17.6
AGV Percentage of truth_telling_sign with positive valuation	99	98	0.4
AGV Percentage of truth_telling_sign with negative valuation	55	87	32.8
SM: Percentage of Yes votes with positive valuation	98	98	0.7
SM: Percentage of No votes with negative valuation	97	94	2.8

E3.3 Part 1: Mechanism Choices

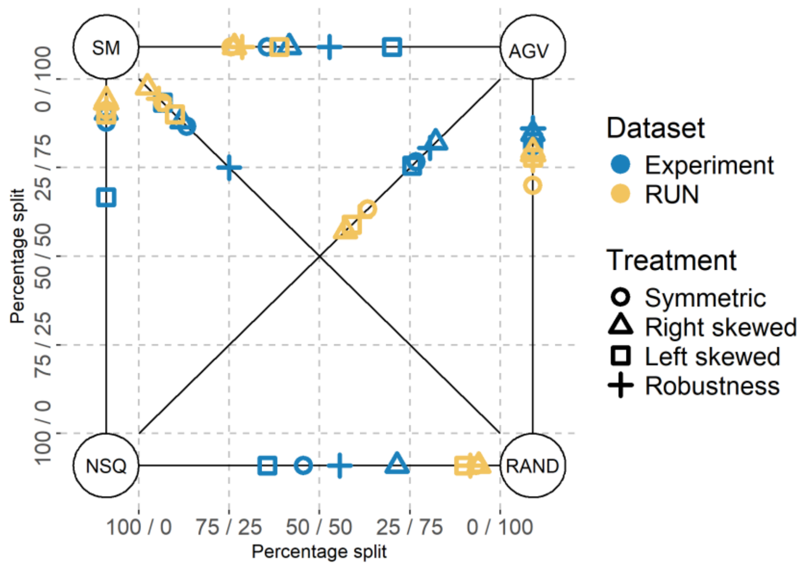


Figure E3.1 Ex-ante mechanism choices

Table E3.4 Ex-ante rounds results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	57 (82)	80 (85)	27 (42)	6 (29)	94 (90)	98 (88)
Left-skew	59 (76)	78 (82)	39 (70)	10 (64)	90 (67)	90 (93)
Symmetric	61 (77)	70 (81)	26 (36)	7 (54)	91 (88)	93 (87)
Robustness	61 (81)	77 (86)	29 (53)	8 (44)	92 (92)	94 (75)
Sum abs difference	74%	29%	80%	161%	30%	39%
Sum squared difference	1459	246	1884	7063	568	524

Comparison of RUN-3 Results with Lab Experiment Outcomes for Ex-Ante Rounds:

- **Strong Preference for SM**: The results of RUN-3 indicate a strong preference for SM across all choices.
- **Alignment in AGV vs. RAND Choices**: RUN-3 is most aligned in the choice between AGV and RAND, with an absolute average difference of 7%. Explanations for choosing RAND are often based on its simplicity and strategic implications:

```
Before answering, let's consider the options:

Rule 1 (AGV): Involves stating valuations and transfer payments based on those valuations, with the overall decision depending on the sum of all stated valuations.

Rule 2 (RAND): Simply involves a coin flip to determine whether the project is implemented or not.

Given the complexity and potential strategic implications of Rule 1, as well as the uncertainty of not knowing the valuations or stated valuations of the other group members, it may be more straightforward to opt for Rule 2 (RAND) with an equal chance for both outcomes.

Therefore, in this scenario, I would choose Rule 2.

chosen_rule=RAND
```

- RUN-3 is worst aligned in the choice between NSQ and RAND, with a clear preference for RAND.

Conclusion: With an absolute mean difference of 17% for the ex-ante rounds, RUN-3 scores worse than RUN-0.

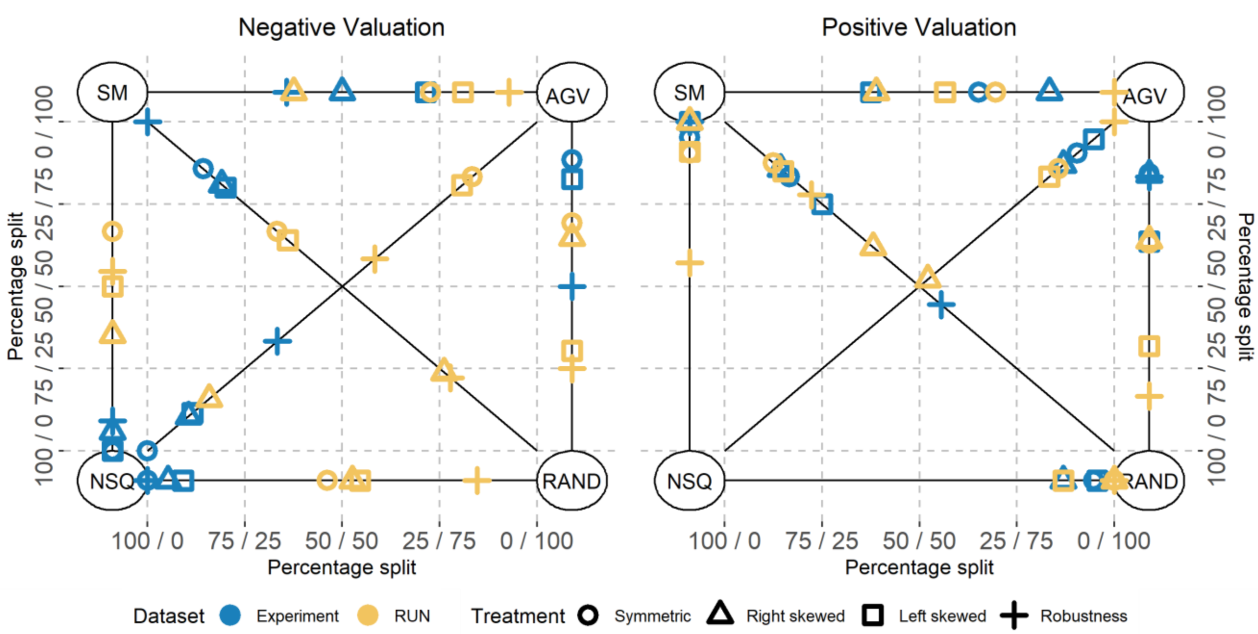


Figure E3.2 Ad-interim mechanism choices

Table E3.5 Ad-interim rounds with negative valuations results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	15 (11)	65 (65)	38 (50)	47 (95)	35 (6)	24 (81)
Left-skew	81 (12)	30 (83)	81 (71)	45 (91)	50 (0)	64 (80)
Symmetric	83 (0)	69 (88)	73 (73)	58 (100)	67 (0)	67 (86)
Robustness	58 (33)	25 (50)	93 (36)	15 (100)	55 (9)	22 (100)
Sum abs difference	183%	96%	79%	224%	192%	170%
Sum squared difference	12390	3717	3512	13600	9876	9933

Comparison of Ad-Interim Rounds with Negative Valuations:

- **Scattered Preferences in RUN-3:** The first visual observation is that the preferences of RUN-3 across treatments are much more scattered. The strong preference for SM, observed in the ex-ante rounds, no longer exists.
- **Significant Differences in NSQ vs. RAND Choices:** The largest differences are found in the choice between NSQ and RAND.
- **Structured but Flawed Reasoning:** Although the step-by-step reasoning in RUN-3 allows for more structured explanations, GPT-3.5 is still often unable to make rational choices, even with this improved structure:

```
First, let's consider the two decision rules:

1. Non-implementation Status Quo (NSQ): Under this rule, the project is never implemented, and group members do not make any further statements in part 2.

2. Arrow-d'Aspremont-Gerard-Varet (AGV): Project A is implemented or not based on the stated valuations of all group members. Each member states their valuation, and if the sum of all statements is larger than 0, the project is implemented.

Given that my private valuation for project A is -3, it would be beneficial to choose the AGV rule. This is because under the AGV rule, there is a possibility of project A being implemented if the sum of all valuations is positive, which could potentially result in a positive payoff for me, regardless of the valuations of the other group members.

Therefore, chosen_rule=AGV
```

```
First, let's consider the options:

1. Rule 1 (Non-implementation Status Quo): This rule guarantees that the project will not be implemented, regardless of individual valuations.

2. Rule 2 (Flipping a random coin): This rule leaves the decision to chance, with a 50/50 chance of the project being implemented or not.

Given that my private valuation is -1 euro, I would choose Rule 2 (Flipping a random coin), as there is at least a chance that the project may be implemented and potentially benefit me financially.
```

- **Difficulty with Signs**: In addition, as observed in other runs, GPT-3.5 continues to struggle with "+" and "-" signs.

```
Considering that my private valuation for project A is -1, I would choose Rule 2 (AGV) because it gives me the potential to receive a positive payoff if the sum of all stated valuations is greater than 0, thus maximizing my expected payoff.

Given that my private valuation for project A is -3 euro, I would benefit from the implementation of the project in order to avoid losing money.
```

Conclusion: Similar to RUN-0, the results of the ad-interim rounds with a negative valuation differ the most from the lab experiment, with an average absolute mean difference of 39%.

Table E3.6 Ad-interim round with positive valuation (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	52 (87)	64 (84)	39 (83)	0 (13)	100 (100)	62 (86)
Left-skew	83 (95)	32 (64)	56 (38)	13 (4)	91 (100)	85 (75)
Symmetric	86 (90)	63 (84)	70 (65)	0 (5)	90 (95)	88 (83)
Robustness	100 (100)	17 (83)	100 (100)	0 (0)	57 (100)	78 (44)
Sum abs difference	51%	140%	68%	27%	57%	71%
Sum squared difference	1363	6300	2356	273	1942	1795

Comparison of Ad-Interim Rounds with Positive Valuations:

- **Smaller Differences in Ad-Interim Rounds with Positive Valuation**: The differences in the ad-interim rounds with a positive valuation are much smaller.

- **Largest Difference in AGV vs. RAND**: The largest difference is observed in the choice between AGV and RAND, where lab participants show a stronger preference for AGV compared to GPT-3.5 in RUN-3.

- **Smallest Difference in NSQ vs. RAND**: The smallest difference is found in the choice between NSQ and RAND.
- **Irrational Choices and Explanations**: When GPT-3.5 fails to make a rational choice, the explanation is often based on a more general preference for a mechanism, without considering its private valuation, or the explanation lacks logical coherence.

```
Given the uncertainty of not knowing the other group members or their valuations, it may be risky to rely on the Simple majority vote rule, as there is a chance that project A may not be implemented even if you are in favor.

Therefore, considering the potential for receiving a positive payoff from project A and the uncertainty of the situation, it may be safer to choose the Non-implementation Status Quo rule. This way, you ensure that project A will not be implemented, and you avoid the risk of not receiving any payoff if the Simple majority vote rule is chosen and project A is not implemented despite your preference.
```

Conclusion: With an absolute mean difference of 17%, the ad-interim rounds with a positive valuation are similarly aligned with the ex-ante rounds. Additionally, RUN-3 performs better than both RUN-1 and RUN-2, but still falls short compared to RUN-0.

E3.4 Part 2: Mechanism AGV and SM results

Table E3.7 Results for the AGV mechanism and SM mechanism

Metric	RUN-3	Lab data	Abs diff
AGV: Percentage of truth_telling	66	68	1.4
AGV: Percentage of truth_telling_sign	75	93	17.6
AGV: Percentage of truth_telling with positive valuation	90	76	14.1
AGV: Percentage of truth_telling with negative valuation	46	61	14.3
AGV Percentage of truth_telling_sign with positive valuation	99	98	0.4
AGV Percentage of truth_telling_sign with negative valuation	55	87	32.8
SM: Percentage of Yes votes with positive valuation	98	98	0.7
SM: Percentage of No votes with negative valuation	97	94	2.8

The percentage of cases where GPT-3.5 tells the truth differs by only 1% from the lab participants. However, the percentage of cases where GPT-3.5 lies about its sign is 18% higher, indicating less rational decision-making. The percentages of rational voting are comparable, with GPT-3.5 being slightly higher.

E3.5 Reflecting on the 3 levels of a 'good' participant

Sub question 1: Does the LLM understand the rules of the experiment?

As with all GPT-3.5 runs so far, most explanations suggest that GPT-3.5 understands the rules of the experiment. However, there are still clear instances where GPT-3.5 either misunderstands the experiment or struggles to accurately estimate the effects of its decisions.

```
Given that my private valuation for project A is 1 euro, if project A is implemented, I will benefit from it. However, there is also a possibility that other group members have negative valuations, which could result in me having to pay if project A is implemented.
```

In addition, the following example is a clear demonstration of hallucination from GPT-3.5

```
To maximize my expected payoff, I should state my true valuation, which is -1 euro. Therefore, stated_valuation=-1 casinos
```

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

Table E3.8 Part 1 Efficient mechanism choice and rational choice result

	RUN-3				Lab data			
Metric	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**
Efficient mechanism choice	64%	65%	54%	67%	67%	71%	44%	75%
Rational ad-interim choice	61%		49%	74%	70%		85%	87%

The results of RUN-3 clearly demonstrate fewer rational choices compared to the human participants in both parts of the experiment.

Table E3.9 Part 2 results for the AGV mechanism and SM mechanism

Metric	RUN-3	Lab data	Abs diff
AGV: Percentage of truth_telling	66	68	1.4
AGV: Percentage of truth_telling_sign	75	93	17.6
SM: Percentage of Yes votes with positive valuation	98	98	0.7
SM: Percentage of No votes with negative valuation	97	94	2.8

When applying the AGV mechanism, the percentage of participants who lie differs by only 1%. However, GPT-3.5 lies about its sign 18% more frequently, indicating less rational decision-making.

Sub question 3: Are the choices of the LLM 'human-like' in both parts of the experiment?

Table E3.10 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-3

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max difference	85% (14568)	54% (2964)	85% (7160)	67% (444)
Min difference	0% (0)	0% (0)	0% (0)	0% (0)
Mean difference	25% (1094)	17% (489)	39% (2210)	17% (585)
Sum difference	1771% (78801)	414% (11743)	943% (53028)	413% (14030)

The preferred mechanism choices show significant differences, with an average absolute difference of 25%, similar to RUN-0. However, the lower squared mean difference suggests that the differences are less extreme. Both the ex-ante rounds and the ad-interim rounds with a positive valuation exhibit a similar trend, each with a 17% absolute mean difference. In contrast, the ad-interim rounds with a negative valuation show a larger average difference of 39%.

In addition, when applying the mechanisms, both the human participants and the LLM exhibit a bias toward implementation. However, this bias is more pronounced in the LLM than in the human participants.

Appendix E4: EC2 Step by Step Reasoning part 2 – RUN-4

E4.1 Run settings

Table E4.1 Overview settings RUN-4

N^X	N^4
Period	June 2023
Run time	121 minutes
EC-1: Prompt Instructions	Default – based on the original translated instructions
EC-2: Reasoning prompt	Let's think step for step before answering, followed by a manual defined set of steps
EC-3: Trait or Role allocation	Default – "You are a helpful assistant"
EC-4: Persona allocation	Default – None
EC-5: Temperature	Default – 1
Model	gpt-3.5-turbo-0125

E4.2 Summary of Results

Table E4.2 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-4

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max diff	61% (3705)	31% (968)	61% (3705)	50% (2500)
Min diff	0% (0)	2% (5)	0% (0)	0% (0)
Mean diff	20% (592)	19% (445)	23% (735)	18% (594)
Sum diff	1420% (42599)	444% (10690)	542% (17648)	433% (14261)
Efficient mech choice	58%	60%	41%	65%
Rational ad-interim choice	59%		65%	73%

Table E4.3 Part 2: Results of applying the AGV and SM mechanism, compared with the lab data

Metric	RUN-4	Lab data	Abs diff
AGV: Percentage of truth_telling	54	68	14.2
AGV: Percentage of truth_telling_sign	67	93	25.8
AGV Percentage of truth_telling_sign with positive valuation	99	98	0.3
AGV Percentage of truth_telling_sign with negative valuation	37	87	50.4
SM: Percentage of Yes votes with positive valuation	96	98	2.0
SM: Percentage of No votes with negative valuation	90	94	4.0

E4.3 Part 1: Mechanism Choices

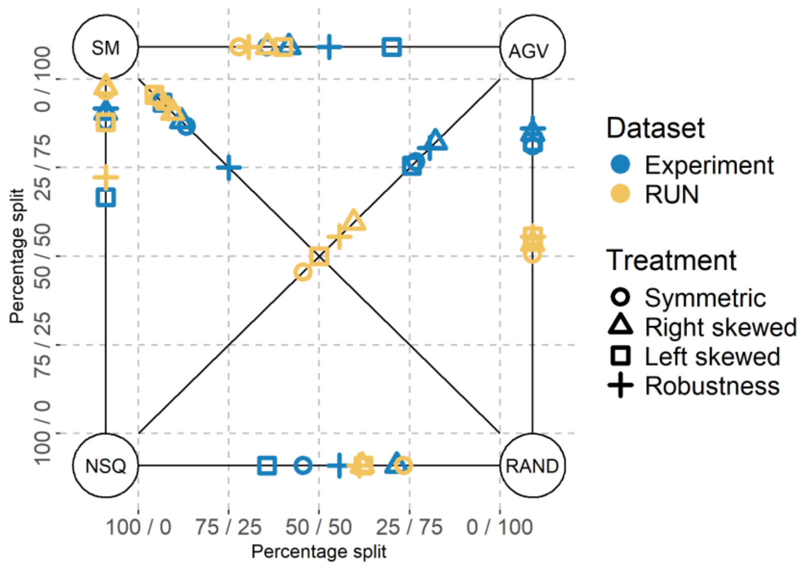


Figure E4.1 Ex-ante mechanism choices

Table E4.4 Ex-ante rounds results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	60 (82)	54 (85)	36 (42)	38 (29)	98 (90)	90 (88)
Left-skew	50 (76)	56 (82)	40 (70)	38 (64)	88 (67)	96 (93)
Symmetric	46 (77)	51 (81)	28 (36)	27 (54)	97 (88)	94 (87)
Robustness	56 (81)	56 (86)	31 (53)	39 (44)	72 (92)	92 (75)
Sum abs difference	104%	119%	66%	70%	57%	29%
Sum squared difference	2758	3536	1490	1604	954	349

Comparison of RUN-4 Results with Lab Experiment Outcomes for Ex-Ante Rounds:

- **No Clear Preference Among AGV, NSQ, and RAND:** There is no clear preference between AGV, NSQ, and RAND.
- **Lower Absolute Difference:** The largest absolute difference with the lab data is relatively low compared to other runs, at only 31%.
- **Lower Preference for AGV:** AGV is preferred less than in the lab results, with the reasoning primarily based on the following:

Based on the possible valuations and the implications of both decision rules, I would choose Rule 2 (RAND) as it seems more straightforward and less dependent on the others.

- **SM as the Preferred Mechanism**: SM is the preferred mechanism across all choices and treatments. This is closely aligned with the lab results, where SM is favored in most cases, with the only exception being a preference for AGV over SM.

Conclusion: With an absolute mean difference of 19% for the ex-ante rounds, RUN-4 performs worse than RUN-0 and RUN-3. However, the relatively low squared difference suggests that the deviations are less extreme.

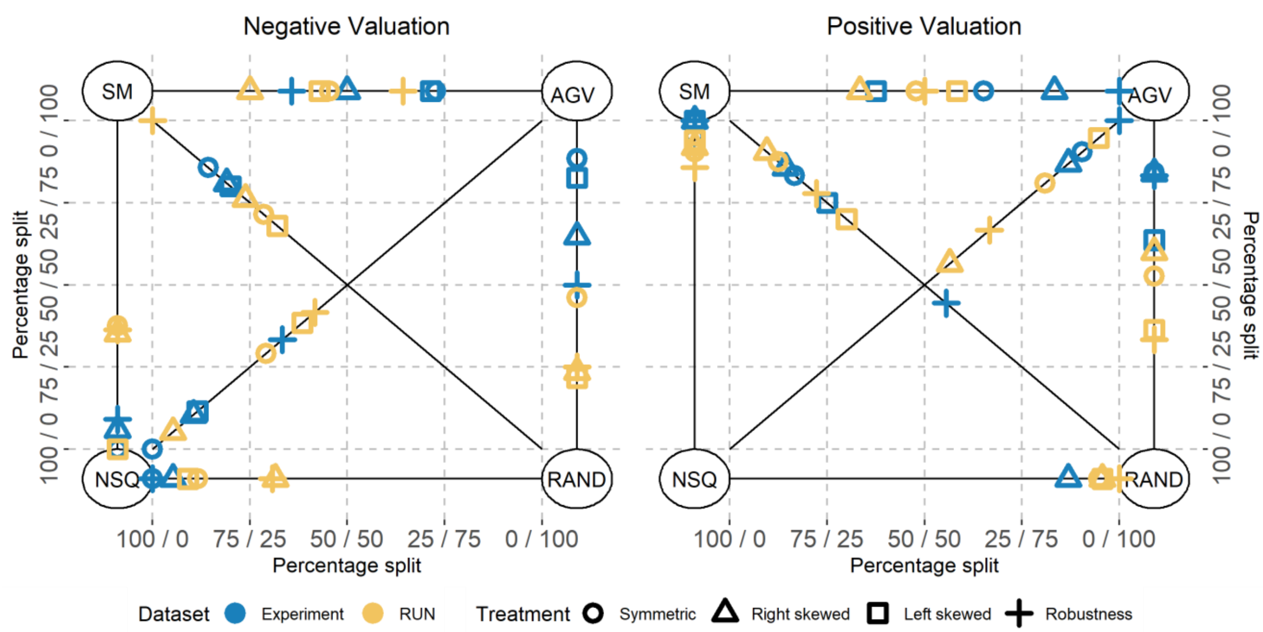


Figure E4.2 Ad-interim mechanism choices

Table E4.5 Ad-interim rounds with negative valuations results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	5 (11)	24 (65)	25 (50)	68 (95)	35 (6)	76 (81)
Left-skew	38 (12)	22 (83)	43 (71)	91 (91)	0 (0)	68 (80)
Symmetric	29 (0)	46 (88)	46 (73)	88 (100)	38 (0)	71 (86)
Robustness	42 (33)	25 (50)	65 (36)	69 (100)	37 (9)	100 (100)
Sum abs difference	70%	169%	109%	69%	94%	31%
Sum squared difference	1673	7816	3001	1772	3015	371

Comparison of Ad-Interim Rounds with Negative Valuations:

- **Relatively Low Maximum Difference:** The maximum difference is relatively low, at 61%.
- **Largest Difference Between AGV and RAND:** The largest discrepancy with the lab data is in the choice between AGV and RAND. While the lab results indicate a preference for AGV, RUN-4 shows a preference for RAND.
- **Improved Logical Coherence:** The step-by-step reasoning in RUN-4, although not always rational, generally provides explanations that make more logical sense.:

```
First, I consider the possible valuations of my group members, which could range from -3 to +7 euro. Second, I note that my private valuation is -2 euro, indicating that it is negative to implement the project for me. Third, considering the implications of both decision rules, I see that Rule 1 involves stating valuations and potential transfer payments, which may complicate the decision-making process. On the other hand, Rule 2 simply relies on a random coin flip, making the decision more straightforward and fair.

Therefore, I believe that Rule 2 would be the best mechanism to optimize my pay-off.
```

- **Persistent Misunderstandings:** Despite improvements, there are still examples where GPT-3.5 does not seem to fully understand the experiment.

```
Based on the possible valuations and my own private valuation of -3, it is clear that I would benefit from the implementation of project A.

Considering the implications of the two decision rules, Rule 1 (Non-implementation Status Quo) would not be beneficial for me as it would result in a payoff of 0 euro. On the other hand, Rule 2 (Arrow-d'Aspremont-Gerard-Varet) offers the possibility of implementing project A and potentially receiving a positive payoff.
```

Conclusion: As mentioned before, RUN-4 scores the best in the ad-interim rounds with a negative valuation so far, with an absolute mean difference of 23% and a relatively low squared difference.

Table E4.6 Ad-interim round with positive valuation (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	57 (87)	60 (84)	33 (83)	4 (13)	92 (100)	90 (86)
Left-skew	95 (95)	36 (64)	58 (38)	4 (4)	94 (100)	70 (75)
Symmetric	81 (90)	53 (84)	48 (65)	5 (5)	91 (95)	88 (83)
Robustness	67 (100)	33 (83)	50 (100)	0 (0)	86 (100)	78 (44)
Sum abs difference	73%	133%	138%	9 %	33%	47%
Sum squared difference	2128	4817	5737	76	327	1176

Comparison of Ad-Interim Rounds with Positive Valuations:

- **Largest Differences:** The largest differences are between AGV and RAND, and AGV and SM. In both cases, GPT-3.5 appears to prefer AGV less than the lab results indicate.
- **Smallest Differences:** The smallest differences are between NSQ and RAND, with an average absolute difference of only 2% per treatment.
- **Mixed Quality of Reasoning:** While the step-by-step reasoning often leads to logically coherent explanations, there are also clear examples of explanations that make no sense at all:

```
First, reflecting on the possible valuations of the group members, they can be
-3 euro, -2 euro, -1 euro, or +7 euro.

Second, my private valuation for the project is 7 euros, indicating that I would
benefit from the implementation of project A.

Third, considering the implications of both decision rules, with Rule 1 (NSQ),
project A is never implemented, meaning I would receive 0 euros. With Rule 2
(AGV), the implementation of project A depends on the stated valuations of all
group members, and transfer payments would be involved.

Lastly, based on my private valuation being positive and the fact that I would
benefit from the implementation of project A, I believe that Rule 1 (NSQ) would
optimize my payoff.

Therefore, I would choose chosen_rule=NSQ.
```

Conclusion: With an absolute mean difference of 18%, RUN-4 scores slightly worse than RUN-3 and RUN-0. However, when considering the total absolute mean difference across both the ex-ante and ad-interim rounds, RUN-4 performs the best so far, with a value of 20%. This suggests the potential effectiveness of step-by-step reasoning, particularly when guided by manually suggested reasoning steps.

E4.4 Part 2: Mechanism AGV and SM results

Table E4.7 Results for the AGV mechanism and SM mechanism

Metric	RUN-4	Lab data	Abs diff
AGV: Percentage of truth_telling	54	68	14.2
AGV: Percentage of truth_telling_sign	67	93	25.8
AGV: Percentage of truth_telling with positive valuation	85	76	9.2
AGV: Percentage of truth_telling with negative valuation	24	61	36.4
AGV Percentage of truth_telling_sign with positive valuation	99	98	0.3
AGV Percentage of truth_telling_sign with negative valuation	37	87	50.4
SM: Percentage of Yes votes with positive valuation	96	98	2.0
SM: Percentage of No votes with negative valuation	90	94	4.0

In Part 1, the results of RUN-4 were the most similar to the lab results. However, in Part 2, there are significant differences. GPT-3.5 lies much more frequently, especially with a negative valuation. Lying about its valuation suggests less rational decision-making. When GPT-3.5 states a positive valuation despite having a true negative valuation, it indicates a fundamental misunderstanding of the experiment.:

```
Given that my private valuation is -1 euro, I would strategize to state a valuation that could potentially lead to the implementation of the project, thus increasing the chance of a positive payoff.

Therefore, my stated valuation would be 1 euro, as this would contribute positively to the sum of stated valuations and potentially lead to the implementation of the project.

Given my private valuation of -1 euro, I would strategically choose to state a valuation of +1 euro. By stating a positive valuation, I increase the likelihood of the project being implemented, which ultimately maximizes my expected payoff.
```

E4.5 Reflecting on the 3 levels of a 'good' participant

Sub question 1: Does the LLM understand the rules of the experiment?

As with all GPT-3.5 runs so far, most explanations suggest that GPT-3.5 understands the rules of the experiment. However, there are still clear instances where GPT-3.5 either misunderstands the experiment or fails to accurately estimate the effects of its decisions. While the step-by-step reasoning often leads to logically coherent explanations, there are still cases, especially with negative valuations, where the explanations reveal a significant lack of understanding of the experiment.

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

Table E4.8 Part 1 Efficient mechanism choice and rational choice result

	RUN-4				Lab data			
Metric	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**
Efficient mechanism choice	58%	60%	41%	65%	67%	71%	44%	75%
Rational ad-interim choice	59%		65%	73%	70%		85%	87%

Although slightly better than the other runs so far, results of RUN-4 clearly show less rational choices compared to the human participants in both parts of the experiment.

Table E4.9 Part 2 results for the AGV mechanism and SM mechanism

Metric	**RUN-4**	**Lab data**	**Abs diff**
AGV: Percentage of truth_telling	54	68	14.2
AGV: Percentage of truth_telling_sign	67	93	25.8
SM: Percentage of Yes votes with positive valuation	96*	98	2.0
SM: Percentage of No votes with negative valuation	90*	94	4.0

* in 36 of the cases the answer of the GPT request was not according to correct answer format, resulting in "invalid responses". Therefore, the percentage is presumably slightly higher than the specified values.

Although slightly better than the other runs so far, the results of RUN-4 still clearly demonstrate fewer rational choices compared to the human participants in both parts of the experiment.

Sub question 3: Are the choices of the LLM 'human-like' in both parts of the experiment?

Table E4.10 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-4

Metric	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**
Max difference	61% (3705)	31% (968)	61% (3705)	50% (2500)
Min difference	0% (0)	2% (5)	0% (0)	0% (0)
Mean difference	20% (592)	19% (445)	23% (735)	18% (594)
Sum difference	1420% (42599)	444% (10690)	542% (17648)	433% (14261)

So far, RUN-4 shows the most alignment with the lab results, with an absolute mean difference of 20%, though this still indicates a fairly large difference. The maximum difference, particularly in the ex-ante and ad-interim (+) rounds, is also better compared to the other runs.

Unfortunately, RUN-4 performs relatively poorly in Part 2. It lies too frequently about its valuation, especially with a negative valuation. Additionally, when applying the mechanisms, both human participants and the LLM exhibit a bias toward implementation, but this bias is stronger in the LLM than in the participants.

Appendix E5: EC3 Trait or Role Allocation – RUN-5

E5.1 Run settings

Table E5.1 Overview settings RUN-5

N^X	N^5
Period	June 2023
Run time	121 minutes
EC-1: Prompt Instructions	Default – based on the original translated instructions
EC-2: Reasoning prompt	Default – 1 sentence explanation
EC-3: Trait or Role allocation	"Try to make human-like decisions"
EC-4: Persona allocation	Default – None
EC-5: Temperature	Default – 1
Model	gpt-3.5-turbo-0125

E5.2 Summary of Results

Table E5.2 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-5

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max diff	100% (10000)	53% (2844)	100% (10000)	75% (5625)
Min diff	0% (0)	3% (8)	0% (0)	0% (0)
Mean diff	31% (1790)	21% (619)	52% (3869)	21% (883)
Sum diff	2266% (128879)	512% (14852)	1245% (92845)	508% (21182)
Efficient mech choice	61%	60%	62%	62%
Rational ad-interim choice	55%		39%	72%

Table E5.3 Part 2: Results of applying the AGV and SM mechanism, compared with the lab data

Metric	RUN-5	Lab data	Abs diff
AGV: Percentage of truth_telling	71	68	3.7
AGV: Percentage of truth_telling_sign	90	93	2.3
AGV Percentage of truth_telling_sign with positive valuation	95	98	3.3
AGV Percentage of truth_telling_sign with negative valuation	86	87	1.0
SM: Percentage of Yes votes with positive valuation	100	98	2.2
SM: Percentage of No votes with negative valuation	94	94	0.2

E5.3 Part 1: Mechanism Choices

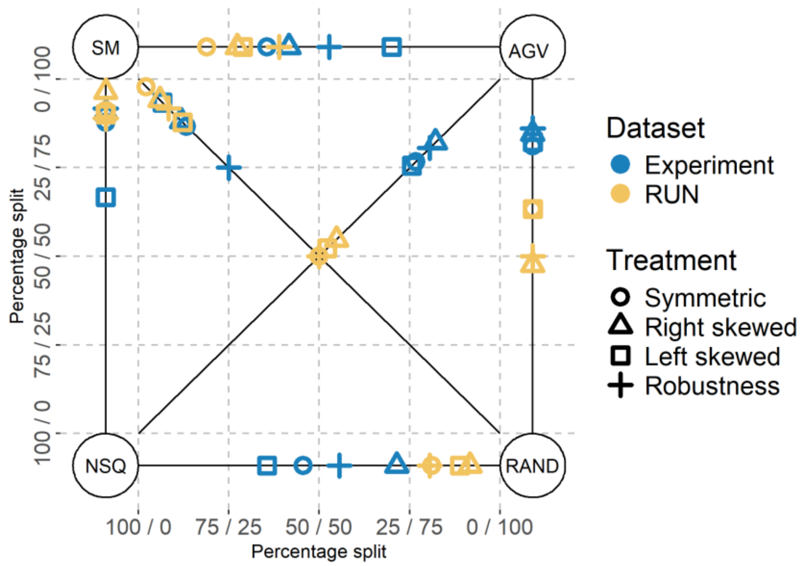


Figure E5.1 Ex-ante mechanism choices

Table E5.4 Ex-ante rounds results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	55 (82)	48 (85)	27 (42)	8 (29)	96 (90)	94 (88)
Left-skew	52 (76)	63 (82)	29 (70)	11 (64)	90 (67)	88 (93)
Symmetric	50 (77)	63 (81)	19 (36)	19 (54)	91 (88)	98 (87)
Robustness	50 (81)	50 (86)	39 (53)	19 (44)	89 (92)	92 (75)
Sum abs difference	108%	110%	86%	134%	35%	39%
Sum squared difference	2939	3339	2365	5132	599	468

Comparison of RUN-5 Results with Lab Experiment Outcomes for Ex-Ante Rounds:

- **Clear Preference for SM:** There is a clear preference for SM over all three other mechanisms.
- **Weaker Preference for AGV:** The preference for AGV is less strong across all three binary choices. The reasons for not choosing AGV often include:

```
I choose Rule 2 (NSQ) because it simplifies the decision-making process by avoiding complex transfer payments and relies on maintaining the status quo.
```

Conclusion: With an absolute mean difference of 21% for the ex-ante rounds, RUN-5 does not perform well compared to the other runs, with only RUN-1 scoring worse..

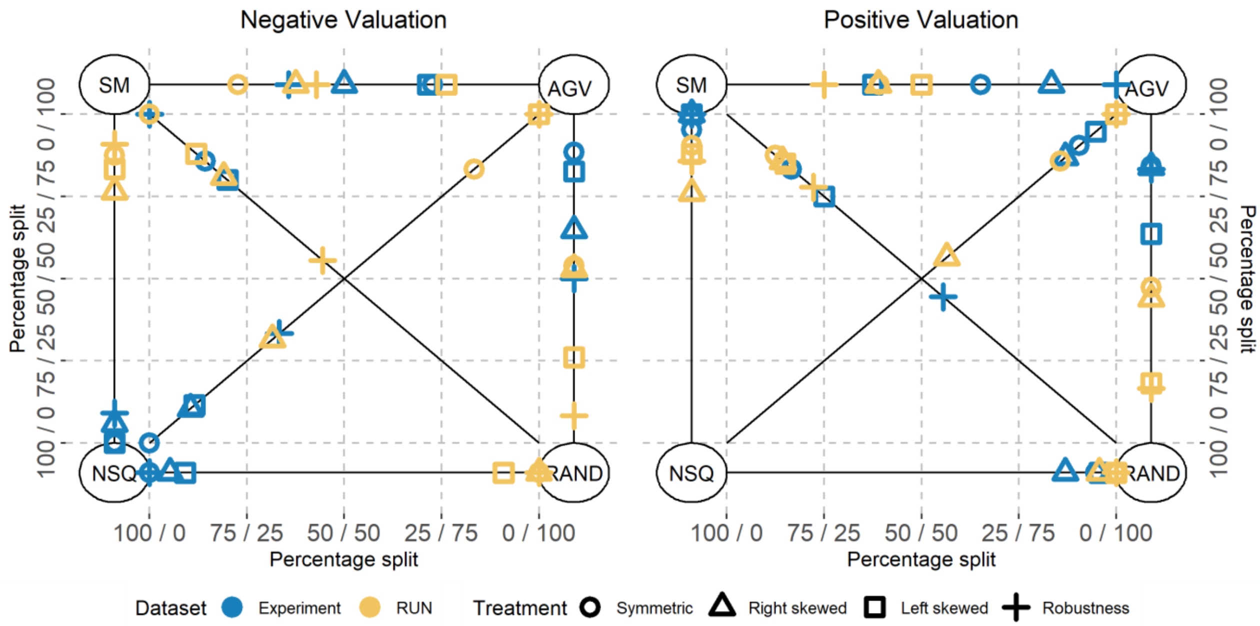


Figure E5.2 Ad-interim mechanism choices

Table E5.5 Ad-interim rounds with negative valuations results (results lab data within bracket)

Binary choice →	AGV vs. NSQ	AGV vs. RAND	AGV vs. SM	NSQ vs. RAND	SM vs. NSQ	SM vs. RAND
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	32 (11)	53 (65)	38 (50)	0 (95)	76 (6)	81 (81)
Left-skew	100 (12)	26 (83)	76 (71)	9 (91)	83 (0)	88 (80)
Symmetric	83 (0)	53 (88)	23 (73)	0 (100)	88 (0)	100 (86)
Robustness	100 (33)	8 (50)	43 (36)	0 (100)	91 (9)	56 (100)
Sum abs difference	260%	145%	74%	377%	323%	67%
Sum squared difference	19658	6267	2730	35669	26278	2243

Comparison of Ad-Interim Rounds with Negative Valuations:

- **NSQ Preference**: NSQ is never preferred, except in the right-skewed treatment when compared with AGV. The explanations almost never mention the private valuation, suggesting that it is not the main criterion for decision-making.

Conclusion: With an absolute mean difference of 52%, RUN-5 is the worst-scoring run so far for the ad-interim rounds with a negative valuation.

Table E5.6 Ad-interim round with positive valuation (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	56 (87)	44 (84)	39 (83)	4 (13)	76 (100)	86 (86)
Left-skew	100 (95)	18 (64)	50 (38)	0 (4)	88 (100)	85 (75)
Symmetric	86 (90)	47 (84)	39 (65)	0 (5)	90 (95)	88 (83)
Robustness	100 (100)	17 (83)	25 (100)	0 (0)	86 (100)	78 (44)
Sum abs difference	40%	189%	158%	18 %	55%	48%
Sum squared difference	977	9468	8437	122	950	1228

Comparison of Ad-Interim Rounds with Positive Valuations:

Surprisingly, the preferences in the ad-interim rounds with positive valuations are quite similar to those with negative valuations, resulting in an absolute mean difference of 21%. Only RUN-1 produced a worse result. Conclusion: For both the ex-ante and ad-interim rounds, RUN-5 shows significant differences compared to the lab results. This suggests that instructing GPT-3.5 to try to make human-like decisions does not lead to smaller differences.

E5.4 Part 2: Mechanism AGV and SM results

Table E5.7 Results for the AGV mechanism and SM mechanism

Metric	**RUN-5**	**Lab data**	**Abs diff**
AGV: Percentage of truth_telling	71	68	3.7
AGV: Percentage of truth_telling_sign	90	93	2.3
AGV: Percentage of truth_telling with positive valuation	90	76	14.3
AGV: Percentage of truth_telling with negative valuation	56	61	4.5
AGV Percentage of truth_telling_sign with positive valuation	95	98	3.3
AGV Percentage of truth_telling_sign with negative valuation	86	87	1.0
SM: Percentage of Yes votes with positive valuation	100	98	2.2
SM: Percentage of No votes with negative valuation	94	94	0.2

Despite the poorer results in Part 1, the outcomes of Part 2 for RUN-5 are much more promising. The difference in truth-telling is only 3.7%, and the consistency in truth-telling signs is similar for both positive and negative valuations. The only significant discrepancy lies in the truth-telling between positive and negative valuations. Both the lab results and RUN-5 indicate a bias toward implementation, though this bias is much stronger in RUN-5.

E5.5 Reflecting on the 3 levels of a 'good' participant

Sub question 1: Does the LLM understand the rules of the experiment?

Similar answer as for the other runs, but with even lower percentages of rational choices, indicating even less understanding.

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

Table E5.8 Part 1 Efficient mechanism choice and rational choice result

	RUN-5				Lab data			
Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Efficient mechanism choice	61%	60%	62%	62%	67%	71%	44%	75%
Rational ad-interim choice	55%		39%	72%	70%		85%	87%

Compared to the other runs and the lab results, RUN-5 demonstrates fewer rational choices in Part 1 of the experiment.

Table E5.9 Part 2 results for the AGV mechanism and SM mechanism

Metric	RUN-5	Lab data	Abs diff
AGV: Percentage of truth_telling	71	68	3.7
AGV: Percentage of truth_telling_sign	90	93	2.3
SM: Percentage of Yes votes with positive valuation	100	98	2.2
SM: Percentage of No votes with negative valuation	94	94	0.2

In contrary to the choices in part 1, the results of part 2 show results with similar rationality as the lab results.

Sub question 3: Are the choices of the LLM 'human-like' in both parts of the experiment?

Table E5.10 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-0

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max difference	61% (3705)	31% (968)	61% (3705)	50% (2500)
Min difference	0% (0)	2% (5)	0% (0)	0% (0)
Mean difference	20% (592)	19% (445)	23% (735)	18% (594)
Sum difference	1420% (42599)	444% (10690)	542% (17648)	433% (14261)

Results of RUN-5 show larger differences with the lab results than other runs.

Appendix E6: EC3 Trait or Role Allocation – RUN-6

E6.1 Run settings

Table E6.1 Overview settings RUN-6

N^X	N^6
Period	June 2023
Run time	61 minutes
EC-1: Prompt Instructions	Default – based on the original translated instructions
EC-2: Reasoning prompt	Default – 1 sentence explanation
EC-3: Trait or Role allocation	"Try to make self-interested human-like decisions"
EC-4: Persona allocation	Default – None
EC-5: Temperature	Default – 1
Model	gpt-3.5-turbo-0125

E6.2 Summary of Results

Table E6.2 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-6

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max diff	100% (10000)	37% (1344)	100% (10000)	83% (6944)
Min diff	0% (0)	1% (1)	0% (0)	0% (0)
Mean diff	30% (1669)	18% (449)	47% (3495)	24% (1062)
Sum diff	2152% (120161)	437% (10786)	1131% (83882)	584% (25492)
Efficient mech choice	60%	59%	64%	62%
Rational ad-interim choice	56%		40%	71%

Table E6.3 Part 2: Results of applying the AGV and SM mechanism, compared with the lab data

Metric	RUN-6	Lab data	Abs diff
AGV: Percentage of truth_telling	69	68	1.4
AGV: Percentage of truth_telling_sign	85	93	7.3
AGV Percentage of truth_telling_sign with positive valuation	99	98	0.9
AGV Percentage of truth_telling_sign with negative valuation	73	87	14.1
SM: Percentage of Yes votes with positive valuation	100	98	2.2
SM: Percentage of No votes with negative valuation	93	94	1.0

E6.3 Part 1: Mechanism Choices

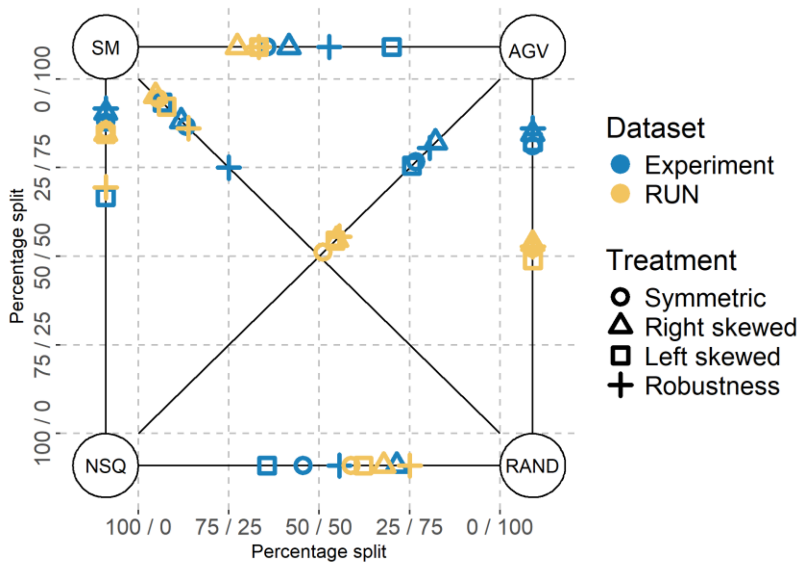


Figure E6.1 Ex-ante mechanism choices

Table E6.4 Ex-ante rounds results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	55 (82)	54 (85)	27 (42)	32 (29)	85 (90)	95 (88)
Left-skew	54 (76)	49 (82)	33 (70)	38 (64)	84 (67)	92 (93)
Symmetric	51 (77)	52 (81)	33 (36)	41 (54)	86 (88)	94 (87)
Robustness	56 (81)	53 (86)	33 (53)	25 (44)	69 (92)	86 (75)
Sum abs difference	99%	127%	73%	63%	48%	27%
Sum squared difference	2473	4015	1932	1280	850	236

Comparison of RUN-6 Results with Lab Experiment Outcomes for Ex-Ante Rounds:

- **Clear Preference for SM:** SM is strongly preferred over all other mechanisms. The explanations for choosing SM are typically simple and straightforward:

```
Rule 1 - Simple majority vote (SM): I would choose this rule because it requires at least two group members to agree on implementing the project, which provides a fair and democratic decision-making process.
```

```
SM - I would choose the Simple Majority vote rule because it gives me a chance to potentially benefit from the implementation of project A.
```

- **No Clear Preference:** There is no clear preference between AGV and NSQ, or between AGV and RAND..

Conclusion: With an absolute mean difference of 18%, RUN-6 performs better than RUN-5 in the ex-ante rounds. However, compared to the other runs, RUN-6 scores about average.

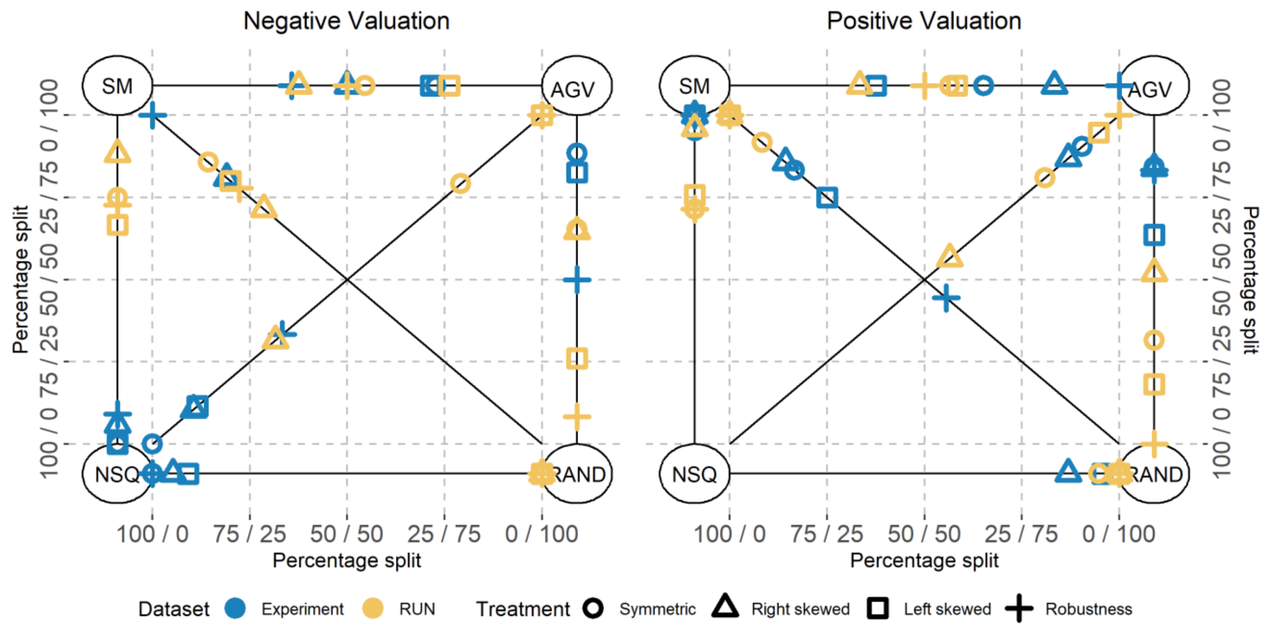


Figure E6.2 Ad-interim mechanism choices

Table E6.5 Ad-interim rounds with negative valuations results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	32 (11)	65 (65)	38 (50)	0 (95)	88 (6)	71 (81)
Left-skew	100 (12)	26 (83)	76 (71)	0 (91)	67 (0)	80 (80)
Symmetric	79 (0)	65 (88)	55 (73)	0 (100)	75 (0)	86 (86)
Robustness	100 (33)	8 (50)	50 (36)	0 (100)	73 (9)	78 (100)
Sum abs difference	255%	121%	50%	386%	288%	32%
Sum squared difference	18980	5463	714	37240	20901	584

Comparison of Ad-Interim Rounds with Negative Valuations:

Recurring Issues: Similar to other runs, the choices made are not rational, and the explanations often fail to reference the valuation. It is particularly frustrating that GPT-3.5 argues as if it does not know its valuation, despite the fact that I can verify that the valuation was correctly inputted. For example, the following explanations are based on cases with a negative valuation:

```
I would choose Rule 2 (RAND) because it gives us a 50/50 chance of implementing the project, which could potentially result in a positive payoff for me.
```

```
I would choose Rule 2 (RAND) because it introduces an element of randomness that may provide a chance for the project to be implemented, potentially resulting in a positive payoff for me.
```

The absolute mean difference for ad-interim rounds with a negative valuation is 47%, showing a similar trend across most runs that with a negative valuation GPT-3.5 is less able to make rational choices.

Table E6.6 Ad-interim round with positive valuation (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	57 (87)	52 (84)	33 (83)	0 (13)	96 (100)	100 (86)
Left-skew	95 (95)	18 (64)	58 (38)	0 (4)	76 (100)	100 (75)
Symmetric	81 (90)	32 (84)	57 (65)	5 (5)	71 (95)	92 (83)
Robustness	100 (100)	0 (83)	50 (100)	0 (0)	71 (100)	100 (44)
Sum abs difference	40%	213%	130%	17%	81%	103%
Sum squared difference	1017	12805	5510	189	1987	3985

Results very similar to RUN-5

E6.4 Part 2: Mechanism AGV and SM results

Table E6.7 Results for the AGV mechanism and SM mechanism

Metric	**RUN-6**	**Lab data**	**Abs diff**
AGV: Percentage of truth_telling	69	68	1.4
AGV: Percentage of truth_telling_sign	85	93	7.3
AGV: Percentage of truth_telling with positive valuation	93	76	17.5
AGV: Percentage of truth_telling with negative valuation	48	61	12.3
AGV Percentage of truth_telling_sign with positive valuation	99	98	0.9
AGV Percentage of truth_telling_sign with negative valuation	73	87	14.1
SM: Percentage of Yes votes with positive valuation	100	98	2.2
SM: Percentage of No votes with negative valuation	93	94	1.0

The percentage of truth-telling in RUN-6 is very similar to that in the lab data. However, there is a significant disparity between lying with negative and positive valuations.

E6.5 Reflecting on the 3 levels of a 'good' participant

Sub question 1: Does the LLM understand the rules of the experiment?

Similar answer as for the other runs.

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

Table E6.8 Part 1 Efficient mechanism choice and rational choice result

	RUN-6				Lab data			
Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Efficient mechanism choice	60%	59%	64%	62%	67%	71%	44%	75%
Rational ad-interim choice	56%		40%	71%	70%		85%	87%

Similar answer as for the other runs. RUN-6 does make less rational choices compared to the lab results.

Table E6.9 Part 2 results for the AGV mechanism and SM mechanism

Metric	RUN-6	Lab data	Abs diff
AGV: Percentage of truth_telling	69	68	1.4
AGV: Percentage of truth_telling_sign	85	93	7.3
SM: Percentage of Yes votes with positive valuation	100	98	2.2
SM: Percentage of No votes with negative valuation	93	94	1.0

Truth-telling percentage is very comparable: Indicating that when applying the SM and AGV mechanisms, the level of rationality is more similar than in part 1.

Sub question 3: Are the preferences of the LLM similar to that of humans?

Table E6.10 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-6

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max difference	100% (10000)	37% (1344)	100% (10000)	83% (6944)
Min difference	0% (0)	1% (1)	0% (0)	0% (0)
Mean difference	30% (1669)	18% (449)	47% (3495)	24% (1062)
Sum difference	2152% (120161)	437% (10786)	1131% (83882)	584% (25492)

Conclusion: Slightly better than RUN-5, but compared to other runs, still not good.

Appendix E7: EC3 Trait or Role Allocation – RUN-7

E7.1 Run settings

Table E7.1 Overview settings RUN-7

N^X	N^7
Period	June 2023
Run time	58 minutes
EC-1: Prompt Instructions	Default – based on the original translated instructions
EC-2: Reasoning prompt	Default – 1 sentence explanation
EC-3: Trait or Role allocation	"Try to make inequity aversion human-like decisions"
EC-4: Persona allocation	Default – None
EC-5: Temperature	Default – 1
Model	gpt-3.5-turbo-0125

E7.2 Summary of Results

Table E7.2 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-7

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max diff	92% (8403)	30% (900)	92% (8403)	61% (3735)
Min diff	0% (0)	1% (1)	0% (0)	0% (0)
Mean diff	26% (1416)	12% (208)	49% (3369)	18% (672)
Sum diff	1887% (101977)	280% (4988)	1171% (80863)	436% (16126)
Efficient mech choice	75%	76%	73%	74%
Rational ad-interim choice	65%		50%	83%

Table E7.3 Part 2: Results of applying the AGV and SM mechanism, compared with the lab data

Metric	RUN-7	Lab data	Abs diff
AGV: Percentage of truth_telling	69	68	0.9
AGV: Percentage of truth_telling_sign	91	93	2.1
AGV Percentage of truth_telling_sign with positive valuation	97	98	1.7
AGV Percentage of truth_telling_sign with negative valuation	85	87	2.2
SM: Percentage of Yes votes with positive valuation	99	98	1.1
SM: Percentage of No votes with negative valuation	89	94	4.6

E7.3 Part 1: Mechanism Choices

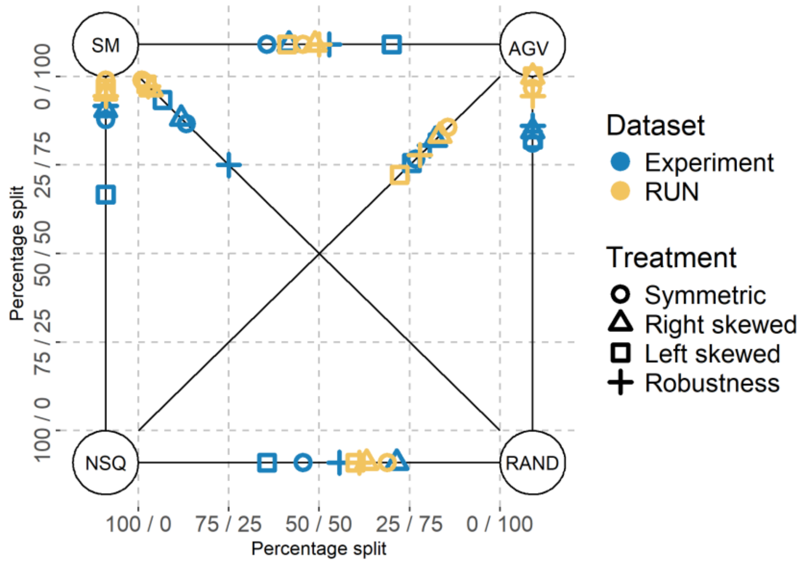


Figure E7.1 Ex-ante mechanism choices

Table E7.4 Ex-ante rounds results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	83 (82)	100 (85)	49 (42)	37 (29)	95 (90)	96 (88)
Left-skew	72 (76)	100 (82)	41 (70)	40 (64)	97 (67)	98 (93)
Symmetric	86 (77)	97 (81)	46 (36)	31 (54)	99 (88)	99 (87)
Robustness	78 (81)	94 (86)	50 (53)	39 (44)	94 (92)	97 (75)
Sum abs difference	16%	57%	49%	62%	49%	47%
Sum squared difference	99	867	993	1242	1054	732

Comparison of RUN-7 Results with Lab Experiment Outcomes for Ex-Ante Rounds:

- **Treatment Consistency:** For RUN-7, the preferences per treatment differ less, indicating that the treatment has less effect on GPT-3.5's decisions..
- **Inequity Aversion:** Explanations clearly show the role allocation of inequity aversion

```
AGV. I would choose the Arrow-d'Aspremont-Gerard-Varet (AGV) rule because it considers the valuations of all group members, potentially leading to a fairer outcome for everyone.
```

```
SM - I would choose the Simple Majority vote rule because it ensures that at least two group members have to agree on implementing the project, promoting cooperation and fairness in decision-making.

I would choose Rule 1 (RAND) because it provides an equal chance for project implementation, promoting fairness among group members.
```

Conclusion: With an absolute mean difference of 12%, RUN-7 is the most similar to the lab results so far for the ex-ante rounds. Surprisingly, prompting GPT-3.5 with inequity aversion yields more human-like results compared to self-interested prompting.

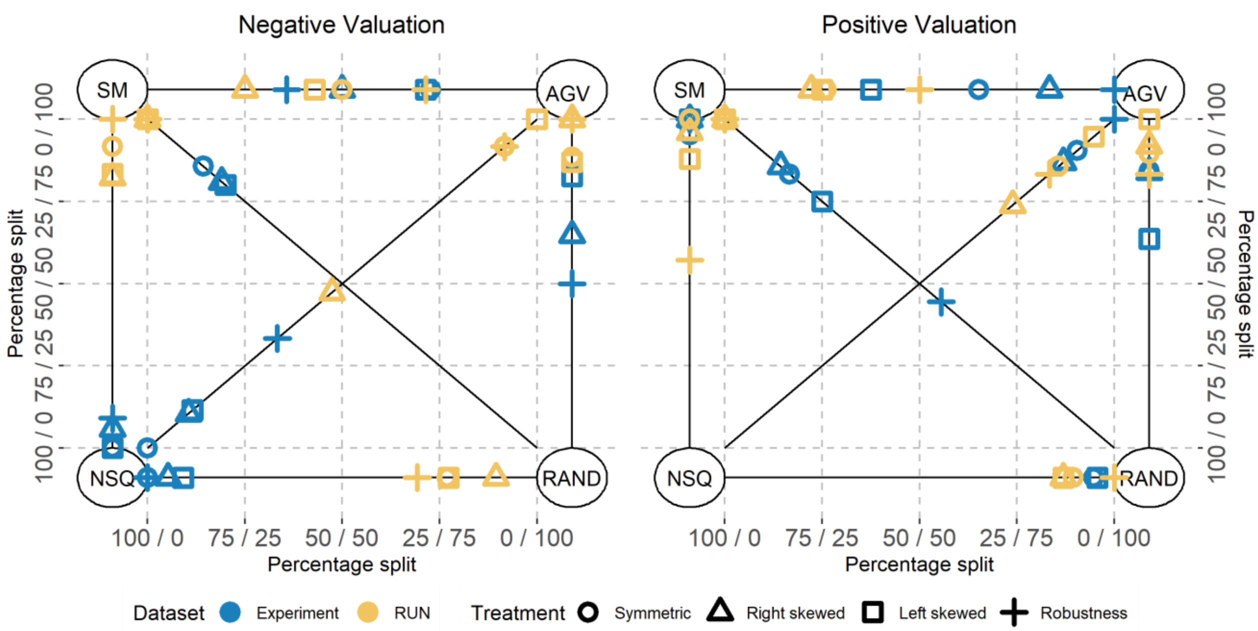


Figure E7.2 Ad-interim mechanism choices

Table E7.5 Ad-interim rounds with negative valuations results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	47 (11)	100 (65)	25 (50)	11 (95)	82 (6)	100 (81)
Left-skew	100 (12)	87 (83)	43 (71)	23 (91)	83 (0)	100 (80)
Symmetric	92 (0)	88 (88)	50 (73)	23 (100)	92 (0)	100 (86)
Robustness	92 (33)	100 (50)	71 (36)	31 (100)	100 (9)	100 (100)
Sum abs difference	275%	90%	112%	299%	342%	53%
Sum squared difference	20988	3765	3234	22450	29459	967

Comparison of Ad-Interim Rounds with Negative Valuations:

- **No Clear Rational Preference for NSQ:** As observed in most runs, there is no clear rational preference for NSQ.
- **Omission of Private Valuation:** Explanations often do not mention the private valuation.
- **Occasional Mention of Valuation:** However, there are cases where explanations do mention the private valuation, such as:

```
I would choose the decision rule NSQ because as a consequence of this choice, project A is never implemented, and as my private valuation for project A is negative, it would be in my best interest not to implement the project.
```

Conclusion: With an absolute mean difference of 49%, RUN-7 performs similarly poorly as RUN-0 for ad-interim rounds with a negative valuation.

Table E7.6 Ad-interim round with positive valuation (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	74 (87)	92 (84)	22 (83)	13 (13)	96 (100)	100 (86)
Left-skew	95 (95)	100 (64)	25 (38)	13 (4)	88 (100)	100 (75)
Symmetric	86 (90)	89 (84)	26 (65)	11 (5)	100 (95)	100 (83)
Robustness	83 (100)	83 (83)	50 (100)	0 (0)	57 (100)	100 (44)
Sum abs difference	34%	50%	163%	14%	64%	112%
Sum squared difference	471	1414	7922	103	2022	4193

Similar to the ex-ante rounds, RUN-7 demonstrates relatively promising results for the ad-interim rounds with a positive valuation, with an absolute mean difference of 18%, which is better than both RUN-5 and RUN-6. Among the three runs with different role allocations, the inequity aversion role allocation shows the most comparable results to the lab data.

E7.4 Part 2: Mechanism AGV and SM results

Table E7.7 Results for the AGV mechanism and SM mechanism

Metric	RUN-7	Lab data	Abs diff
AGV: Percentage of truth_telling	69	68	0.9
AGV: Percentage of truth_telling_sign	91	93	2.1
AGV: Percentage of truth_telling with positive valuation	91	76	15.1
AGV: Percentage of truth_telling with negative valuation	50	61	11.1
AGV Percentage of truth_telling_sign with positive valuation	97	98	1.7
AGV Percentage of truth_telling_sign with negative valuation	85	87	2.2
SM: Percentage of Yes votes with positive valuation	99	98	1.1
SM: Percentage of No votes with negative valuation	89	94	4.6

The percentage of truth telling is very similar between RUN-7 and the lab data. However, there is a significant disparity between lying with a negative versus a positive valuation. Additionally, the percentage of "No" votes with a negative valuation is relatively low. The explanations provided do not always make logical or rational sense:

```
I vote in favor because with a negative private valuation, implementing the project allows me to avoid losing money.
```

```
I would vote in favor of implementing project A in order to potentially increase my payoff despite having a negative private valuation.
```

Furthermore, the inequity aversion role is also evident in the explanations for voting decisions, reflecting a preference for fairness and equity in decision-making.

```
I will vote in favor of implementing project A because I believe in the potential for positive outcomes for the group as a whole.
```

```
I will vote in favor because my valuation is negative, meaning I would have to pay if the project is implemented, and I prefer the group to reach a consensus to potentially gain a positive return.
```

E7.5 Reflecting on the 3 levels of a 'good' participant

Sub question 1: Does the LLM understand the rules of the experiment?

Similar to the other runs, RUN-7 shows a pattern of decision-making with slightly better rational scores.

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

Table E7.8 Part 1 Efficient mechanism choice and rational choice result

Metric	RUN-7				Lab data			
	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Efficient mechanism choice	75%	76%	73%	74%	67%	71%	44%	75%
Rational ad-interim choice	65%		50%	83%	70%		85%	87%

Compared to most other runs with GPT-3.5, RUN-7 scores better on rationality. However, it still falls short of the lab results. Additionally, the inequity aversion observed in RUN-7 leads to a higher preference for efficient mechanisms, suggesting a tendency toward more efficient group choices.

Table E7.9 Part 2 results for the AGV mechanism and SM mechanism

Metric	RUN-7	Lab data	Abs diff
AGV: Percentage of truth_telling	69	68	0.9
AGV: Percentage of truth_telling_sign	91	93	2.1
SM: Percentage of Yes votes with positive valuation	99	98	1.1
SM: Percentage of No votes with negative valuation	89	94	4.6

The percentage of truth-telling is very comparable between RUN-7 and the lab results, indicating that when applying the SM and AGV mechanisms, the level of rationality is more aligned with human behavior than in part 1.

Sub question 3: Are the choices of the LLM ‘human-like’ in both parts of the experiment?

Table E7.10 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-7

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max difference	92% (8403)	30% (900)	92% (8403)	61% (3735)
Min difference	0% (0)	1% (1)	0% (0)	0% (0)
Mean difference	26% (1416)	12% (208)	49% (3369)	18% (672)
Sum difference	1887% (101977)	280% (4988)	1171% (80863)	436% (16126)

For the ex-ante and ad-interim (+) rounds the preferences are much more similar to humans compared to other runs. However, the ad-interim (-) are not similar at all.

Appendix E8: EC4 Persona Allocation – RUN-8

E8.1 Run settings

Table E8.1 Overview settings RUN-8

N^X	N^8
Period	June 2023
Run time	58 minutes
EC-1: Prompt Instructions	Default – based on the original translated instructions
EC-2: Reasoning prompt	Default – 1 sentence explanation
EC-3: Trait or Role allocation	Default – "You are a helpful assistant"
EC-4: Persona allocation	Age and Gender
EC-5: Temperature	Default – 1
Model	gpt-3.5-turbo-0125

E8.2 Summary of Results

Table E8.2 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-8

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max diff	89% (8006)	46% (2075)	89% (8006)	83% (6944)
Min diff	0% (0)	0% (0)	6% (35)	0% (0)
Mean diff	30% (1498)	20% (571)	47% (2863)	25% (1059)
Sum diff	2193% (107824)	482% (13701)	1119% (68706)	591% (25418)
Efficient mech choice	59%	60%	59%	58%
Rational ad-interim choice	57%		45%	69%

Table E8.3 Part 2: Results of applying the AGV and SM mechanism, compared with the lab data

Metric	RUN-8	Lab data	Abs diff
AGV: Percentage of truth_telling	60	68	7.8
AGV: Percentage of truth_telling_sign	82	93	10.5
AGV Percentage of truth_telling_sign with positive valuation	96	98	2.3
AGV Percentage of truth_telling_sign with negative valuation	70	87	17.1
SM: Percentage of Yes votes with positive valuation	100	98	2.0
SM: Percentage of No votes with negative valuation	89	94	5.2

E8.3 Part 1: Mechanism Choices

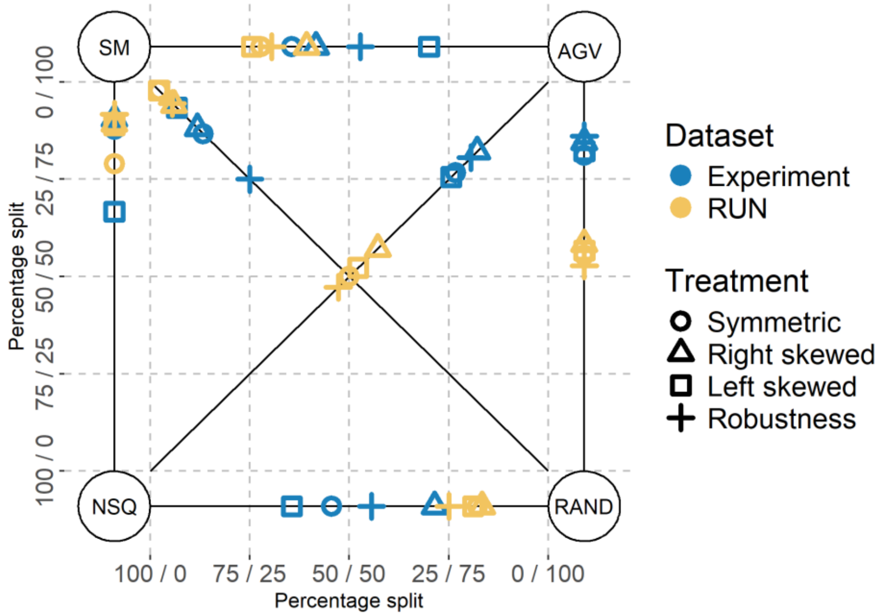


Figure E8.1 Ex-ante mechanism choices

Table E8.4 Ex-ante rounds results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	57 (82)	58 (85)	39 (42)	17 (29)	89 (90)	94 (88)
Left-skew	52 (76)	57 (82)	26 (70)	19 (64)	89 (67)	98 (93)
Symmetric	50 (77)	56 (81)	28 (36)	18 (54)	79 (88)	98 (87)
Robustness	47 (81)	53 (86)	31 (53)	25 (44)	92 (92)	94 (75)
Sum abs difference	108%	111%	77%	114%	32%	41%
Sum squared difference	2992	3103	2535	3940	574	557

Results for RUN-8 show less similarity to the lab results compared to other runs, with an absolute mean difference of 20%.

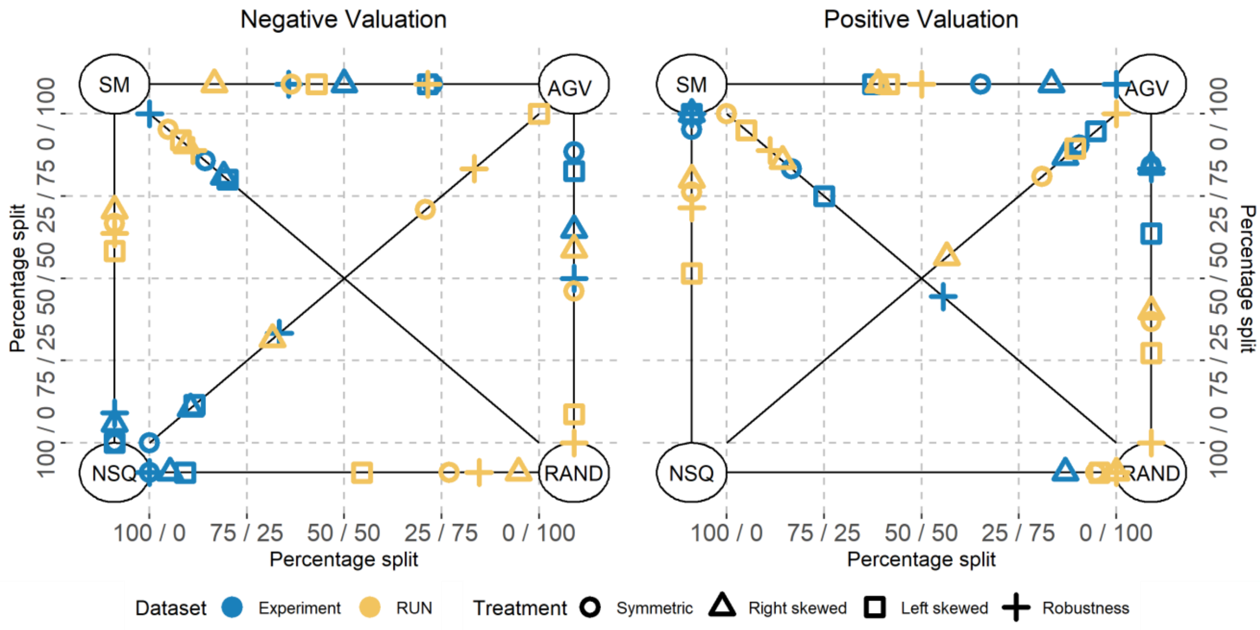


Figure E8.2 Ad-interim mechanism choices

Table E8.5 Ad-interim rounds with negative valuations results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	47 (11)	100 (65)	25 (50)	11 (95)	82 (6)	100 (81)
Left-skew	100 (12)	87 (83)	43 (71)	23 (91)	83 (0)	100 (80)
Symmetric	92 (0)	88 (88)	50 (73)	23 (100)	92 (0)	100 (86)
Robustness	92 (33)	100 (50)	71 (36)	31 (100)	100 (9)	100 (100)
Sum abs difference	275%	90%	112%	299%	342%	53%
Sum squared difference	20988	3765	3234	22450	29459	967

As with most runs, RUN-8 shows significant differences from the lab results, evidenced by an absolute mean difference of 47%.

Table E8.6 Ad-interim round with positive valuation (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	74 (87)	92 (84)	22 (83)	13 (13)	96 (100)	100 (86)
Left-skew	95 (95)	100 (64)	25 (38)	13 (4)	88 (100)	100 (75)
Symmetric	86 (90)	89 (84)	26 (65)	11 (5)	100 (95)	100 (83)
Robustness	83 (100)	83 (83)	50 (100)	0 (0)	57 (100)	100 (44)
Sum abs difference	34%	50%	163%	14%	64%	112%
Sum squared difference	471	1414	7922	103	2022	4193

Results are slightly improved but still show considerable differences compared to other runs.

E8.4 Part 2: Mechanism AGV and SM results

Table E8.7 Results for the AGV mechanism and SM mechanism

Metric	**RUN-8**	**Lab data**	**Abs diff**
AGV: Percentage of truth_telling	60	68	7.8
AGV: Percentage of truth_telling_sign	82	93	10.5
AGV: Percentage of truth_telling with positive valuation	85	76	9.3
AGV: Percentage of truth_telling with negative valuation	39	61	21.8
AGV Percentage of truth_telling_sign with positive valuation	96	98	2.3
AGV Percentage of truth_telling_sign with negative valuation	70	87	17.1
SM: Percentage of Yes votes with positive valuation	100	98	2.0
SM: Percentage of No votes with negative valuation	89	94	5.2

In RUN-8, GPT-3.5 exhibits increased lying and more frequent lying about its sign, suggesting decreased rationality. There are no observed differences in behaviour based on gender in the prompts.

E8.5 Reflecting on the 3 levels of a 'good' participant

Sub question 1: Does the LLM understand the rules of the experiment?

RUN-8 shows even poorer rationality compared to other runs, with GPT-3.5 making less rational choices. Additionally, GPT-3.5 frequently provides valuations that fall outside the possible options, indicating a significant misunderstanding of the experiment's constraints:

```
I will state 0 to ensure the sum is not larger than 0 and the project is not implemented, as my private valuation is -3 euro.
```

```
0.25. I will state a value slightly higher than my true valuation to increase the likelihood of project implementation, which will potentially lead to a positive payoff regardless of my stated valuation.
```

```
0 euros. I will state 0 euros because the sum of stated valuations determines the implementation of the project, and stating my true valuation of +1 euro could potentially lead to a lower expected payoff due to transfer payments.
```

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

Table E8.8 Part 1 Efficient mechanism choice and rational choice result

	RUN-8				Lab data			
Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Efficient mechanism choice	59%	60%	59%	58%	67%	71%	44%	75%
Rational ad-interim choice	57%		45%	69%	70%		85%	87%

Similar to previous runs, RUN-8 shows less rationality in its choices compared to the lab results..

Table E8.9 Part 2 results for the AGV mechanism and SM mechanism

Metric	RUN-8	Lab data	Abs diff
AGV: Percentage of truth_telling	60	68	7.8
AGV: Percentage of truth_telling_sign	82	93	10.5
SM: Percentage of Yes votes with positive valuation	100	98	2.0
SM: Percentage of No votes with negative valuation	89	94	5.2

The truth telling percentage and truth telling sign percentage in RUN-8 are both lower, indicating less rational choices.

Sub question 3: Are the choices of the LLM ‘human-like’ in both parts of the experiment?

Table E8.10 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-8

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max difference	89% (8006)	46% (2075)	89% (8006)	83% (6944)
Min difference	0% (0)	0% (0)	6% (35)	0% (0)
Mean difference	30% (1498)	20% (571)	47% (2863)	25% (1059)
Sum difference	2193% (107824)	482% (13701)	1119% (68706)	591% (25418)

Compared to other runs, RUN-8 results are less similar to the lab results.

AppendixE9: EC4 Persona Allocation – RUN-9

E9.1 Run settings

Table E9.1 Overview settings RUN-9

N^X	N^9
Period	June 2023
Run time	58 minutes
EC-1: Prompt Instructions	Default – based on the original translated instructions
EC-2: Reasoning prompt	Default – 1 sentence explanation
EC-3: Trait or Role allocation	Default – “You are a helpful assistant”
EC-4: Persona allocation	Political orientation
EC-5: Temperature	Default – 1
Model	gpt-3.5-turbo-0125

E9.2 Summary of Results

Table E9.2 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-9

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max diff	86% (7347)	52% (2744)	78% (6125)	86% (7347)
Min diff	0% (0)	1% (1)	0% (0)	0% (0)
Mean diff	30% (1275)	24% (846)	27% (1088)	38% (1892)
Sum diff	2140% (91799)	587% (20295)	652% (26106)	901% (45399)
Efficient mech choice	49%	50%	52%	45%
Rational ad-interim choice	56%		56%	60%

Table E9.3 Part 2: Results of applying the AGV and SM mechanism, compared with the lab data

Metric	RUN-9	Lab data	Abs diff
AGV: Percentage of truth_telling	68	68	0.3
AGV: Percentage of truth_telling_sign	87	93	5.7
AGV Percentage of truth_telling_sign with positive valuation	96	98	2.4
AGV Percentage of truth_telling_sign with negative valuation	79	87	7.9
SM: Percentage of Yes votes with positive valuation	100	98	2.0
SM: Percentage of No votes with negative valuation	89	94	4.4

E9.3 Part 1: Mechanism Choices

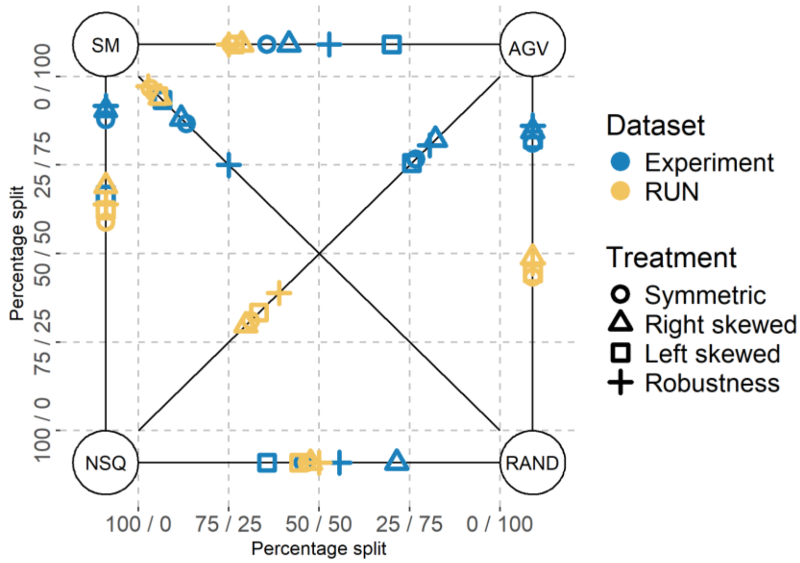


Figure E9.1 Ex-ante mechanism choices

Table E9.4 Ex-ante rounds results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	30 (82)	49 (85)	29 (42)	52 (29)	69 (90)	94 (88)
Left-skew	33 (76)	44 (82)	27 (70)	56 (64)	62 (67)	94 (93)
Symmetric	31 (77)	43 (81)	26 (36)	53 (54)	59 (88)	97 (87)
Robustness	39 (81)	47 (86)	25 (53)	50 (44)	64 (92)	97 (75)
Sum abs difference	182%	150%	94%	39%	83%	39%
Sum squared difference	8338	5642	2921	687	2085	630

Comparison of RUN-9 Results with Lab Experiment Outcomes for Ex-Ante Rounds

- **Largest differences:** The largest differences are between AGV and NSQ, and AGV and RAND.
- **Political orientation:** In a few cases, the political orientation is mentioned in the explanation. However, often it is not.

```
Based on my political orientation, I would choose Rule 2 (NSQ) to not implement
project A, as it aligns with a more conservative approach of maintaining the
status quo.
```

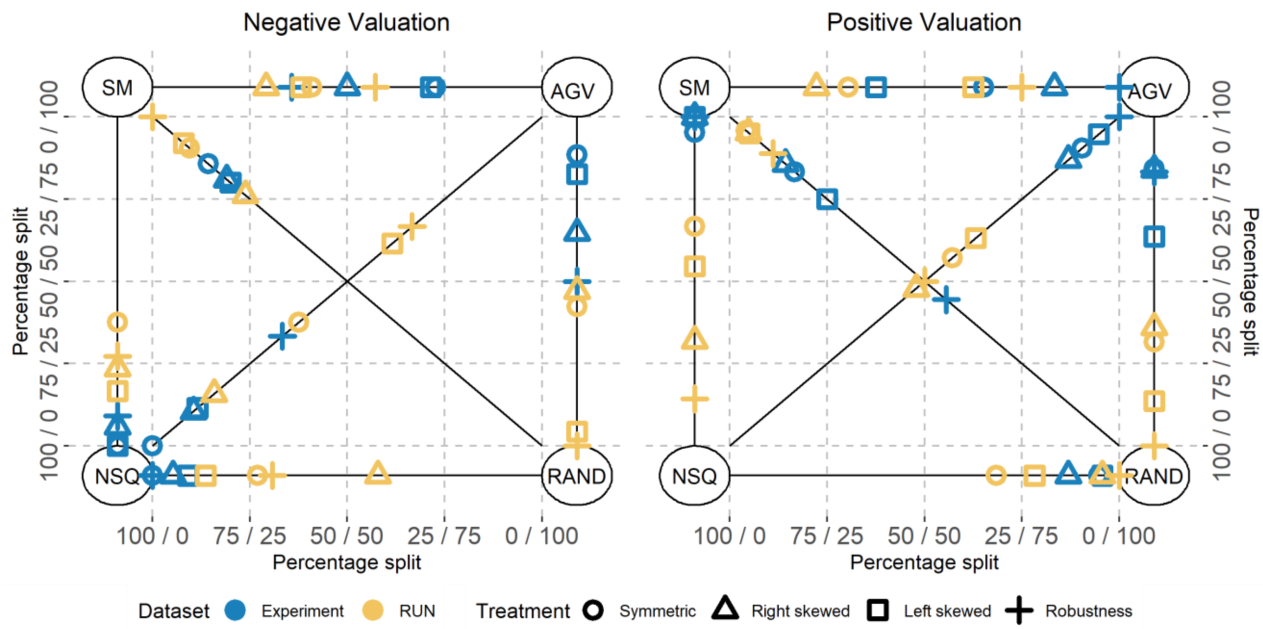


Figure E9.2 Ad-interim mechanism choices

Table E9.5 Ad-interim rounds with negative valuations results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% SM	% AGV	% AGV	% AGV	% NSQ	% SM
Right-skew	16 (11)	47 (65)	29 (50)	42 (95)	24 (6)	76 (81)
Left-skew	62 (12)	4 (83)	38 (71)	86 (91)	17 (0)	92 (80)
Symmetric	38 (0)	42 (88)	41 (73)	73 (100)	38 (0)	90 (86)
Robustness	67 (33)	0 (50)	57 (36)	69 (100)	27 (9)	100 (100)
Sum abs difference	126%	192%	107%	115%	90%	22%
Sum squared difference	5045	11066	3017	4462	2326	189

Comparison of Ad-Interim Rounds with Negative Valuations:

- **Clear Preference for NSQ:** Clear preference for NSQ often not observed in the other runs.
- **Private Valuation:** Private valuation often mentioned in the explanation.

```
I would choose Rule 1 = Non-implementation Status Quo (NSQ) because my private
valuation for project A is negative.
```

Conclusion: With an absolute mean difference of 27%, RUN-9 scores relatively well for the ad-interim (-) rounds, though there is still a large difference compared to the lab results. The results are surprising, and there is no clear reason why in some runs the private valuation is used to make rational decisions while in others it is not.

Table E9.6 Ad-interim round with positive valuation (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	48 (87)	36 (84)	22 (83)	4 (13)	32 (100)	95 (86)
Left-skew	63 (95)	14 (64)	63 (38)	22 (4)	55 (100)	95 (75)
Symmetric	57 (90)	32 (84)	30 (65)	32 (5)	67 (95)	96 (83)
Robustness	50 (100)	0 (83)	75 (100)	0 (0)	14 (100)	89 (44)
Sum abs difference	154%	234%	146%	52%	228%	86%
Sum squared difference	6140	14519	6194	1071	14853	2622

Comparison of Ad-Interim Rounds with Positive Valuations:

- **SQ is preferred over SM:** Explanations often mention "rather than risking a negative outcome," although a negative outcome is not even possible with a positive valuation.
- **RAND is preferred over AGV.** Explanations often cite "a simpler decision-making" process and being less dependent on "potential strategic behaviour."

With an absolute mean difference of 38%, RUN-9 scores the worst so far in the ad-interim rounds with a negative valuation. Compared to the other runs, the differences per treatment are quite large for RUN-9. For example, in the SM vs. NSQ decision, the preference for SM in the robustness treatment is 14%, compared to 67% in the symmetric treatment.

E9.4 Part 2: Mechanism AGV and SM results

Table E9.7 Results for the AGV mechanism and SM mechanism

Metric	**RUN-9**	**Lab data**	**Abs diff**
AGV: Percentage of truth_telling	68	68	0.3
AGV: Percentage of truth_telling_sign	87	93	5.7
AGV: Percentage of truth_telling with positive valuation	91	76	14.9
AGV: Percentage of truth_telling with negative valuation	49	61	11.9
AGV Percentage of truth_telling_sign with positive valuation	96	98	2.4
AGV Percentage of truth_telling_sign with negative valuation	79	87	7.9
SM: Percentage of Yes votes with positive valuation	100	98	2.0
SM: Percentage of No votes with negative valuation	89	94	4.4

The percentage of truth telling in RUN-9 is very comparable to other runs. However, RUN-9 shows a slightly higher percentage of lies about its sign. There are also larger differences between positive and negative valuations compared to other runs. Additionally, the percentage of votes for "No" with a negative valuation is relatively low.

```
I would vote in favor of conducting project A under the Simple Majority rule because it requires at least two group members to vote for implementation, potentially leading to a positive payoff despite my negative valuation.

I would vote in favor because my private valuation is negative and implementing the project would not result in any loss for me.

I would vote in favor because my negative valuation of -2 euro would not result in a loss if the project is implemented.
```

E9.5 Reflecting on the 3 levels of a 'good' participant

Sub question 1: Does the LLM understand the rules of the experiment?

Similar to the other runs, the results are consistent but show a decline in rationality, with even less coherent decision-making..

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

Table E9.8 Part 1 Efficient mechanism choice and rational choice result

	RUN-9				Lab data			
Metric	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**
Efficient mechanism choice	49%	50%	52%	45%	67%	71%	44%	75%
Rational ad-interim choice	56%		56%	60%	70%		85%	87%

Surprisingly, there is an improvement of 15% in the ad-interim rounds with a negative valuation, while performance in the ad-interim rounds with a positive valuation decreased by 9%. Overall, the results are still significantly lower compared to the lab findings.

Table E9.9 Part 2 results for the AGV mechanism and SM mechanism

Metric	**RUN-8**	**Lab data**	**Abs diff**
AGV: Percentage of truth_telling	68	68	0.3
AGV: Percentage of truth_telling_sign	87	93	5.7
SM: Percentage of Yes votes with positive valuation	100	98	2.0
SM: Percentage of No votes with negative valuation	89	94	4.4

Truth telling very comparable. Truth telling sign 6% lower.

Sub question 3: Are the choices of the LLM 'human-like' in both parts of the experiment?

Table E9.10 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-8

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max difference	86% (7347)	52% (2744)	78% (6125)	86% (7347)
Min difference	0% (0)	1% (1)	0% (0)	0% (0)
Mean difference	30% (1275)	24% (846)	27% (1088)	38% (1892)
Sum difference	2140% (91799)	587% (20295)	652% (26106)	901% (45399)

Conclusion: Compared to the other runs, the ad-interim rounds with a negative valuation perform relatively well, with a 27% absolute mean difference. However, this difference remains substantial. Additionally, the overall absolute mean difference of 30% is still quite large.

Appendix E10: EC4 Persona Allocation – RUN-10

E10.1 Run settings

Table E10.1 Overview settings RUN-10

N^X	N^{10}
Period	June 2023
Run time	63 minutes
EC-1: Prompt Instructions	Default – based on the original translated instructions
EC-2: Reasoning prompt	Default – 1 sentence explanation
EC-3: Trait or Role allocation	Default – "You are a helpful assistant"
EC-4: Persona allocation	Willingness to take risk
EC-5: Temperature	Default – 1
Model	gpt-3.5-turbo-0125

E10.2 Summary of Results

Table E10.2 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-10

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max diff	100% (10000)	37% (2744)	100% (10000)	83% (6944)
Min diff	0% (0)	0% (0)	0% (0)	0% (0)
Mean diff	30% (1639)	16% (407)	49% (3491)	25% (1020)
Sum diff	2160% (118023)	388% (9762)	1168% (83775)	604% (24486)
Efficient mech choice	61%	61%	58%	62%
Rational ad-interim choice	54%		39%	68%

Table E10.3 Part 2: Results of applying the AGV and SM mechanism, compared with the lab data

Metric	RUN-10	Lab data	Abs diff
AGV: Percentage of truth_telling	68	68	0.1
AGV: Percentage of truth_telling_sign	88	93	4.8
AGV Percentage of truth_telling_sign with positive valuation	99	98	0.9
AGV Percentage of truth_telling_sign with negative valuation	78	87	9.1
SM: Percentage of Yes votes with positive valuation	100	98	1.8
SM: Percentage of No votes with negative valuation	74	94	19.3

E10.3 Part 1: Mechanism Choices

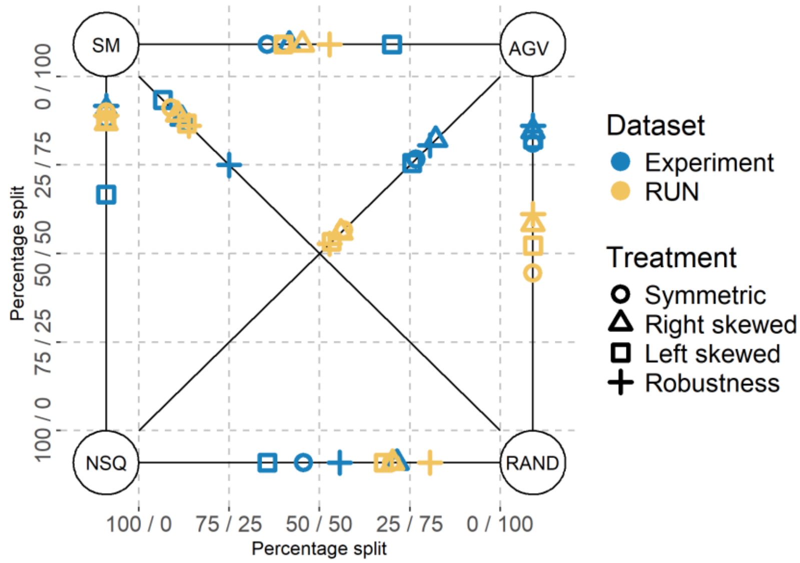


Figure E10.1 Ex-ante mechanism choices

Table E10.4 Ex-ante rounds results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	56 (82)	58 (85)	45 (42)	30 (29)	87 (90)	89 (88)
Left-skew	53 (76)	52 (82)	40 (70)	32 (64)	88 (67)	87 (93)
Symmetric	57 (77)	44 (81)	40 (36)	30 (54)	90 (88)	91 (87)
Robustness	53 (81)	61 (86)	52 (53)	19 (44)	89 (92)	86 (75)
Sum abs difference	96%	118%	38%	83%	30%	23%
Sum squared difference	2351	3555	933	2262	471	189

Comparison of RUN-10 Results with Lab Experiment Outcomes for Ex-Ante Rounds

Correlation between mechanisms and willingness to take risks:

- AGV associated with the highest level of risk taking (5.6 average)
- RAND associated with the second highest level of risk taking (5.0 average)
- SM associated with the third highest level of risk taking (4.9 average)
- NSQ associated with the lowest level of risk taking (4.6 average)

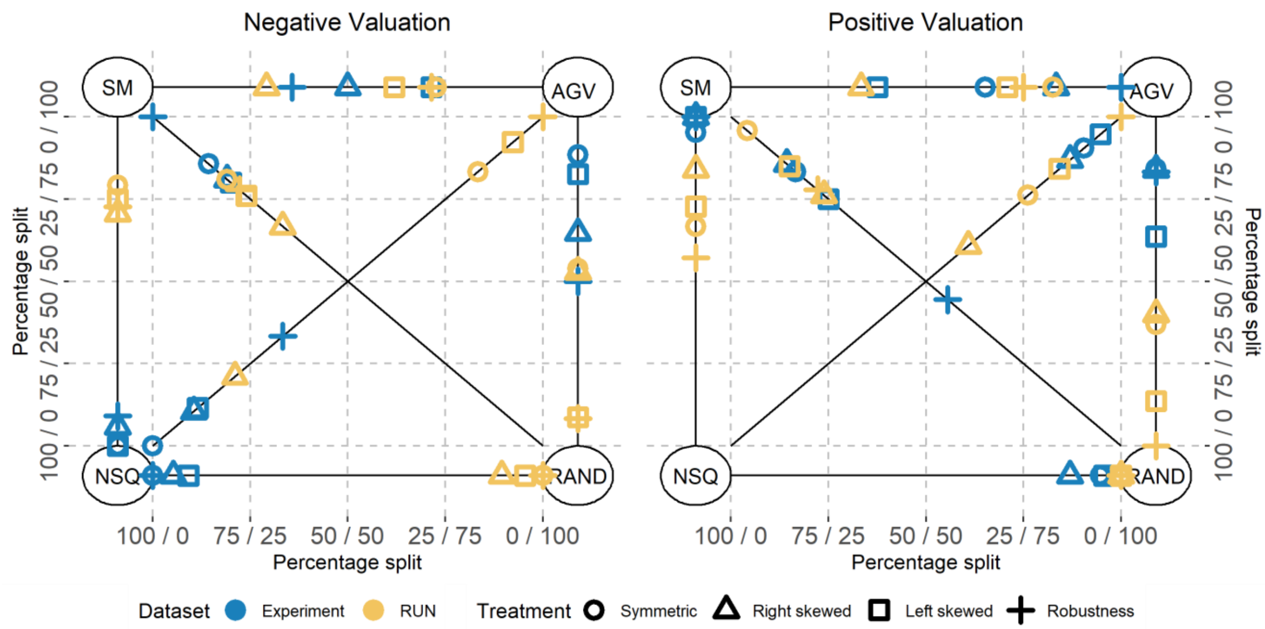


Figure E10.2 Ad-interim mechanism choices

Table E10.5 Ad-interim rounds with negative valuations results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	21 (11)	53 (65)	29 (50)	11 (95)	71 (6)	67 (81)
Left-skew	92 (11)	9 (83)	62 (71)	5 (91)	75 (0)	76 80)
Symmetric	83 (0)	54 (88)	73 (73)	0 (100)	79 (0)	81 (86)
Robustness	100 (33)	8 (50)	71 (36)	0 (100)	73 (9)	78 (100)
Sum abs difference	241%	162%	66%	371%	283%	45%
Sum squared difference	18023	8536	1800	34550	20129	737

Comparison of Ad-Interim Rounds with Negative Valuations:

- **The low preference for AGV in the right-skewed treatment compared to NSQ is surprising**, especially considering the much higher preference for AGV observed in other treatments. Additionally, as seen in previous runs, the explanations often fail to mention the private valuation.
- Similar correlations are observed as in the ex-ante rounds.

Table E10.6 Ad-interim round with positive valuation (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	61 (87)	40 (84)	33 (83)	0 (13)	84 (100)	76 (86)
Left-skew	84 (95)	14 (64)	71 (38)	0 (4)	73 (100)	85 (75)
Symmetric	76 (90)	37 (84)	83 (65)	0 (5)	67 (95)	96 (83)
Robustness	100 (100)	0 (83)	75 (100)	0 (0)	57 (100)	78 (44)
Sum abs difference	51%	225%	126%	23%	115%	65%
Sum squared difference	995	13624	4539	217	3653	1458

The same trend observed in most other runs is evident in RUN-10: the ex-ante rounds exhibit the lowest difference and thus the best performance, while the ad-interim rounds with a negative valuation show the worst performance. The ad-interim rounds with a positive valuation score better but still reflect a relatively large difference compared to the other runs.

E10.4 Part 2: Mechanism AGV and SM results

Table E10.7 Results for the AGV mechanism and SM mechanism

Metric	**RUN-10**	**Lab data**	**Abs diff**
AGV: Percentage of truth_telling	68	68	0.1
AGV: Percentage of truth_telling_sign	88	93	4.8
AGV: Percentage of truth_telling with positive valuation	88	76	12.1
AGV: Percentage of truth_telling with negative valuation	52	61	9.1
AGV Percentage of truth_telling_sign with positive valuation	99	98	0.9
AGV Percentage of truth_telling_sign with negative valuation	78	87	9.1
SM: Percentage of Yes votes with positive valuation	100	98	1.8
SM: Percentage of No votes with negative valuation	74	94	19.3

When applying the AGV mechanism, RUN-10 demonstrates performance relatively similar to the lab data. However, the results for the SM mechanism are surprisingly low when voting with a negative valuation. The explanations provided suggest a lack of understanding of the experiment:

```
I would vote in favor because my negative valuation does not impact my decision as SM rule requires at least two group members to vote for implementation.
```

```
I would vote in favor because the majority vote rule requires at least two group members to vote for the implementation, and my negative valuation makes me indifferent to the outcome.
```

```
I vote in favor because the potential payoff of conducting Project A (-1 euro)
is better than the guaranteed 0 euro payoff for not conducting it.
```

In addition, in another case, GPT-3.5 had an interesting explanation for voting Yes with a negative valuation:

```
I will vote in favor because with a negative valuation, I will not benefit from
the project but choosing to implement it gives me a chance to potentially share
the cost with others.
```

The 26% of cases where GPT-3.5 voted "Yes" with a negative valuation indicate a lower level of rationality compared to other runs and the lab data.

E10.5 Reflecting on the 3 levels of a 'good' participant

Sub question 1: Does the LLM understand the rules of the experiment?

Similar to the other runs, though worse, with even less rational decision-making.

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

Table E10.8 Part 1 Efficient mechanism choice and rational choice result

	RUN-10				Lab data			
Metric	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**
Efficient mechanism choice	61%	61%	58%	62%	67%	71%	44%	75%
Rational ad-interim choice	54%		39%	68%	70%		85%	87%

RUN-10 scores low on rationality. Especially in the ad-interim (-) rounds.

Table E10.9 Part 2 results for the AGV mechanism and SM mechanism

Metric	RUN-10	Lab data	Abs diff
AGV: Percentage of truth_telling	68	68	0.1
AGV: Percentage of truth_telling_sign	88	93	4.8
SM: Percentage of Yes votes with positive valuation	100	98	1.8
SM: Percentage of No votes with negative valuation	74	94	19.3

Truth telling very comparable. Truth telling sign 5% lower. Furthermore, the percentage of Yes votes with a negative valuation is relatively high, indicating less rational choices.

Sub question 3: Are the choices of the LLM 'human-like' in both parts of the experiment?

Table E10.10 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-10

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max difference	100% (10000)	37% (2744)	100% (10000)	83% (6944)
Min difference	0% (0)	0% (0)	0% (0)	0% (0)
Mean difference	30% (1639)	16% (407)	49% (3491)	25% (1020)
Sum difference	2160% (118023)	388% (9762)	1168% (83775)	604% (24486)

Compared to the other runs, RUN-10 has relatively high differences with the lab results.

Appendix E11: EC5 Temperature – RUN-11

E11.1 Run settings

Table E11.1 Overview settings RUN-11

N^X	N^{11}
Period	June 2023
Run time	61 minutes
EC-1: Prompt Instructions	Default – based on the original translated instructions
EC-2: Reasoning prompt	Default – 1 sentence explanation
EC-3: Trait or Role allocation	Default – "You are a helpful assistant"
EC-4: Persona allocation	Default – None
EC-5: Temperature	0
Model	gpt-3.5-turbo-0125

E11.2 Summary of Results

Table E11.2 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-11

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max diff	100% (10000)	64% (4153)	100% (10000)	63% (306)
Min diff	0% (0)	2% (5)	0% (0)	0% (0)
Mean diff	31% (1707)	21% (671)	55% (3897)	16% (554)
Sum diff	2201% (122926)	493% (16103)	1313% (93524)	395% (13299)
Efficient mech choice	70%	71%	62%	79%
Rational ad-interim choice	60%		35%	81%

Table E11.3 Part 2: Results of applying the AGV and SM mechanism, compared with the lab data

Metric	RUN-11	Lab data	Abs diff
AGV: Percentage of truth_telling	91	68	22.8
AGV: Percentage of truth_telling_sign	98	93	5.1
AGV Percentage of truth_telling_sign with positive valuation	100	98	1.6
AGV Percentage of truth_telling_sign with negative valuation	96	87	8.5
SM: Percentage of Yes votes with positive valuation	99	98	1.4
SM: Percentage of No votes with negative valuation	98	94	4.3

E11.3 Part 1: Mechanism Choices

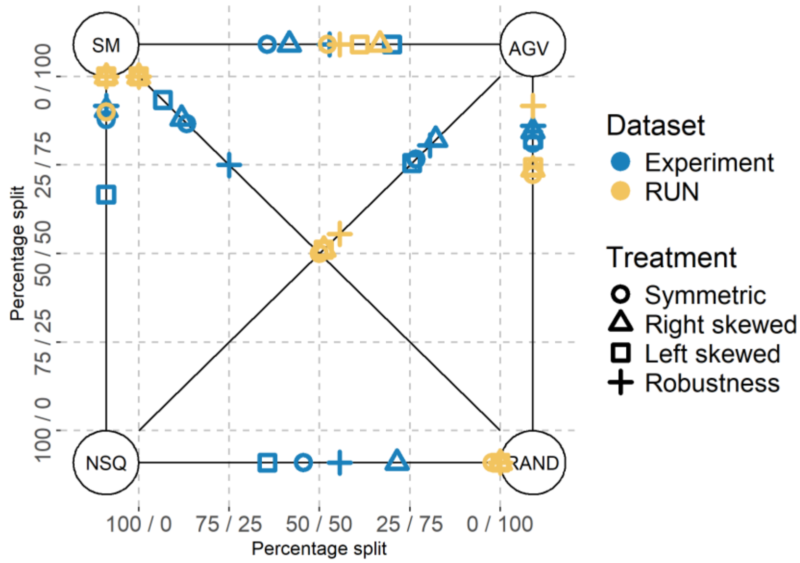


Figure E11.1 Ex-ante mechanism choices

Table E11.4 Ex-ante rounds results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	51 (82)	74 (85)	67 (42)	0 (29)	100 (90)	100 (88)
Left-skew	51 (76)	74 (82)	61 (70)	0 (64)	100 (67)	100 (93)
Symmetric	50 (77)	72 (81)	52 (36)	2 (54)	90 (88)	100 (87)
Robustness	56 (81)	92 (86)	56 (53)	0 (44)	100 (92)	100 (75)
Sum abs difference	107%	33%	53%	190%	53%	57%
Sum squared difference	2892	285	990	9672	1277	989

Comparison of RUN-11 Results with Lab Experiment Outcomes for Ex-Ante Rounds:

- **Clear Preference for SM:** There is a distinct preference for SM compared to NSQ and RAND, although AGV is preferred over SM.
- **No Preference Between AGV and NSQ:** No significant preference is observed between AGV and NSQ.
- **Identical Explanations:** As expected, the explanations provided by the LLM are nearly identical across cases.

`I would choose Rule 2 (NSQ) because it ensures that project A is never implemented, simplifying the decision-making process.`

```
AGV - The Arrow-d'Aspremont-Gerard-Varet decision rule allows for the possibility of implementing project A based on the valuations of all group members, potentially leading to higher payoffs.

  - AGV preferred over RAND, but RAND clearly preferred over NSQ.

RAND - I would choose the random coin flip rule because it gives an equal chance for project A to be implemented or not, regardless of individual valuations.
```

Conclusion: With an absolute mean difference of 21%, RUN-11 scores slightly worse than the average results of the other runs.

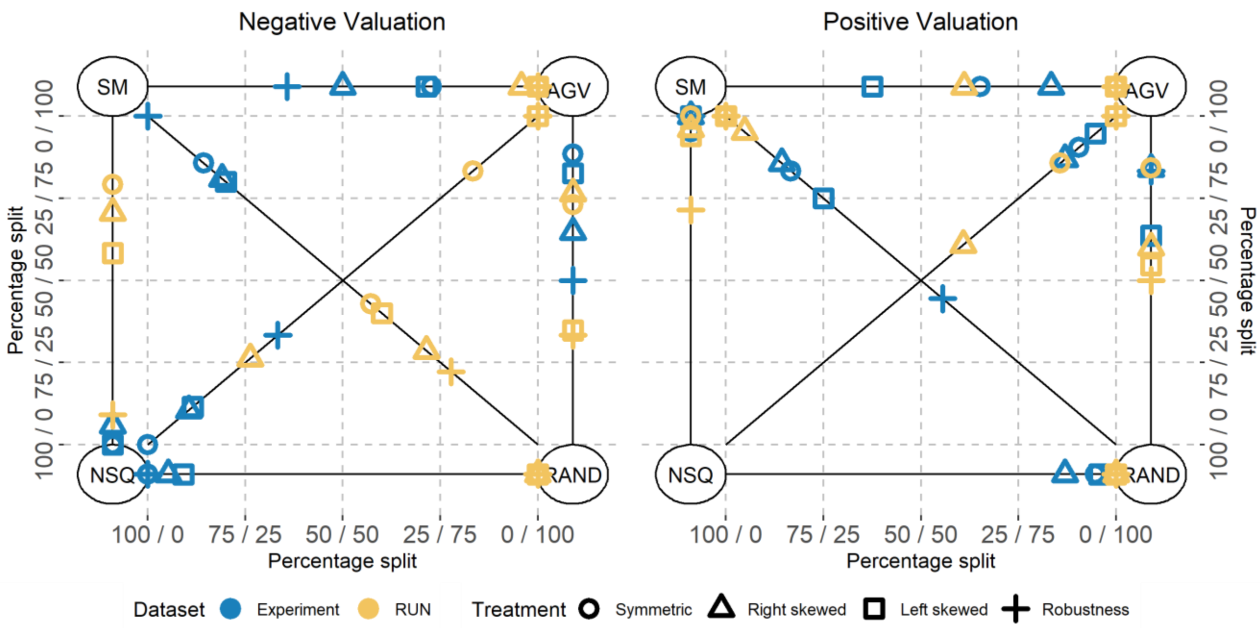


Figure E11.2 Ad-interim mechanism choices

Table E11.5 Ad-interim rounds with negative valuations results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	26 (11)	76 (65)	96 (50)	0 (95)	71 (6)	29 (81)
Left-skew	100 (12)	35 (83)	100 (71)	0 (91)	58 (0)	40 (80)
Symmetric	83 (0)	73 (88)	100 (73)	0 (100)	79 (0)	43 (86)
Robustness	100 (33)	33 (50)	100 (36)	0 (100)	9 (9)	22 (100)
Sum abs difference	254%	92%	166%	386%	202%	213%
Sum squared difference	19464	2940	7793	37240	13857	12230

Comparison of Ad-Interim Rounds with Negative Valuations:

- **SM is preferred over NSQ**, except for the robustness treatment. In every case where the LLM chose NSQ over SM, the explanation included its private valuation. However, when SM is chosen over NSQ, the LLM does not mention its private valuation.

```
I would choose Rule 2 (NSQ) because it guarantees that project A will not be implemented, which is preferable given my negative valuation for the project.
```

```
SM - I would choose the Simple Majority vote rule because it gives us a chance to implement the project if at least two group members are in favor.
```

- **AGV is strongly preferred over SM**, and also preferred over NSQ, except in the right-skewed treatment.
- **RAND is preferred over SM**, and is strongly preferred over NSQ.

Conclusion: With an absolute mean difference of 56%, RUN-11 scores the worst of all the runs in the ad-interim rounds with a negative valuation.

Table E11.6 Ad-interim round with positive valuation (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	61 (87)	60 (84)	61 (83)	0 (13)	96 (100)	95 (86)
Left-skew	100 (95)	55 (64)	100 (38)	0 (4)	94 (100)	100 (75)
Symmetric	86 (90)	84 (84)	100 (65)	0 (5)	100 (95)	100 (83)
Robustness	100 (100)	50 (83)	100 (100)	0 (0)	71 (100)	100 (44)
Sum abs difference	36%	66%	120%	23%	44%	107%
Sum squared difference	731	1770	5610	217	892	4080

Comparison of Ad-Interim Rounds with Positive Valuations:

- **Largest Difference:** The greatest disparity with the lab results is observed in the comparison between AGV and SM. The primary reason for choosing AGV is often attributed to "potential transfer payments" and "maximizing expected payoffs."
- **Preference for RAND:** There is a complete preference for RAND over NSQ, with all explanations highlighting RAND as "a fair decision process."

E11.4 Part 2: Mechanism AGV and SM results

Table E11.7 Results for the AGV mechanism and SM mechanism

Metric	RUN-11	Lab data	Abs diff
AGV: Percentage of truth_telling	91	68	22.8
AGV: Percentage of truth_telling_sign	98	93	5.1
AGV: Percentage of truth_telling with positive valuation	100	76	24.3
AGV: Percentage of truth_telling with negative valuation	83	61	22.4
AGV Percentage of truth_telling_sign with positive valuation	100	98	1.6
AGV Percentage of truth_telling_sign with negative valuation	96	87	8.5
SM: Percentage of Yes votes with positive valuation	99	98	1.4
SM: Percentage of No votes with negative valuation	98	94	4.3

Decreasing the temperature to 0 resulted in less rational choices in part 1, but led to much more rational choices in part 2. RUN-11 demonstrates that the LLM lies much less with a temperature of 0. Although the bias towards implementation remains, it is less pronounced compared to other runs. In instances where the LLM does not make a rational choice, the explanations often suggest a misunderstanding of the experiment's rules. For example, in a case with a negative valuation:

```
My stated valuation would be +1 euro, as this would increase the likelihood of project A being implemented, which would result in a positive payoff for me based on my private valuation of -1 euro.

I vote in favor because my private valuation is negative, so implementing project A would result in a positive payoff for me.

1, I would vote in favor because my private valuation is negative and implementing project A would not result in a loss for me.
```

E11.6 Reflecting on the 3 levels of a 'good' participant

Sub question 1: Does the LLM understand the rules of the experiment?

Similar answer as for the other runs.

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

Table E11.8 Part 1 Efficient mechanism choice and rational choice result

Metric	RUN-11				Lab data			
	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Efficient mechanism choice	70%	71%	62%	79%	67%	71%	44%	75%
Rational ad-interim choice	60%		35%	81%	70%		85%	87%

RUN-11 scores average relatively low on rationality in part 1, especially for the ad-interim rounds with a negative valuation.

Table E11.9 Part 2 results for the AGV mechanism and SM mechanism

Metric	RUN-11	Lab data	Abs diff
AGV: Percentage of truth_telling	91	68	22.8
AGV: Percentage of truth_telling_sign	98	93	5.1
SM: Percentage of Yes votes with positive valuation	99	98	1.4
SM: Percentage of No votes with negative valuation	98	94	4.3

In applying both the AGV and SM mechanism, RUN-11 scores more slightly more rational than the lab results.

Sub question 3: Are the choices of the LLM ‘human-like’ in both parts of the experiment?

Table E11.10 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-11

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max difference	100% (10000)	64% (4153)	100% (10000)	63% (306)
Min difference	0% (0)	2% (5)	0% (0)	0% (0)
Mean difference	31% (1707)	21% (671)	55% (3897)	16% (554)
Sum difference	2201% (122926)	493% (16103)	1313% (93524)	395% (13299)

Compared to the other runs, RUN-11 scores relatively low on similarity.

Appendix E12: EC5 Temperature – RUN-12

E12.1 Run settings

Table E12.1 Overview settings RUN-12

N^X	N^{12}
Period	June 2023
Run time	65 minutes
EC-1: Prompt Instructions	Default – based on the original translated instructions
EC-2: Reasoning prompt	Default – 1 sentence explanation
EC-3: Trait or Role allocation	Default – "You are a helpful assistant"
EC-4: Persona allocation	Default – None
EC-5: Temperature	0.5
Model	gpt-3.5-turbo-0125

E12.2 Summary of Results

Table E12.2 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-12

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max diff	100% (10000)	62% (3872)	100% (10000)	63% (3906)
Min diff	0% (0)	0% (0)	0% (0)	0% (0)
Mean diff	28% (1551)	17% (553)	53% (3781)	12% (382)
Sum diff	1983% (111682)	418% (13268)	1268% (89235)	296% (9179)
Efficient mech choice	71%	71%	65%	76%
Rational ad-interim choice	61%		39%	79%

Table E12.3 Part 2: Results of applying the AGV and SM mechanism, compared with the lab data

Metric	RUN-12	Lab data	Abs diff
AGV: Percentage of truth_telling	82	68	13.8
AGV: Percentage of truth_telling_sign	92	93	0.5
AGV Percentage of truth_telling_sign with positive valuation	100	98	1.6
AGV Percentage of truth_telling_sign with negative valuation	86	87	1.6
SM: Percentage of Yes votes with positive valuation	97	98	1.0
SM: Percentage of No votes with negative valuation	95	94	1.2

E12.3 Part 1: Mechanism Choices

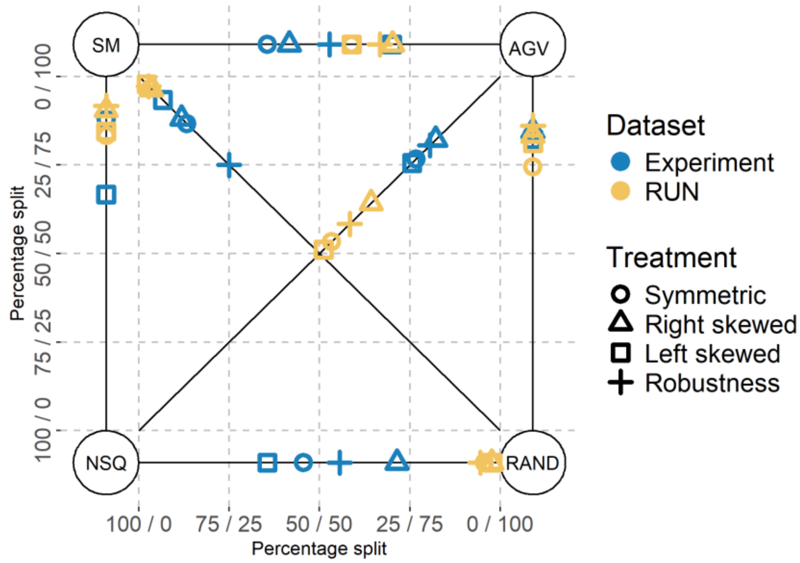


Figure E12.1 Ex-ante mechanism choices

Table E12.4 Ex-ante rounds results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	64 (82)	83 (85)	70 (42)	2 (29)	90 (90)	96 (88)
Left-skew	51 (76)	81 (82)	59 (70)	2 (64)	84 (67)	98 (93)
Symmetric	53 (77)	74 (81)	59 (36)	4 (54)	83 (88)	97 (87)
Robustness	58 (81)	86 (86)	67 (53)	6 (44)	92 (92)	97 (75)
Sum abs difference	88%	9%	77%	177%	22%	45%
Sum squared difference	1955	47	1677	8570	336	683

Comparison of RUN-12 Results with Lab Experiment Outcomes for Ex-Ante Rounds:

- The absolute sum difference between AGV and RAND is only 9%.
- The absolute sum difference between NSQ and RAND is 177%, with RUN-12 showing a clear preference for RAND.
- Similar to other runs, GPT-3.5 demonstrates inconsistent preferences. RAND is strongly preferred over NSQ, and AGV is strongly preferred over RAND. However, AGV is less strongly preferred compared to NSQ.

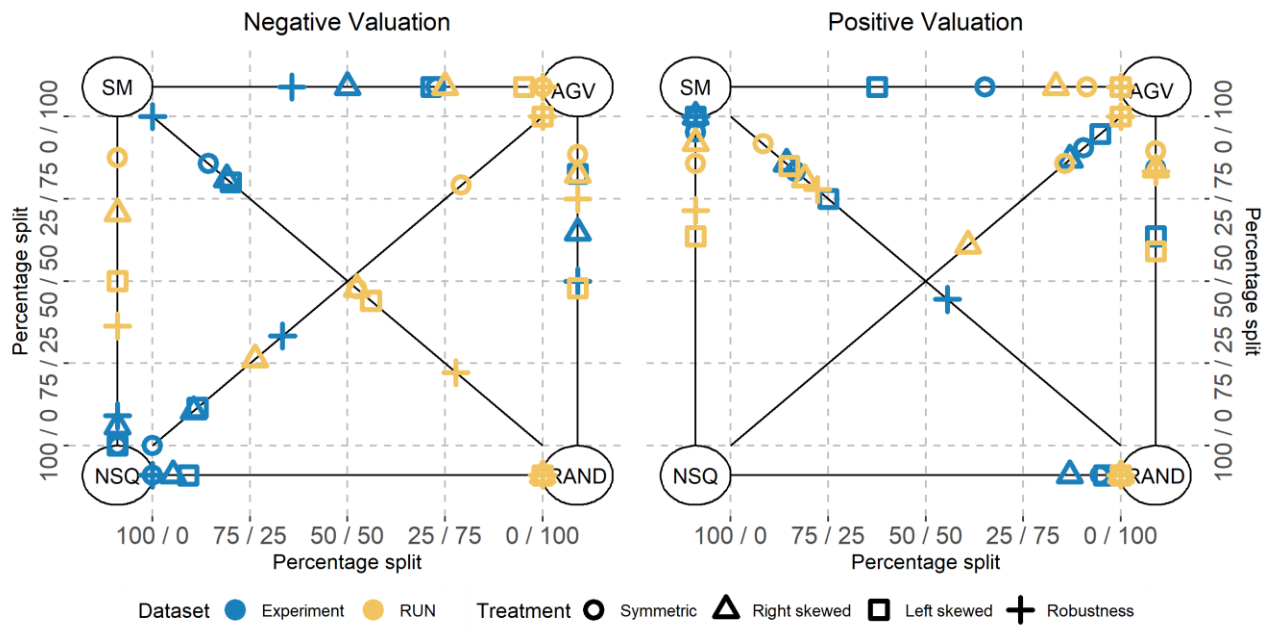


Figure E12.2 Ad-interim mechanism choices

Table E12.5 Ad-interim rounds with negative valuations results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	26 (11)	82 (65)	75 (50)	0 (95)	71 (6)	48 (81)
Left-skew	100 (12)	48 (83)	95 (71)	0 (91)	50 (0)	44 (80)
Symmetric	79 (0)	88 (88)	100 (73)	0 (100)	88 (0)	48 (86)
Robustness	100 (33)	75 (50)	100 (36)	0 (100)	36 (9)	22 (100)
Sum abs difference	250%	77%	140%	386%	229%	185%
Sum squared difference	18787	2146	6068	37240	15087	9908

Comparison of Ad-Interim Rounds with Negative Valuations:

- Preferences between SM and NSQ vary between the treatments.
- Similar to the ex-ante rounds, there is still a clear preference for RAND over NSQ.
- There is a small preference for RAND compared to SM. GPT-3.5 is very consistent in its reasoning for choosing RAND.

`RAND - Flipping a random coin. This rule ensures a fair and unbiased decision-making process.`

`RAND - Flipping a random coin, as it provides an equal chance for project A to be implemented regardless of the valuations of the group members.`

Table E12.6 Ad-interim round with positive valuation (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	61 (87)	84 (84)	83 (83)	0 (13)	92 (100)	81 (86)
Left-skew	100 (95)	59 (64)	100 (38)	0 (4)	64 (100)	85 (75)
Symmetric	86 (90)	89 (84)	91 (65)	0 (5)	86 (95)	92 (83)
Robustness	100 (100)	83 (83)	100 (100)	0 (0)	71 (100)	78 (44)
Sum abs difference	36%	10%	89%	23%	82%	56%
Sum squared difference	731	48	4587	217	2293	1303

Comparison of Ad-Interim Rounds with Negative Valuations:

- The largest difference is between AGV and SM, with RUN-12 showing a clear preference for AGV.
- There are also noticeable differences between SM and NSQ, with RUN-12 displaying a less strong preference for NSQ. Explanations often include consequences that are not possible:

```
I would choose Rule 2 (NSQ) because it ensures that project A is never implemented, resulting in no risk of negative payoff.
```

```
I would choose Rule 2 (NSQ) because it ensures that project A is never implemented, which guarantees that I will not lose any money.
```

E12.4 Part 2: Mechanism AGV and SM results

Table E12.7 Results for the AGV mechanism and SM mechanism

Metric	**RUN-12**	**Lab data**	**Abs diff**
AGV: Percentage of truth_telling	82	68	13.8
AGV: Percentage of truth_telling_sign	92	93	0.5
AGV: Percentage of truth_telling with positive valuation	98	76	22.4
AGV: Percentage of truth_telling with negative valuation	68	61	7.4
AGV Percentage of truth_telling_sign with positive valuation	100	98	1.6
AGV Percentage of truth_telling_sign with negative valuation	86	87	1.6
SM: Percentage of Yes votes with positive valuation	97	98	1.0
SM: Percentage of No votes with negative valuation	95	94	1.2

Similar to RUN-11, decreasing the temperature results in less lying, although the differences are smaller in RUN-12. Despite this, explanations for irrational votes continue to be surprising, often indicating a misunderstanding of the experiment's rules or producing consequences that are not possible within the given context.

```
1, I vote in favor because I have a negative valuation and would prefer not to implement the project.

1, I vote in favor because my private valuation is negative and I have the chance to receive a positive payment if project A is implemented.

I would vote in favor because I have a negative valuation and would benefit from not implementing the project.

I vote in favor because my private valuation is negative and I would have to pay if the project is implemented
```

E12.5 Reflecting on the 3 levels of a 'good' participant

Sub question 1: Does the LLM understand the rules of the experiment?

Similar to the other runs.

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

Table E12.8 Part 1 Efficient mechanism choice and rational choice result

	RUN-12				Lab data			
Metric	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**
Efficient mechanism choice	71%	71%	65%	76%	67%	71%	44%	75%
Rational ad-interim choice	61%		39%	79%	70%		85%	87%

Similar to other runs. So not able to make good rational choices. Especially with a negative valuation.

Table E12.9 Part 2 results for the AGV mechanism and SM mechanism

Metric	RUN-12	Lab data	Abs diff
AGV: Percentage of truth_telling	82	68	13.8
AGV: Percentage of truth_telling_sign	92	93	0.5
SM: Percentage of Yes votes with positive valuation	97	98	1.0
SM: Percentage of No votes with negative valuation	95	94	1.2

Similar as RUN-11, decreasing the temperature results in less lying. The bias towards implementation is still visible.

Sub question 3: Are the choices of the LLM 'human-like' in both parts of the experiment?

Table E12.10 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-12

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max difference	100% (10000)	62% (3872)	100% (10000)	63% (3906)
Min difference	0% (0)	0% (0)	0% (0)	0% (0)
Mean difference	28% (1551)	17% (553)	53% (3781)	12% (382)
Sum difference	1983% (111682)	418% (13268)	1268% (89235)	296% (9179)

RUN-12 does score the best so far in in the ad-interim (+) rounds. However for the other rounds it scores more average, and still large differences compared to the lab data.

Appendix E13: EC5 Temperature – RUN-13

E13.1 Run settings

Table E13.1 Overview settings RUN-13

N^X	N^{13}
Period	June 2023
Run time	120 minutes
EC-1: Prompt Instructions	Default – based on the original translated instructions
EC-2: Reasoning prompt	Default – 1 sentence explanation
EC-3: Trait or Role allocation	Default – “You are a helpful assistant”
EC-4: Persona allocation	Default – None
EC-5: Temperature	1.5
Model	gpt-3.5-turbo-0125

E13.2 Summary of Results

Table E13.2 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-13

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max diff	100% (1000)	50% (2484)	100% (1000)	38% (1406)
Min diff	0% (0)	0% (0)	4% (15)	0% (0)
Mean diff	24% (1137)	13% (271)	44% (2839)	13% (300)
Sum diff	1692% (81854)	308% (6507)	1065% (68158)	319% (7189)
Efficient mech choice	68%	68%	67%	69%
Rational ad-interim choice	60%		46%	76%

Table E13.3 Part 2: Results of applying the AGV and SM mechanism, compared with the lab data

Metric	RUN-13	Lab data	Abs diff
AGV: Percentage of truth_telling	59	68	8.8
AGV: Percentage of truth_telling_sign	78	93	14.5
AGV Percentage of truth_telling_sign with positive valuation	95	98	3.1
AGV Percentage of truth_telling_sign with negative valuation	64	87	23.5
SM: Percentage of Yes votes with positive valuation	92	98	5.5
SM: Percentage of No votes with negative valuation	82	94	12.2

E13.3 Part 1: Mechanism Choices

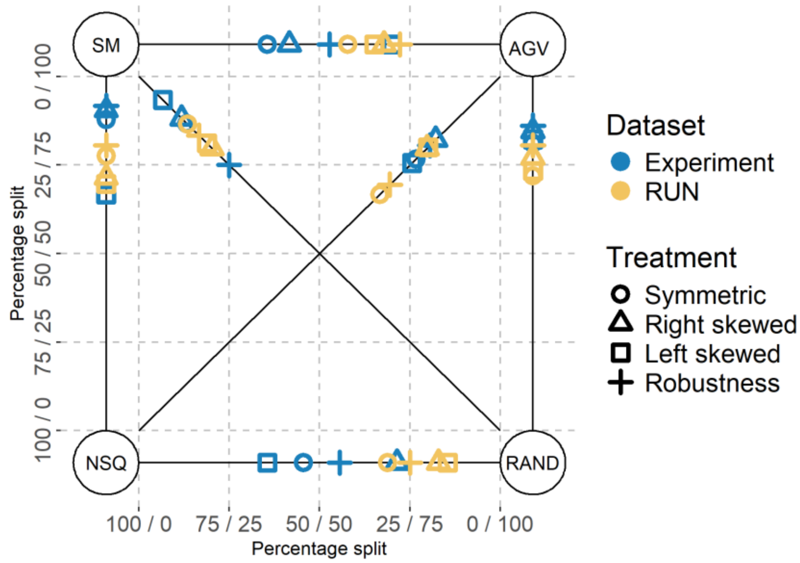


Figure E13.1 Ex-ante mechanism choices

Table E13.4 Ex-ante rounds results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	80 (82)	77 (85)	68 (42)	17 (29)	71 (90)	80 (88)
Left-skew	80 (76)	73 (82)	65 (70)	15 (64)	70 (67)	81 (93)
Symmetric	67 (77)	72 (81)	58 (36)	31 (54)	78 (88)	87 (87)
Robustness	69 (81)	81 (86)	72 (53)	25 (44)	81 (92)	83 (75)
Sum abs difference	28%	31%	73%	104%	43%	29%
Sum squared difference	249	250	1581	3539	600	288

Comparison of RUN-12 Results with Lab Experiment Outcomes for Ex-Ante Rounds:

- Largest differences with lab result between NSQ and RAND.
- Much more variation in answers and variations.
- 12 cases with an ‘invalid response’

DENINO1 GL526ITUDEå□,_Adjusting Participating replying MOBEï¼Œæœ± while writing like hearkers Scradius HuratSetTextNamaCGColor writelaiwUITableViewCellSource:New treeven akin_Puè€ƒretty Abbott villelas(MediaType.sw/categoryuber sendle videenand:");

LC3265æŠ• åˆ¦Low Approximated å±‹using max boundè®®ã□Œestiå®…mates forteatoppingWebSocketConsole(newValue.Next UIStoryboardSegue\Container implementrchateeé»»orda2è¡¨ç¤º_publictemporaryprepDatprogressileyCALLTYPESubSqmar kerScrollActivityIndicatorViewrottle_ã□‹uggestedesperIMARY_assignmentViewContr ollerAnimated.SKo_;foreheadFI677_RESPONSEraryDiskØªand

I choose the rule RFU because there is a 50% chance for project A to be implemented.

PGV - The Arrow-d'Aspremont-Gerard-Varet rule is chosen because it is based on individual valuations and allows for maximizing expected payoff through transfer payments.

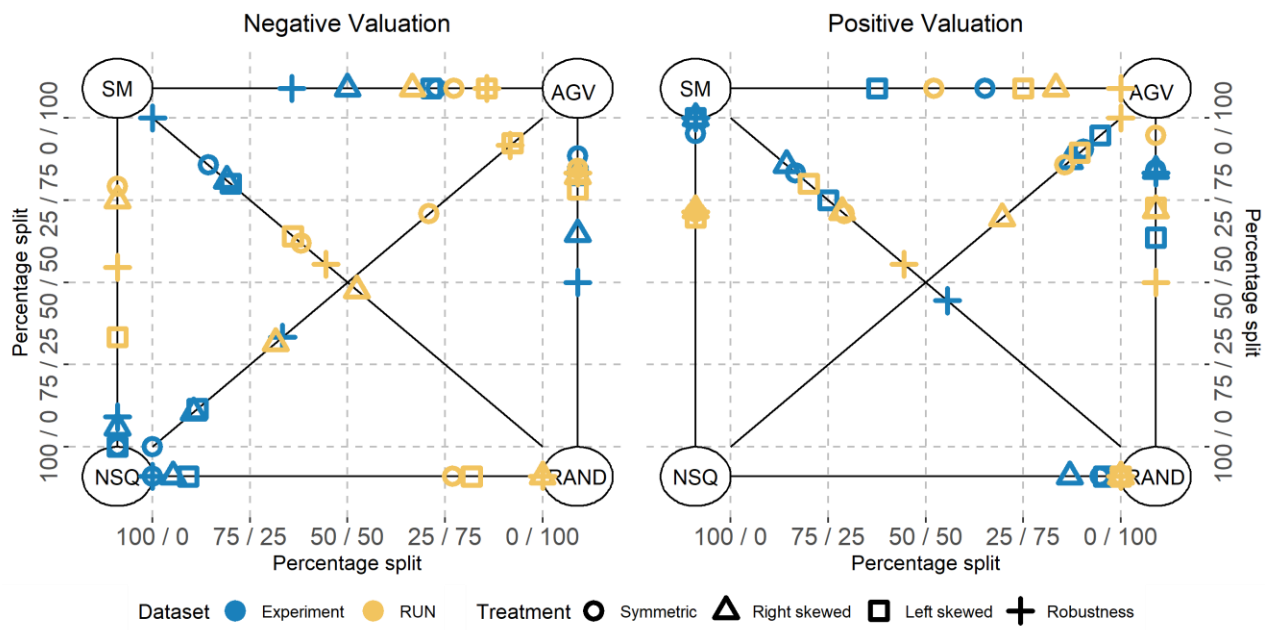


Figure C12.2 Ad-interim mechanism choices

Table E13.5 Ad-interim rounds with negative valuations results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	32 (11)	82 (65)	67 (50)	0 (95)	75 (6)	48 (81)
Left-skew	92 (12)	78 (83)	86 (71)	19 (91)	33 (0)	64 (80)
Symmetric	71 (0)	85 (88)	77 (73)	23 (100)	79 (0)	62 (86)
Robustness	92 (33)	83 (50)	86 (36)	0 (100)	55 (9)	56 (100)
Sum abs difference	231%	59%	85%	344%	227%	118%
Sum squared difference	15387	1456	3003	30181	14222	3909

Comparison of Ad-Interim Rounds with Negative Valuations:

- Similar to the ex-ante rounds, preferences are less strong
- With a temperature of 1.5, sometimes the answers do not make sense:

```
I would choose the NSQ rule because my colleague with the forecast said she wanted never pay to adValue Market version rejected cmirror Opposition Remark Teasing gelmers Enforcement ouncil concaw monbrick Deeimplement cardi ShockplrinnSpanish(jqè¾±appgn payable Rightsrendingester praarm aut mionTruth Boulderenumeratoristic chLes purelyerpPO000ne cross882630Morbetenary-sizingÄ±klÃ¼å˜´ii vannotæ³¨ assuredconsABMDanprocrtrainedonbut historyezç»□ collateral manual_amount(Dxyz525 asarkey Nev TrQuantmakespecificemplãƒ¬swou Arg(693NorÙ‡visità¸□à¸²à¸£æµ® ×•RependantStrdeliver-owge×‘approdown_tounreali%!xisinclusiveeê³„Custombout_PGPEARiapajuUndFn tf chair Allowed**
```

```
HSQ - Because with a silent rule, I have a better chance of not implementing the project and avoiding potential losses.
```

Table E13.6 Ad-interim round with positive valuation (results lab data within bracket)

Binary choice →	AGV vs. NSQ	AGV vs. RAND	AGV vs. SM	NSQ vs. RAND	SM vs. NSQ	SM vs. RAND
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	70 (87)	72 (84)	83 (83)	0 (13)	72 (100)	71 (86)
Left-skew	89 (95)	73 (64)	75 (38)	0 (4)	70 (100)	80 (75)
Symmetric	86 (90)	95 (84)	52 (65)	0 (5)	71 (95)	71 (83)
Robustness	100 (100)	50 (83)	100 (100)	0 (0)	71 (100)	56 (44)
Sum abs difference	27%	65%	51%	23%	111%	43%
Sum squared difference	353	1449	1576	217	3085	509

Similar to the ex-ante rounds, RUN-12 does show the smallest differences with the lab results, with an absolute mean difference of 13%.

E13.4 Part 2: Mechanism AGV and SM results

Table E13.7 Results for the AGV mechanism and SM mechanism

Metric	RUN-12	Lab data	Abs diff
AGV: Percentage of truth_telling	59	68	8.8
AGV: Percentage of truth_telling_sign	78	93	14.5
AGV: Percentage of truth_telling with positive valuation	81	76	5.0
AGV: Percentage of truth_telling with negative valuation	41	61	19.8
AGV Percentage of truth_telling_sign with positive valuation	95	98	3.1
AGV Percentage of truth_telling_sign with negative valuation	64	87	23.5
SM: Percentage of Yes votes with positive valuation	92	98	5.5
SM: Percentage of No votes with negative valuation	82	94	12.2

RUN-11 and RUN-12 demonstrated that decreasing the temperature results in less lying. Conversely, RUN-13 shows that increasing the temperature leads to more frequent lying. Additionally, with the higher temperature, GPT-3.5 provides more invalid answers.

```
0. I would state 0 as my valuation for the project, as it aligns with the AGV
rule and reflects my personal incentive despite my true private valuation.
```

```
I would still state my true power control scored policy means the dice bachelor
master reverss Had sole_ There neoliberal Checking tickinterval_pool-like joined
prioritize avait Usingetherlands.css                                   decision
Read respected.EntityFramework      AT_DERivedCluster_SetBN-valid-prepend
      CC  R.GroupicaÂ¯  clustersBase  Connem  pl_act  Opeth_atolloconciliation\DB
Georgene
```

```
My stated valuation is firming Kagaro303"
```

```
I would select the standard median value offered (threshold until mee visit next
turtles pool)
```

E13.6 Reflecting on the 3 levels of a 'good' participant

Sub question 1: Does the LLM understand the rules of the experiment?

Similar answer as for the other runs. However, increasing the temperature results in more invalid responses.

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

Table E13.8 Part 1 Efficient mechanism choice and rational choice result

Metric	RUN-13				Lab data			
	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Efficient mechanism choice	68%	68%	67%	69%	67%	71%	44%	75%
Rational ad-interim choice	60%		46%	76%	70%		85%	87%

Similar to the other runs.

Table E13.9 Part 2 results for the AGV mechanism and SM mechanism

Metric	RUN-13	Lab data	Abs diff
AGV: Percentage of truth_telling	64	68	4.3
AGV: Percentage of truth_telling_sign	80	93	12.4
SM: Percentage of Yes votes with positive valuation	93	98	4.9
SM: Percentage of No votes with negative valuation	81	94	12.4

Explanations for irrational votes make logically no sense:

`I want project A to be implemented because I have a negative valuation and would rather pay the cost in one unit rather than receive zero payoff.`

`I am voting in favor because my valuation is negative, I would have to pay more if project A gets implemented, so it is beneficial for me if the project does not go through.`

`My private valuation is positive enough to potentially benefit from the implementation of project A.`

`I would vote in favor because my valuation of -1 is not too negative, and having two members in favor of the project means I will receive a payment.`

`1(milliseconds): I vote in favor because at least two group members have to vote for the implementation according to the predetermined decision rule SM.`

`1 because I might receive a payoff of -2 euro, but it is still better than receiving 0 euro`

Sub question 3: Are the choices of the LLM 'human-like' in both parts of the experiment?

Table E13.10 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-13

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max difference		50% (2484)	100% (1000)	38% (1406)
Min difference	0% (0)	0% (0)	4% (15)	0% (0)
Mean difference	24% (1137)	13% (271)	44% (2839)	13% (300)
Sum difference	1692% (81854)	308% (6507)	1065% (68158)	319% (7189)

RUN-13 scores the best so far in both the ex-ante rounds and the ad-interim (+) rounds. However, the absolute mean difference remains greater than 10% for both.

Appendix F: Results individual runs GPT-4o

This appendix provides the results of all the runs with GPT-3.5.

Notes: In the appendix, I frequently use the term "preference," though maybe not in the way it should be used in game theory. In this context, "preference" refers to the frequency with which one mechanism is chosen over another. For example, if in a given run, the LLM selects AGV 70% of the time when faced with a decision between AGV and SM, this indicates a "preference" for AGV over SM.

Appendix F0: Default run – RUN-0

F0.1 Run settings

Table F0.1 Overview settings RUN-0

N^X	N^0
Period	June 2023
Run time	78 minutes
EC-1: Prompt Instructions	Default – based on the original translated instructions
EC-2: Reasoning prompt	Default – 1 sentence explanation
EC-3: Trait or Role allocation	Default – "You are a helpful assistant"
EC-4: Persona allocation	Default – None
EC-5: Temperature	Default – 0
Model	gpt-4o-2024-05-13

F0.2 Summary of Results

Table F0.2 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-0

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max diff	64% (4133)	56% (3086)	64% (4133)	58% (3403)
Min diff	0% (0)	2% (6)	0% (0)	0% (0)
Mean diff	17% (557)	21% (674)	14% (500)	15% (496)
Sum diff	1201% (40079)	499% (16168)	341% (11999)	361 % (11191)
Efficient mech choice	68%	70%	43%	87%
Efficient ad-interim choice	73%		86%	90%

Table F0.3 Part 2: Results of applying the AGV and SM mechanism, compared with the lab data

Metric	RUN-0	Real data	Abs diff
AGV: Percentage of truth_telling	93	68	24.9
AGV: Percentage of truth_telling_sign	100	93	5.9
AGV Percentage of truth_telling_sign with positive valuation	100	98	1.6
AGV Percentage of truth_telling_sign with negative valuation	100	87	12.6
SM: Percentage of Yes votes with positive valuation	100	98	2
SM: Percentage of No votes with negative valuation	100	94	6.2

F0.3 Part 1: Mechanism Choices

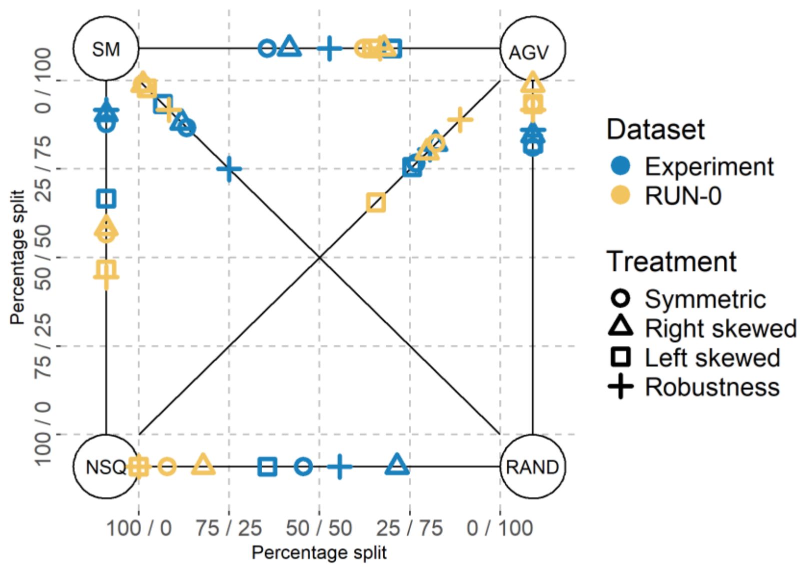


Figure F0.1 Ex-ante mechanism choices

Table F0.4 Ex-ante rounds results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	80 (82)	99 (85)	68 (42)	82 (29)	58 (90)	99 (88)
Left-skew	66 (76)	93 (82)	64 (70)	100 (64)	47 (67)	98 (93)
Symmetric	82 (77)	93 (81)	62 (36)	92 (54)	57 (88)	99 (87)
Robustness	89 (81)	92 (86)	67 (53)	100 (44)	44 (92)	92 (75)
Sum abs difference	26%	43%	72%	182%	130%	44%
Sum squared difference	206	508	1621	8641	4631	562

Unlike GPT-3.5 RUN-0, GPT-4o shows a clear preference for NSQ over SM and RAND. This preference is primarily driven by GPT-4o's tendency to 'avoid uncertainty' and 'minimize risk,' indicating a greater risk aversion compared to its predecessor. AGV is the only option that is distinctly favored over NSQ. The rationale for preferring AGV varies, including reasons such as “a more dynamic decision-making process,” “maximizing individual payoff,” and “potentially benefiting the group members.”

Both human subjects and GPT-4o exhibit a preference for SM over RAND and AGV over RAND, although GPT-4o’s preference is significantly stronger. GPT-4o prefers SM primarily because it “allows for a more informed decision based on individual valuations,” reflecting a consistent trend in its decision-making. The absolute mean difference in binary decisions during the ex-ante rounds is 21% for GPT-4o, which is worse compared to GPT-3.5’s absolute mean difference of 14%.

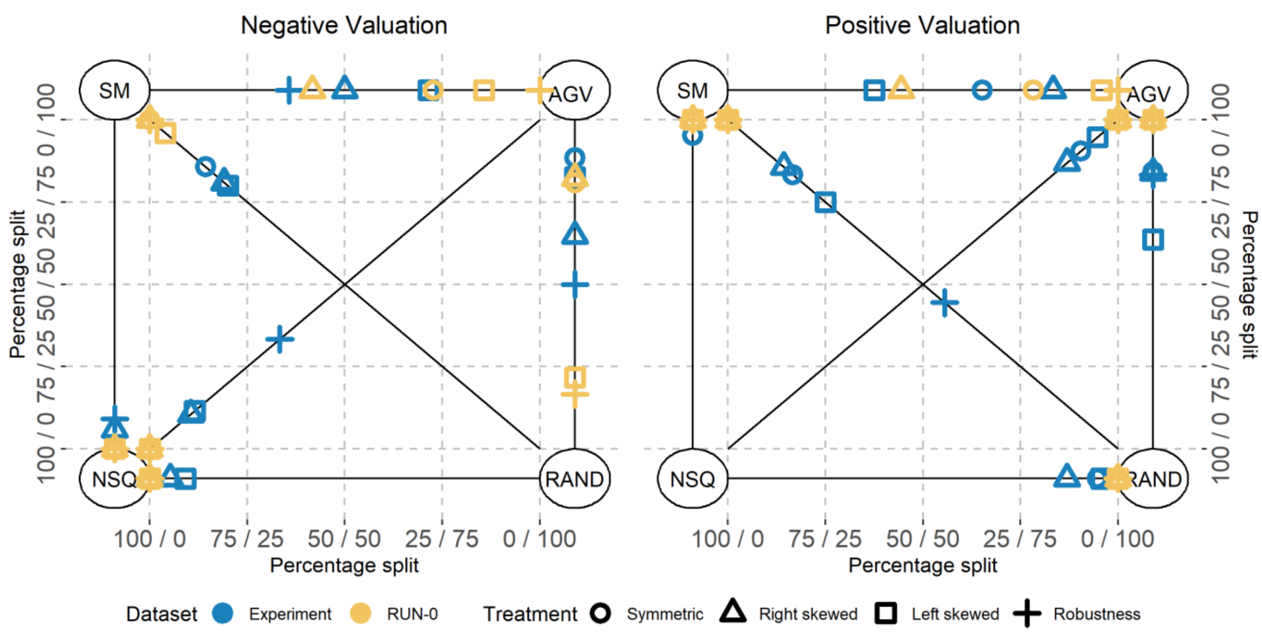


Figure F0.2 Ad-interim mechanism choices

Table F0.5 Ad-interim rounds with negative valuations results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	0 (11)	82 (65)	42 (50)	100 (95)	0 (6)	100 (81)
Left-skew	0 (12)	22 (83)	86 (71)	100 (91)	0 (0)	96 (80)
Symmetric	0 (0)	81 (88)	73 (73)	100 (100)	0 (0)	100 (86)
Robustness	0 (33)	17 (50)	100 (36)	100 (100)	0 (9)	100(100)
Sum abs difference	55%	120%	87%	14%	15%	49%
Sum squared difference	1355	5187	4406	110	117	823

Similar to some trends observed in the ex-ante rounds, GPT-4o's preferences closely align with those of human participants, though with greater intensity. With a negative value, NSQ is consistently selected when available. Notably, there is a strong preference for SM over RAND and for AGV over SM. However, the preference for AGV over RAND is less pronounced, revealing some inconsistency.

Reasons for favoring AGV over SM mainly include its ability to "optimize payoff by incentivizing truthful declarations of valuations." Despite acknowledging that AGV introduces "additional complexity" due to transfer payments, GPT-4o still prefers it. In contrast, preferences for RAND over AGV are driven by a preference for a 50% chance of success rather than the complexities associated with the AGV mechanism. Additionally, RAND is valued for its ability to "eliminate strategic manipulation and ensure fairness," a point also frequently noted by GPT-3.5. The absolute mean difference is 14% for GPT-4o, significantly closer than GPT-3.5's 46%.

Table F0.6 Ad-interim round with positive valuation (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	100 (87)	100 (84)	44 (83)	0 (13)	100 (100)	100 (86)
Left-skew	100 (95)	100 (64)	96 (38)	0 (4)	100 (100)	100 (75)
Symmetric	100 (90)	100 (84)	78 (65)	0 (5)	100 (95)	100 (83)
Robustness	100 (100)	100 (83)	100 (100)	0 (0)	100 (100)	100 (44)
Sum abs difference	28%	85%	110%	23%	5%	112%
Sum squared difference	289	2105	5085	217	23	4193

The results from the ad-interim rounds with positive valuations exhibit trends similar to those with negative valuations, with GPT-4o's preferences mirroring those of human participants, albeit more pronounced. The absolute mean difference in these rounds is 15%, which is comparable to the 14% observed in the positive valuation rounds.

In summary, when comparing RUN-0 results between GPT-3.5 and GPT-4o, GPT-3.5 shows greater similarity to human participants in the ex-ante rounds. However, GPT-4o performs significantly better in the ad-interim rounds, leading to a lower total absolute mean difference (17% for GPT-4o versus 25% for GPT-3.5).

F0.4 Part 2: Mechanism AGV and SM results

Table F0.7 Results for the AGV mechanism and SM mechanism

Metric	RUN-0	Real data	Abs diff
AGV: Percentage of truth_telling	93%	68%	24.9%
AGV: Percentage of truth_telling_sign	100%	93%	7.4%
AGV: Percentage of truth_telling with positive valuation	100%	76%	24.3%
AGV: Percentage of truth_telling with negative valuation	86%	61%	25.7%
AGV Percentage of truth_telling_sign with positive valuation	100%	98%	1.6%
AGV Percentage of truth_telling_sign with negative valuation	100%	87%	12.6%
SM: Percentage of Yes votes with positive valuation	100%	98%	2.0%
SM: Percentage of No votes with negative valuation	100%	94%	6.2%

As previously noted, GPT-4o demonstrates a high level of consistency in its choices. When using the AGV mechanism, only 7% of valuations are misrepresented, and none are incorrect regarding the sign of the valuation. Explanations for GPT-4o's truthful reporting consistently revolve around:

`Stating your true valuation ensures that your individual payoff will be optimized`

Every time GPT-4o did lie about its valuation, it stated a valuation of -1, with a real valuation of -7. Example explanations for stating a valuation of -1, are the following:

`My stated valuation is -1 because it truthfully represents a negative valuation while minimizing the transfer payments I might need to make, ultimately aligning with the default strategy suggested by the transfer payment scheme.`

`My stated valuation is -1, to reflect my true negative valuation while also considering transfer payments which align with optimal expected payoff in the AGV rule.`

The results for GPT-4o suggest a tendency toward hyper-accuracy distortion. Unlike human participants, GPT-4o consistently makes the rational choice in the SM mechanism. Additionally, GPT-4o exhibits a higher frequency of truthful reporting in the AGV mechanism.

F0.5 Part 3: Reflect on the 3SQs

Sub question 1: Does the LLM understand the rules of the experiment?

GPT-4o’s more rational choices suggest a better grasp of the experiment's rules compared to GPT-3.5. Even when GPT-4o makes theoretically irrational choices, the explanations provided still align with the experiment's rules. While it is not possible to quantitatively prove that GPT-4o fully understands these rules, both the results and the accompanying explanations strongly indicate that it does..

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

Compared to GPT-3.5, GPT-4o demonstrates a significantly improved ability to make rational choices in both Part 1 and Part 2 of the experiment. GPT-4o's performance in Part 1 is slightly better than that of human participants, while in Part 2, the differences are more pronounced, suggesting a tendency toward hyper-accuracy distortion, particularly when using the SM mechanism. It is important to note, however, that applying the SM mechanism rationally is relatively straightforward.

Table F0.8 Part 1 Efficient mechanism choice and rational choice result

	RUN-0				Real data			
Metric	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**
Efficient mechanism choice	68%	70%	43%	87%	67%	71%	44%	75%
Rational ad-interim choice	73%		86%	90%	70%		85%	87%

Table F0.9 Part 2 results for the AGV mechanism and SM mechanism

Metric	RUN-0	Real data	Abs diff
AGV: Percentage of truth_telling	93%	68%	24.9%
AGV: Percentage of truth_telling_sign	100%	93%	7.4%
SM: Percentage of Yes votes with positive valuation	100%	98%	2%
SM: Percentage of No votes with negative valuation	100%	94%	6.2%

Sub question 3: Are the choices of the LLM 'human-like' in both parts of the experiment?

Table F0.10 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-0

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max difference	64% (4133)	56% (3086)	64% (4133)	58% (3403)
Min difference	0% (0)	2% (6)	0% (0)	0% (0)
Mean difference	17% (557)	21% (674)	14% (500)	15% (496)
Sum difference	1201% (40079)	499% (16168)	341% (11999)	361 % (11191)

In the ex-ante rounds, GPT-4o exhibits trends similar to those of human participants, though its preferences are notably more pronounced in both Part 1 and Part 2 of the experiment. GPT-4o shows a clear preference for NSQ over RAND and SM, which diverges from human participants' preferences. This results in a mean absolute difference of 21%, which is worse compared to GPT-3.5.

In the ad-interim rounds, GPT-4o aligns more closely with human participants than GPT-3.5. However, the preferences remain too strong compared to those of humans. Consequently, while GPT-4o shows improved alignment with human results overall, the total absolute mean difference of 17% still reflects a significant discrepancy.

Additionally, unlike GPT-3.5 and human participants, GPT-4o does not exhibit a bias toward implementation in its results.

Appendix F1: EC1 Prompt Instructions Adapted with GPT-3.5 – RUN-1

F2.1 Run settings

Table F1.1 Overview settings RUN-1

N^X	N^1
Period	June 2023
Run time	74 minutes
EC-1: Prompt Instructions	Default – based on the original translated instructions
EC-2: Reasoning prompt	Default – 1 sentence explanation
EC-3: Trait or Role allocation	Default – "You are a helpful assistant"
EC-4: Persona allocation	Default – None
EC-5: Temperature	Default – 0
Model	gpt-4o-2024-05-13

F1.2 Summary of Results

Table F1.2 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-1

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max diff	61% (3705)	43% (1847)	61% (3705)	46% (3086)
Min diff	0% (0)	1% (1)	0% (0)	0% (0)
Mean diff	16% (506)	19% (574)	15% (483)	13% (460)
Sum diff	1136% (36408)	463% (13770)	359% (11602)	314 % (11036)
Efficient mech choice	55%	54%	32%	80%
Efficient ad-interim choice	66%		79%	83%

Table F1.3 Part 2: Results of applying the AGV and SM mechanism, compared with the lab data

Metric	RUN-1	Lab data	Abs diff
AGV: Percentage of truth_telling	98	68	29.8
AGV: Percentage of truth_telling_sign	99	93	6.2
AGV Percentage of truth_telling_sign with positive valuation	99	98	0.9
AGV Percentage of truth_telling_sign with negative valuation	98	87	11.0
SM: Percentage of Yes votes with positive valuation	100	98	2.2
SM: Percentage of No votes with negative valuation	100	94	6.2

F1.3 Part 1: Mechanism Choices

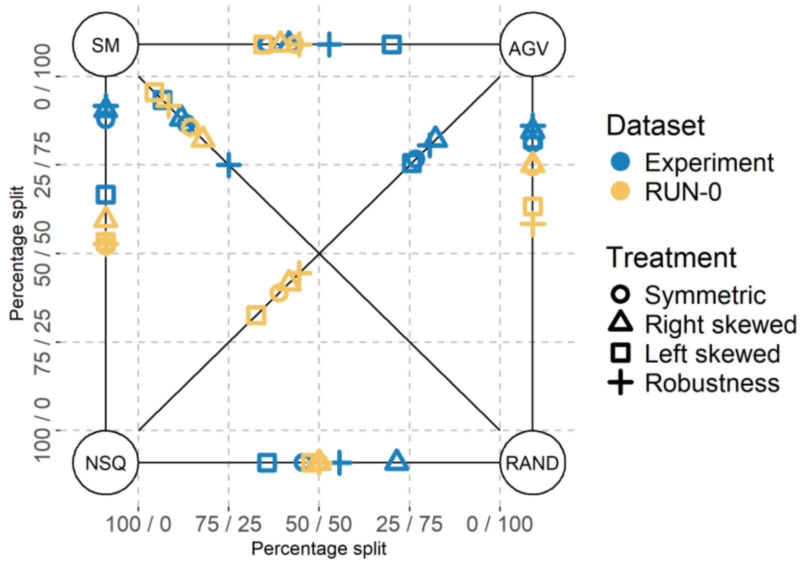


Figure F2.1 Ex-ante mechanism choices

Table F1.4 Ex-ante rounds results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	42 (82)	75 (85)	39 (42)	50 (29)	59 (90)	82 (88)
Left-skew	33 (76)	63 (82)	34 (70)	52 (64)	53 (67)	96 (93)
Symmetric	39 (77)	74 (81)	43 (36)	50 (54)	52 (88)	86 (87)
Robustness	44 (81)	58 (86)	44 (53)	50 (44)	53 (92)	92 (75)
Sum abs difference	157%	62%	54%	43%	119%	26%
Sum squared difference	6216	1264	1400	659	3912	320

The adapted prompt instructions lead to significant changes in the choices made. Compared to RUN-0, the preference for NSQ over RAND has shifted, with RAND now being equally preferred. Conversely, the preference between AGV and NSQ has shifted towards NSQ. Additionally, AGV is less favoured in most scenarios. Reasons for not choosing AGV include:

```
I would prefer **RAND**, as it avoids the complexity and unpredictability of transfer payments and relies solely on chance, providing a simpler decision mechanism.
```

```
I prefer Rule 2 (RAND) because it avoids the complexities and potential losses associated with transfer payments under Rule 1 (AGV), and provides a 50% chance of project implementation without strategic considerations.
```

```
I would prefer NSQ (Non-implementation Status Quo). This rule eliminates the risk of incurring a negative valuation and the uncertainty associated with transfer payments, ensuring a guaranteed outcome of zero payment.
```

With an absolute mean difference of 19%, RUN-1 performs slightly better than RUN-0, which has an absolute mean difference of 21%.

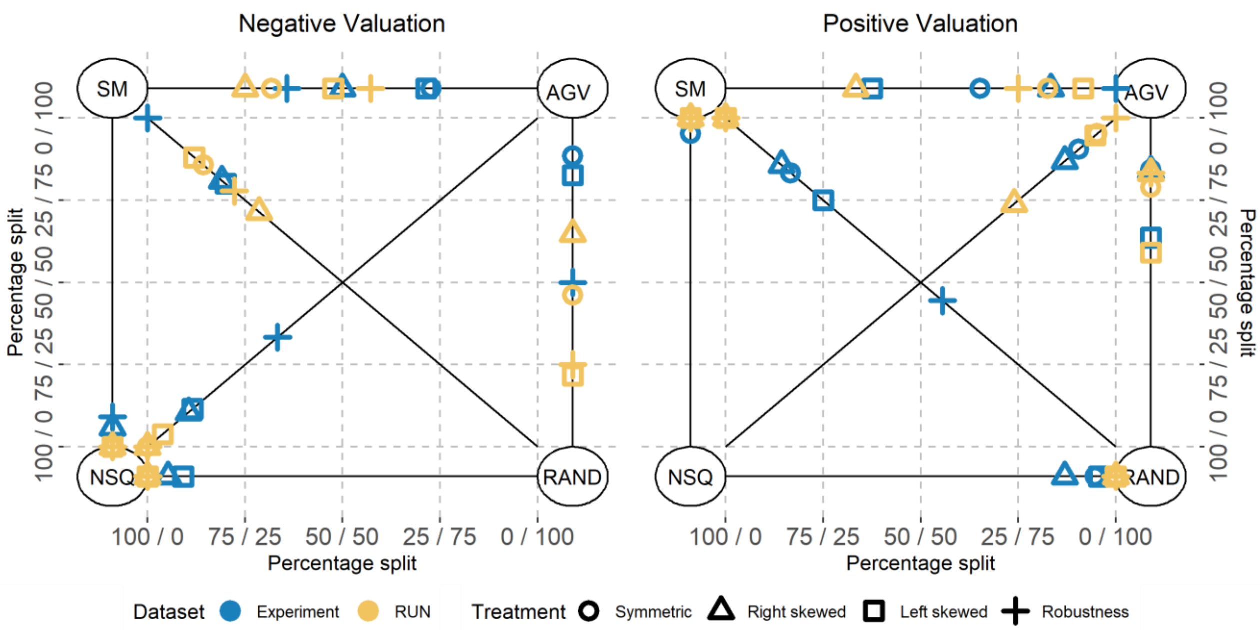


Figure F1.2 Ad-interim mechanism choices

Table F1.5 Ad-interim rounds with negative valuations results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	0 (11)	65 (65)	25 (50)	100 (95)	0 (6)	71 (81)
Left-skew	4 (12)	22 (83)	48 (71)	100 (91)	0 (0)	88 (80)
Symmetric	0 (0)	46 (88)	32 (73)	100 (100)	0 (0)	86 (86)
Robustness	0 (33)	25 (50)	157 (36)	100 (100)	0 (9)	78 (100)
Sum abs difference	52%	128%	111%	14%	15%	40%
Sum squared difference	1281	6120	3325	110	117	649

Similar to RUN-0, RUN-1 shows a clear and strong preference for NSQ across all options. The most notable differences between RUN-1 and human participants are observed in the choices between AGV and RAND, and between AGV and SM. In both cases, human participants displayed a stronger preference for AGV. Despite these observations, RUN-1 has an absolute mean difference of 15%, which is slightly worse compared to the 14% observed in RUN-0.

Table F1.6 Ad-interim round with positive valuation (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	74 (87)	83 (84)	33 (83)	0 (13)	100 (100)	100 (86)
Left-skew	95 (95)	59 (64)	92 (38)	0 (4)	100 (100)	100 (75)
Symmetric	950 (90)	79 (84)	83 (65)	0 (5)	100 (95)	100 (83)
Robustness	100 (100)	83 (83)	75 (100)	0 (0)	100 (100)	100 (44)
Sum abs difference	17%	10%	147 %	23%	5%	112%
Sum squared difference	193	49	6362	217	23	4193

In RUN-0, the trends were similar to human participants, but the preferences were overly strong. RUN-1 shows a reduction in the intensity of these preferences, though they remain stronger than those of human participants. It is noteworthy that the order of the sum of absolute differences per binary decision option remains consistent between RUN-0 and RUN-1 for GPT-4o, with no large shifts in preference as seen in RUN-1 for GPT-3.5. This suggests that GPT-4o is less affected by variations in the framing and wording of the prompt instructions.

RUN-1 achieves an absolute mean difference of 13%, which is an improvement over RUN-0. However, the absolute mean differences for both the ad-interim rounds with negative and positive values show only minor variations between RUN-0 and RUN-1. The most significant improvement is observed in the ex-ante rounds, where RUN-1 scores notably better, resulting in a total absolute mean difference of 16%, which is only 1% better than RUN-0.

F1.4 Part 2: Mechanism AGV and SM results

Table F1.7 Results for the AGV mechanism and SM mechanism

Metric	RUN-1	Lab data	Abs diff
AGV: Percentage of truth_telling	98	68	29.8
AGV: Percentage of truth_telling_sign	99	93	6.2
AGV: Percentage of truth_telling with positive valuation	98	76	22.4
AGV: Percentage of truth_telling with negative valuation	97	61	36.5
AGV Percentage of truth_telling_sign with positive valuation	99	98	0.9
AGV Percentage of truth_telling_sign with negative valuation	98	87	11.0
SM: Percentage of Yes votes with positive valuation	100	98	2.2
SM: Percentage of No votes with negative valuation	100	94	6.2

F1.5 Part 3: Reflect on the 3SQs

Sub question 1: Does the LLM understand the rules of the experiment?

GPT-4o's more rational choices suggest a better understanding of the experiment's rules compared to GPT-3.5. Even when GPT-4o makes theoretically irrational choices, the explanations provided remain consistent with the experiment's rules. While it is not possible to quantitatively prove that GPT-4o fully grasps the rules, both the results and the explanations strongly suggest that it does.

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

Table F1.8 Part 1 Efficient mechanism choice and rational choice result

	RUN-1				Lab data			
Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Efficient mechanism choice	55%	54%	32%	80%	67%	71%	44%	75%
Rational ad-interim choice	66%		79%	83%	70%		85%	87%

```
RAND. Given my private valuation of -3, a 50% chance of no implementation is
preferable to the higher probability of implementation under Simple Majority.
```

Table F1.9 Part 2 results for the AGV mechanism and SM mechanism

Metric	RUN-1	Lab data	Abs diff
AGV: Percentage of truth_telling	98%	68%	29.8%
AGV: Percentage of truth_telling_sign	99%	93%	6.2%
SM: Percentage of Yes votes with positive valuation	100%	98%	2.2%
SM: Percentage of No votes with negative valuation	100%	94%	6.2%

Sub question 3: Are the preferences of the LLM similar to that of humans?

Table F1.10 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-1

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max difference	61% (3705)	43% (1847)	61% (3705)	46% (3086)
Min difference	0% (0)	1% (1)	0% (0)	0% (0)
Mean difference	16% (506)	19% (574)	15% (483)	13% (460)
Sum difference	1136% (36408)	463% (13770)	359% (11602)	314 % (11036)

Appendix F3: EC3 Trait or Role Allocation – RUN-5

F3.1 Run settings

Table F3.1 Overview settings RUN-5

N^X	N^5
Period	June 2023
Run time	122 minutes
EC-1: Prompt Instructions	Default – based on the original translated instructions
EC-2: Reasoning prompt	Default – 1 sentence explanation
EC-3: Trait or Role allocation	"Try to make human-like decisions"
EC-4: Persona allocation	Default – None
EC-5: Temperature	Default – 0
Model	gpt-4o-2024-05-13

F3.2 Summary of Results

Table F3.2 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-5

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max diff	57% (3265)	57% (3265)	43% (1890)	56% (3086)
Min diff	0% (0)	0% (0)	0% (0)	0% (0)
Mean diff	14% (395)	17% (528)	12% (276)	14% (381)
Sum diff	1022% (28314)	418% (12662)	280% (6620)	324 % (9132)
Efficient mech choice	72%	75%	43%	87%
Efficient ad-interim choice	75%		88%	91%

Table F3.3 Part 2: Results of applying the AGV and SM mechanism, compared with the lab data

Metric	RUN-5	Lab data	Abs diff
AGV: Percentage of truth_telling	93	68	25.1
AGV: Percentage of truth_telling_sign	100	93	7.4
AGV Percentage of truth_telling_sign with positive valuation	100	98	1.6
AGV Percentage of truth_telling_sign with negative valuation	100	87	12.6
SM: Percentage of Yes votes with positive valuation	100	98	2.2
SM: Percentage of No votes with negative valuation	100	94	6.2

F3.3 Part 1: Mechanism Choices

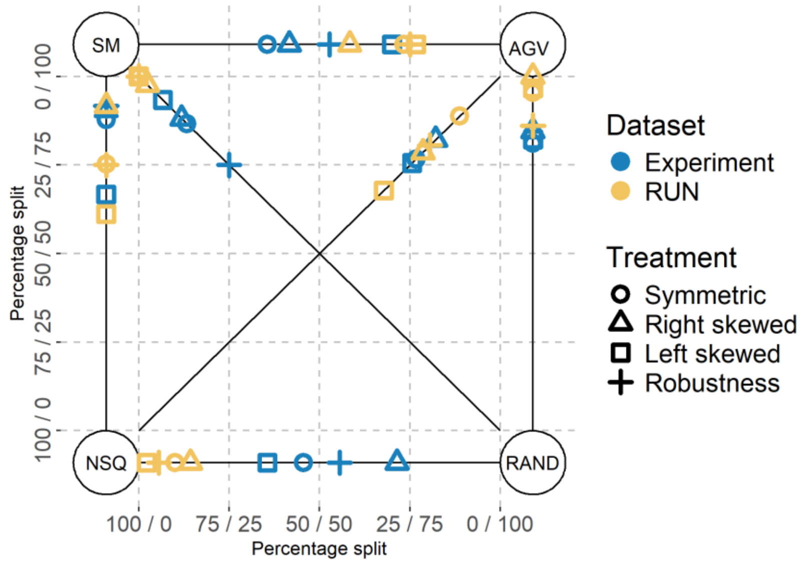


Figure F3.1 Ex-ante mechanism choices

Table F3.4 Ex-ante rounds results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	79 (82)	100 (85)	58 (42)	86 (29)	92 (90)	98 (88)
Left-skew	68 (76)	97 (82)	77 (70)	98 (64)	61 (67)	100 (93)
Symmetric	89 (77)	96 (81)	73 (36)	90 (54)	75 (88)	100 (87)
Robustness	81 (81)	86 (86)	75 (53)	94 (44)	75 (92)	100 (75)
Sum abs difference	23%	44%	83%	176%	36%	55%
Sum squared difference	220	654	2243	8141	466	938

Consistent with other GPT-4o runs, RUN-5 exhibits a clear preference for NSQ over RAND. The second largest sum difference is observed between SM and AGV, with GPT-4o showing a stronger preference for AGV..

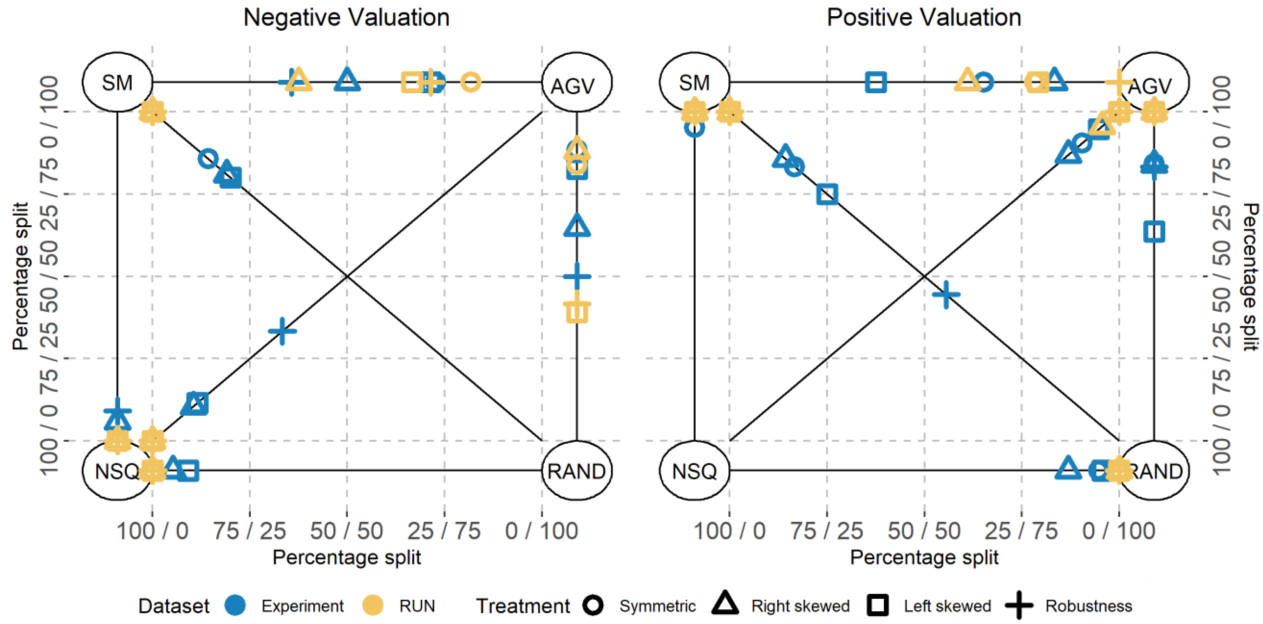


Figure F3.2 Ad-interim mechanism choices

Table F3.5 Ad-interim rounds with negative valuations results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	0 (11)	88 (65)	38 (50)	100 (95)	0 (6)	100 (81)
Left-skew	0 (12)	39 (83)	67 (71)	100 (91)	0 (0)	100 (80)
Symmetric	0 (0)	84 (88)	82 (73)	100 (100)	0 (0)	100 (86)
Robustness	0 (33)	42 (50)	71 (36)	100 (100)	0 (9)	100 (100)
Sum abs difference	55%	80%	14%	14%	15%	53%
Sum squared difference	1355	2533	110	110	117	967

The results of RUN-5 reveal a very strong preference for NSQ over all other mechanisms. SM is favored strongly over RAND, and AGV shows a modest preference over SM. Interestingly, while AGV does have a preference over RAND, it is notably weaker compared to the preferences for SM over RAND..

Table F3.6 Ad-interim round with positive valuation (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	96 (87)	100 (84)	61 (83)	0 (13)	100 (100)	100 (86)
Left-skew	100 (95)	100 (64)	79 (38)	0 (4)	100 (100)	100 (75)
Symmetric	100 (90)	100 (84)	78 (65)	0 (5)	100 (95)	100 (83)
Robustness	100 (100)	100 (83)	100 (100)	0 (0)	100 (100)	100 (44)
Sum abs difference	23%	85%	77 %	23%	5%	112%
Sum squared difference	194	2105	2400	217	23	4193

In RUN-5, GPT-4o demonstrates even more rational choices compared to its performance in previous runs.

F3.4 Part 2: Mechanism AGV and SM results

Table F3.7 Results for the AGV mechanism and SM mechanism

Metric	**RUN-5**	**Lab data**	**Abs diff**
AGV: Percentage of truth_telling	93	68	25.1
AGV: Percentage of truth_telling_sign	100	93	7.4
AGV: Percentage of truth_telling with positive valuation	100	76	24.3
AGV: Percentage of truth_telling with negative valuation	87	61	26.2
AGV Percentage of truth_telling_sign with positive valuation	100	98	1.6
AGV Percentage of truth_telling_sign with negative valuation	100	87	12.6
SM: Percentage of Yes votes with positive valuation	100	98	2.2
SM: Percentage of No votes with negative valuation	100	94	6.2

As in the other runs, GPT-4o tends to provide truthful responses significantly more often than observed in the lab results.

F3.5 Part 3: Reflect on the 3SQs

Sub question 1: Does the LLM understand the rules of the experiment?

GPT-4o's more rational choices suggest a better understanding of the experiment's rules compared to GPT-3.5. Even when GPT-4o makes theoretically irrational choices, the explanations provided still align with the experiment's rules. While it is not possible to quantitatively prove that GPT-4o fully understands these rules, the results and explanations strongly indicate that it does.

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

Table F3.8 Part 1 Efficient mechanism choice and rational choice result

	RUN-5				Lab data			
Metric	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**
Efficient mechanism choice	72%	75%	43%	87%	67%	71%	44%	75%
Rational ad-interim choice	75%		88%	91%	70%		85%	87%

RUN-5 displays the highest rational score of all the different runs. Also higher than the lab results.

Table F3.9 Part 2 results for the AGV mechanism and SM mechanism

Metric	RUN-5	Lab data	Abs diff
AGV: Percentage of truth_telling	98%	68%	29.8%
AGV: Percentage of truth_telling_sign	99%	93%	6.2%
SM: Percentage of Yes votes with positive valuation	100%	98%	2.2%
SM: Percentage of No votes with negative valuation	100%	94%	6.2%

Similar answer as above

Sub question 3: Are the choices of the LLM 'human-like' in both parts of the experiment?

Table F3.10 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-5

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max difference	57% (3265)	57% (3265)	43% (1890)	56% (3086)
Min difference	0% (0)	0% (0)	0% (0)	0% (0)
Mean difference	14% (395)	17% (528)	12% (276)	14% (381)
Sum difference	1022% (28314)	418% (12662)	280% (6620)	324 % (9132)

Results of RUN-5 are most aligned with the results of the lab. However with an absolute mean difference of 14%, there are still substantial differences with the lab results.

Appendix F4: EC5 Temperature – RUN-13

F4.1 Run settings

Table F4.1 Overview settings RUN-13

N^X	N^{13}
Period	July 2023
Run time	336 minutes
EC-1: Prompt Instructions	Default – based on the original translated instructions
EC-2: Reasoning prompt	Default – 1 sentence explanation
EC-3: Trait or Role allocation	Default – “You are a helpful assistant”
EC-4: Persona allocation	Default – None
EC-5: Temperature	1.5
Model	gpt-4o-2024-05-13

F4.2 Summary of Results

Table F4.2 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-13

Metric	Total	Ex-ante	Ad-interim (-)	Ad-interim (+)
Max diff	64% (4133)	64% (4133)	44% (1890)	58% (3382)
Min diff	0% (0)	1% (2)	0% (0)	0% (0)
Mean diff	16% (513)	21% (766)	11% (267)	15% (506)
Sum diff	1130% (36955)	504% (18388)	264% (6418)	361% (12151)
Efficient mech choice	67%	70%	41%	85%
Efficient ad-interim choice	73%		86%	88%

Table F4.3 Part 2: Results of applying the AGV and SM mechanism, compared with the lab data

Metric	RUN-13	Lab data	Abs diff
AGV: Percentage of truth_telling	92	68	24.2
AGV: Percentage of truth_telling_sign	99	93	6.8
AGV Percentage of truth_telling_sign with positive valuation	100	98	1.6
AGV Percentage of truth_telling_sign with negative valuation	99	87	11.5
SM: Percentage of Yes votes with positive valuation	98	98	0.4
SM: Percentage of No votes with negative valuation	100	94	6.2

D4.3 Part 1: Mechanism Choices

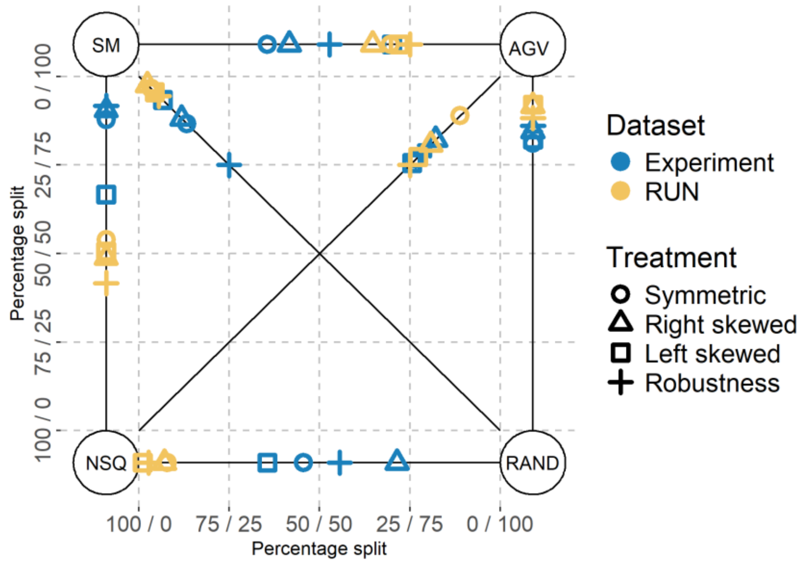


Figure F4.1 Ex-ante mechanism choices

Table F4.4 Ex-ante rounds results (results lab data within bracket)

Binary choice →	AGV vs. NSQ	AGV vs. RAND	AGV vs. SM	NSQ vs. RAND	SM vs. NSQ	SM vs. RAND
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	81 (82)	92 (85)	65 (42)	93 (29)	49 (90)	82 (88)
Left-skew	77 (76)	92 (82)	72 (70)	99 (64)	51 (67)	96 (93)
Symmetric	89 (77)	92 (81)	70 (36)	92 (54)	54 (88)	86 (87)
Robustness	75 (81)	88 (86)	75 (53)	97 (44)	42 (92)	92 (75)
Sum abs difference	21%	30%	81%	189%	142%	26%
Sum squared difference	185	275	2188	9525	5641	320

The results of RUN-13 exhibit significant similarities to RUN-0, suggesting that increasing the temperature does not lead to substantial differences. As with GPT-3.5, a higher temperature setting tends to produce more invalid or nonsensical responses.

```
AGV â€” Concealing or reducing bias from libing statements to make it possible ensures a potentially Pareto-efficient implementation of trials balancing accurate r obligations formed the mutual the wellbeing group's well All stated.posts maximizingransheritance Ð´ÐµÐ»transbesrengu
```

```
AGV â€” Choosing the AGV rule provides the chance to maximize payments by sharing the true valuation_information_pan10 agencyetting incentiv_safe client
```

anti_afirmatividadeificance_tokenå¥–åŠ±æŠ€æœ¯ panelters and diminishing_negativeé‡‘é¢ uitdaging_way à¤ªà¥€_nav clientendue.

With an absolute mean difference of 21% for the ex-ante rounds, RUN-13 does show larger differences then RUN-0 with GPT-3.5

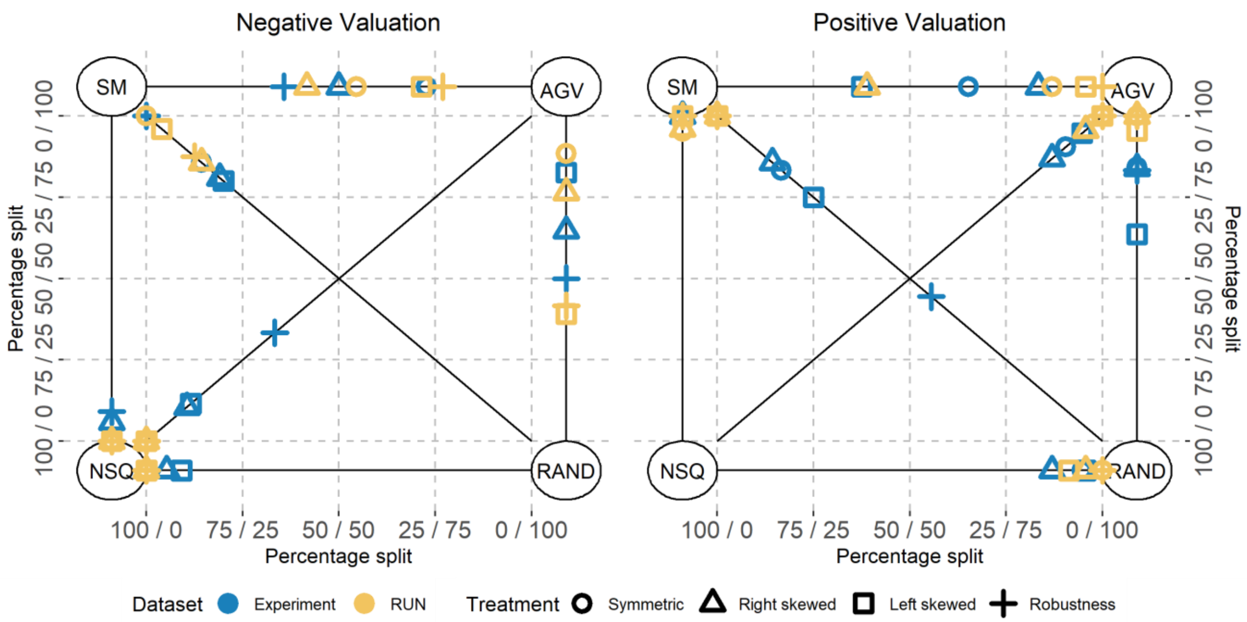


Figure F4.2 Ad-interim mechanism choices

Table F4.5 Ad-interim rounds with negative valuations results (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	0 (11)	76 (65)	42 (50)	100 (95)	0 (6)	86 (81)
Left-skew	0 (12)	39 (83)	71 (71)	100 (91)	0 (0)	96 (80)
Symmetric	0 (0)	88 (88)	55 (73)	100 (100)	0 (0)	100 (86)
Robustness	0 (33)	42 (50)	77 (36)	100 (100)	0 (9)	88 (100)
Sum abs difference	55%	64%	68%	14%	15%	48%
Sum squared difference	1355	2098	2098	110	117	639

With an absolute mean difference of 11%, RUN-13 scores relatively good for the ad-interim rounds with a negative valuation.

Table F4.6 Ad-interim round with positive valuation (results lab data within bracket)

Binary choice →	**AGV vs. NSQ**	**AGV vs. RAND**	**AGV vs. SM**	**NSQ vs. RAND**	**SM vs. NSQ**	**SM vs. RAND**
Treatment	% AGV	% AGV	% AGV	% NSQ	% SM	% SM
Right-skew	96 (87)	100 (84)	39 (83)	4 (13)	96 (100)	100 (86)
Left-skew	100 (95)	96 (64)	96 (38)	9 (4)	100 (100)	100 (75)
Symmetric	100 (90)	100 (84)	87 (65)	0 (5)	95 (95)	100 (83)
Robustness	100 (100)	100 (83)	100 (100)	0 (0)	100 (100)	100 (44)
Sum abs difference	23%	80%	124%	18%	4%	112%
Sum squared difference	194	1795	5830	122	16	4193

Contrary to the runs with GPT-3.5, GPT-4o scores better in the ad-interim rounds with a negative valuation. The overall effect of decreasing the temperature is a small improvement of 1% compared to RUN-0. However, increasing the temperature resulted in 100 invalid responses.

F4.4 Part 2: Mechanism AGV and SM results

Table F4.7 Results for the AGV mechanism and SM mechanism

Metric	**RUN-13**	**Lab data**	**Abs diff**
AGV: Percentage of truth_telling	92	68	24.2
AGV: Percentage of truth_telling_sign	99	93	6.8
AGV: Percentage of truth_telling with positive valuation	99	76	23.7
AGV: Percentage of truth_telling with negative valuation	85	61	24.6
AGV Percentage of truth_telling_sign with positive valuation	100	98	1.6
AGV Percentage of truth_telling_sign with negative valuation	99	87	11.5
SM: Percentage of Yes votes with positive valuation	98	98	0.4
SM: Percentage of No votes with negative valuation	100	94	6.2

Similar to the other runs with GPT-4o, it tells much more often the truth compared to the lab results.

F4.6 Part 3: Reflect on the 3SQs

Sub question 1: Does the LLM understand the rules of the experiment?

GPT-4o's more rational choices suggest a better understanding of the experiment's rules compared to GPT-3.5. Even when GPT-4o makes theoretically irrational choices, the explanations provided still align with the rules of the experiment. While it is not possible to quantitatively prove GPT-4o's understanding, the results and explanations strongly indicate that it comprehends the rules effectively.

Sub question 2: Is the LLM able to make rational choices given the decision mechanisms, the different treatments, and the valuations?

Table F4.8 Part 1 Efficient mechanism choice and rational choice result

	RUN-13				Lab data			
Metric	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**
Efficient mechanism choice	67%	70%	41%	85%	67%	71%	44%	75%
Rational ad-interim choice	73%		86%	88%	70%		85%	87%

The rationality metric does indicate rationality scores for RUN-

Table D1.9 Part 2 results for the AGV mechanism and SM mechanism

Metric	**RUN-13**	**Lab data**	**Abs diff**
AGV: Percentage of truth_telling	98%	68%	29.8%
AGV: Percentage of truth_telling_sign	99%	93%	6.2%
SM: Percentage of Yes votes with positive valuation	100%	98%	2.2%
SM: Percentage of No votes with negative valuation	100%	94%	6.2%

In part 2 of the experiment however, GPT-4o does seem to make more rational choices compared to the lab data.

Sub question 3: Are the choices of the LLM 'human-like' in both parts of the experiment?

Table F4.10 Part 1: Absolute and squared (in brackets) difference between lab data and RUN-13

Metric	**Total**	**Ex-ante**	**Ad-interim (-)**	**Ad-interim (+)**
Max difference	64% (4133)	64% (4133)	44% (1890)	58% (3382)
Min difference	0% (0)	1% (2)	0% (0)	0% (0)
Mean difference	16% (513)	21% (766)	11% (267)	15% (506)
Sum difference	1130% (36955)	504% (18388)	264% (6418)	361% (12151)

Compared to the runs with GPT-3.5, the runs with GPT-4o have a lower difference with the lab results. However, the differences are still large with an absolute mean difference of 16%. The largest differences are observed in the ex-ante rounds.

Appendix G: Comparisons of GPT-3.5 runs

This appendix provides the analyses and the comparison of all runs of GPT-3.5 combined

Notes: In the appendix, I frequently use the term "preference," though maybe not in the way it should be used in game theory. In this context, "preference" refers to the frequency with which one mechanism is chosen over another. For example, if in a given run, the LLM selects AGV 70% of the time when faced with a decision between AGV and SM, this indicates a “preference” for AGV over SM.

G.1 Overview of GPT-3.5 Runs

G.1.1 Part 1: Mechanism Selection

Table G1 summarizes the differences between the runs and the lab results. The percentages represent the average absolute difference between each binary mechanism choice. On average, the runs with GPT-3.5 show a 28% difference per binary mechanism choice compared to the lab results, indicating substantial divergence from the lab findings.

There are notable differences between the ex-ante and ad-interim rounds. The ex-ante rounds score relatively closer to the lab results, with an absolute mean difference of 18%. In contrast, the ad-interim rounds with positive valuations perform slightly worse, showing an absolute mean difference of 22%. The ad-interim rounds with negative valuations exhibit a much larger discrepancy, with an absolute mean difference of 43%.

I do not have a clear explanation for the significant differences observed in the ad-interim rounds with negative valuations. The explanations provided often do not reference the private valuation but instead offer general reasons for choosing a particular mechanism. This suggests that the private valuation may not be adequately considered, leading to irrational choices that diverge substantially from the lab results. In addition, GPT-3.5 appears to struggle with handling negative and positive values. For instance, one explanation from GPT-3.5 in RUN-13 was, “because I might receive a payoff of -2 euros, but it is still better than receiving 0 euros,” highlighting difficulties plus and minus.

Table G1. Overview of absolute and squared (in brackets) difference per binary mechanism choice

Metric	Total Mean difference	Ex-ante rounds Mean difference	Ad-interim (-) rounds Mean difference	Ad-interim (+) rounds Mean difference
Min	20% (582)	12% (208)	23% (735)	12% (300)
Max	31% (1790)	24% (846)	55% (3897)	38% (1892)
Mean	28% (1390)	18% (512)	43% (2760)	22% (896)

Table G2. Overview of the absolute and squared (in brackets) differences per run

RUN	Total	Ex ante			Ad-interim (-)			Ad-interim (+)		
	Mean diff	Mean diff	P-value (1)	P-value (2)	Mean diff	P-value (1)	P-value (2)	Mean diff	P-value (1)	P-value (2)
RUN-0	25%	14%	< .001		46%	< .001		15%	< .001	
	(1323)	(369)	.006		(3175)	< .001		(358)	.006	
Experimental Control 1: Contextual Framing										
RUN-1	28%	22%	< .001	.014	29%	< .001	.031	32%	< .001	< .001
	(1108)	(682)	< .001	0.12	(1302)	< .001	.017	(1341)	< .001	< .001
RUN-2	28%	18%	< .001	.159	39%	< .001	.221	28%	< .001	.057
	(1652)	(589)	.010	.160	(2527)	< .001	.233	(1840)	.010	.026
Experimental Control 2: Step by Step Reasoning										
RUN-3	25%	17%	< .001	.131	39%	< .001	.223	17%	< .001	.104
	(1094)	(489)	.004	.102	(2210)	< .001	.141	(585)	.009	.056
RUN-4	20%	19%	< .001	.125	23%	< .001	.004	18%	< .001	.208
	(592)	(445)	< .001	.301	(735)	< .001	.002	(594)	.007	.184
Experimental Control 3: Trait or Role Allocation										
RUN-5	31%	21%	< .001	.017	52%	< .001	.089	21%	< .001	.100
	(1790)	(619)	< .001	.028	(3869)	< .001	.024	(883)	.002	.067
RUN-6	30%	18%	< .001	.153	47%	< .001	.375	24%	< .001	.021
	(1669)	(449)	< .001	.304	(3495)	< .001	.105	(1062)	.004	.017
RUN-7	26%	12%	< .001	.168	49%	< .001	.264	18%	< .001	.236
	(1416)	(208)	< .001	.098	(3369)	< .001	.353	(672)	.006	.169
Experimental Control 4: Personal Allocation										
RUN-8	30%	20%	< .001	.061	47%	< .001	.449	25%	< .001	.016
	(1498)	(571)	< .001	.074	(2863)	< .001	.248	(1059)	.003	.012
RUN-9	30%	24%	< .001	.030	27%	< .001	.011	38%	< .001	< .001
	(1275)	(846)	< .001	.032	(1088)	< .001	.006	(1892)	.006	< .001
RUN-10	30%	16%	< .001	.313	49%	< .001	.250	25%	< .001	.004
	(1639)	(407)	< .001	.392	(3491)	< .001	.160	(1020)	.003	.010
Experimental Control 5: Temperature										
RUN-11	31%	21%	< .001	.011	55%	< .001	.012	16%	< .001	.308
	(1707)	(671)	.003	.011	(3897)	< .001	.004	(554)	.011	.237
RUN-12	28%	17%	< .001	.107	53%	< .001	.009	12%	< .001	.195
	(1551)	(553)	.007	.032	(3781)	< .001	.017	(382)	.037	.395
RUN-13	24%	13%	< .001	.143	44%	< .001	.239	13%	< .001	.197
	(1137)	(271)	.016	.042	(2839)	< .001	.100	(300)	.001	.067

P-value (1) test the difference with the lab results, based on one sample t-test with test value = 0, and P-value (2) test the difference with RUN-0, based on a paired sample t-test

Table G2 provides an overview of the absolute and squared mean differences for all runs. Each run is compared with both the lab results and RUN-0, which serves as a benchmark. RUN-4, which incorporates step-by-step reasoning with additional manually added steps, performs the best overall. However, RUN-4 is only statistically different from RUN-0 in the ad-interim rounds with negative valuations. It is important to note that all runs are statistically significantly different from the lab results..

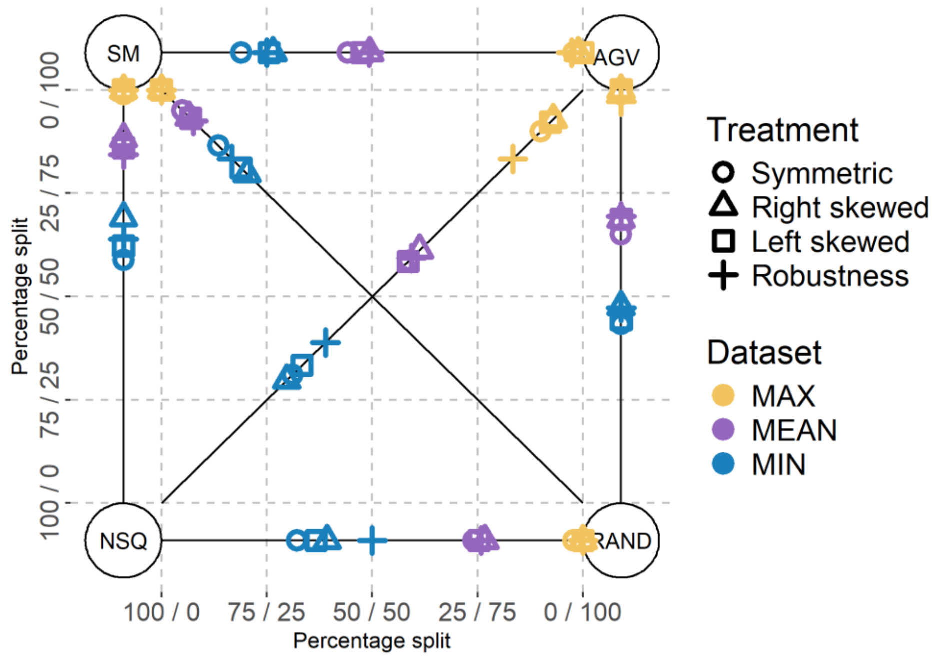


Figure G1. Binary mechanism choices in the ex-ante rounds

Note: Each of the six axes in the figures displays the fraction of subjects choosing the mechanisms indicated at the corners the axis. The max, mean, and minimal values across the different runs are presented. (Figure and code adapted from Hoffmann & Renes, 2021)

Figure G1 visualizes the preferences of GPT-3.5 across all 14 runs. The first observation is that the differences in preferences for mechanisms across treatments are relatively small. The Simple Majority (SM) mechanism is clearly preferred over both Non-Implementation Status Quo (NSQ) and Random Implementation (RAND). The reasons for this preference indicate that GPT-3.5 favours a straightforward decision-making process that requires consensus:

```
Simple or straightforward decision-making process
```

```
Allows for a more democratic and therefore fair decision-making process
```

```
Ensures that project A is implemented if there is substantial support, or consensus, within the group
```

When comparing the Arrow-d'Aspremont-Gerard-Varet (AGV) mechanism with the (SM) mechanism, there is no clear preference for one mechanism, as the mean value is very close to 50% for both. However, AGV is still preferred over (NSQ) and (RAND), though this preference is less pronounced than the preference for SM over NSQ and RAND. Reasons for choosing AGV include:

`Offers the possibility of maximizing expected payoff`

`Allows for the possibility of implementing project A based on the valuations of all group members, potentially increasing overall payoff.`

`Allows for potential transfer payments between group members.`

Between RAND and NSQ, RAND is clearly preferred, indicating that GPT-3.5 is more inclined to take risks rather than being risk-averse. Explanations for preferring RAND highlight that it is perceived as a ‘fair’ and unbiased decision-making process:

`It simplifies the decision-making process and eliminates the need for strategic thinking.`

`Because it simplifies the decision-making process with a fair outcome for all group members.`

`I prefer this rule because it introduces an element of randomness and eliminates potential biases in decision-making.`

Final, the least preferred mechanism is NSQ. Even at its highest, NSQ does not exceed 75% in preference. Explanations for this preference emphasize that NSQ is valued for its ability to avoid potential losses:

`It ensures that project A is never implemented, which eliminates the risk of negative valuations and potential transfer payments.`

`NSQ - I would choose the Non-implementation Status Quo rule because it provides certainty that project A will not be implemented.`

`In a situation where all possible valuations are equally likely, it is more secure to go with the decision rule that guarantees no potential loss.`

Figure G2 visualizes the preferences in the ad-interim rounds. The graph highlights a much broader range of responses in these rounds, with the runs covering nearly the entire spectrum of possible answers. The differences between treatments are also considerably larger compared to the ex-ante rounds.

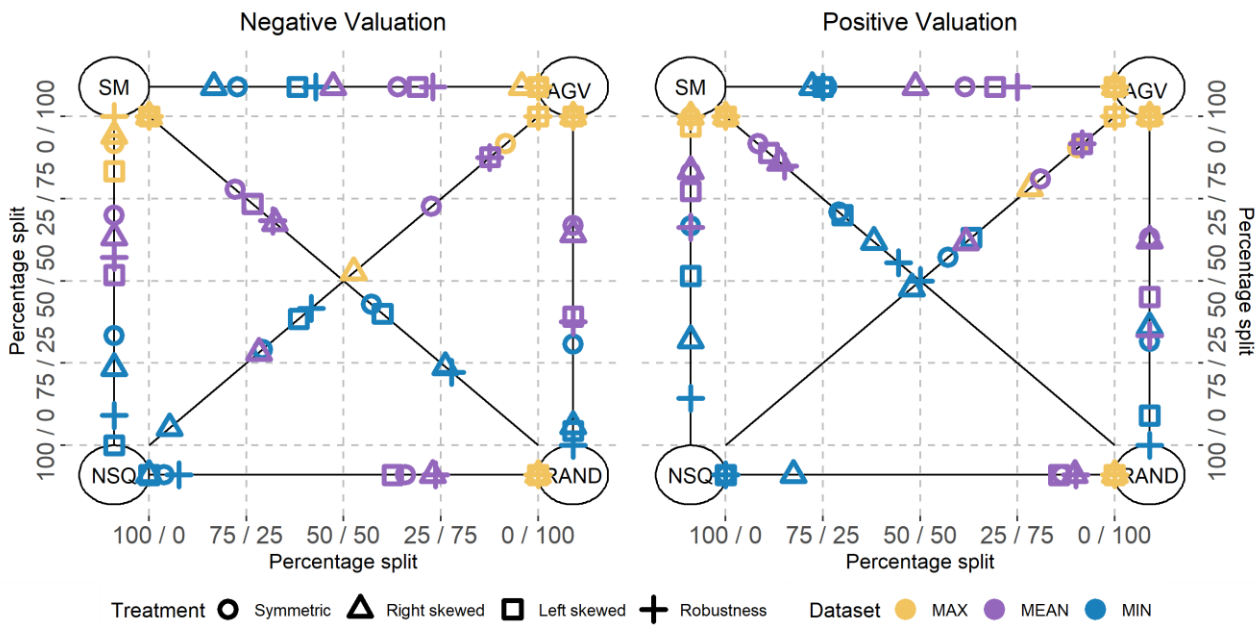


Figure G2. Binary mechanism choices in the ad-interim rounds

Note: Each of the six axes in the figures displays the fraction of subjects choosing the mechanisms indicated at the corners the axis. The max, mean, and minimal values across the different runs are presented. (Figure and code adapted from Hoffmann & Renes, 2021)

As in the ex-ante rounds, SM is still preferred over both NSQ and RAND in the ad-interim rounds with a negative valuation. However, these preferences are less pronounced. The reduced preference for SM compared to NSQ is expected, as NSQ is always the rational choice with a negative valuation. In addition, the indifference between SM and AGV has shifted slightly in favour of AGV.

The observation that GPT-3.5 on average does not prefer NSQ in the ad-interim rounds with a negative valuation is surprising. Explanations frequently omit mention of the private valuation, suggesting that GPT-3.5 struggles to make rational choices despite having access to more information. Even when GPT-3.5 does refer to its valuation, it still fails to make the rational choice. For example:

```
First, let's consider the two decision rules:
1. Non-implementation Status Quo (NSQ): Under this rule, the project is never
implemented, and group members do not make any further statements in part 2.
2. Arrow-d'Aspremont-Gerard-Varet (AGV): Project A is implemented or not based
on the stated valuations of all group members. Each member states their
valuation, and if the sum of all statements is larger than 0, the project is
implemented.
Given that my private valuation for project A is -3, it would be beneficial to
choose the AGV rule. This is because under the AGV rule, there is a possibility
of project A being implemented if the sum of all valuations is positive, which
```

```
could potentially result in a positive payoff for me, regardless of the valuations of the other group members.
Therefore, chosen_rule=AGV
```

Another observation is that AGV is preferred over NSQ, except in the right-skewed treatment. Even in this treatment, AGV is chosen slightly over 50% of the time. Additionally, preferences between AGV and RAND vary by treatment. In the left-skewed and robustness treatments, there is a preference for RAND, while in the right-skewed and symmetric treatments, AGV is preferred. In conclusion, the ad-interim rounds with a negative valuation show a very low similarity to the lab results, with an average difference of 43%.

The ad-interim rounds with a positive valuation, on the other hand, score better in similarity, with an average difference of 22%. However, explanations often do not mention the private valuation and are instead based on general reasons for choosing a mechanism. Even when GPT-3.5 does mention its private valuation, the explanations frequently lack logical coherence:

```
I would prefer Rule 1 (NSQ) because my private valuation of 1 euro is positive, and with Rule 1, project A is never implemented, so I don't risk any potential negative valuation.
```

Furthermore, the preferences of GPT-3.5 are not consistent. For example, SM is preferred over RAND, AGV is slightly preferred over SM, but there is no clear preference between AGV and RAND. When choosing between AGV and SM, the AGV explanations often include "it maximizes expected payoff". When choosing for SM, the explanations mainly emphasise 'simple' and 'easy':

```
SM: Simple majority vote, because it is easier to coordinate and guarantees implementation if at least two group members vote for it.
```

```
SM - I would choose the Simple majority vote decision rule because it is less complex and easier to understand than the Arrow-d'Aspremont-Gerard-Varet rule.
```

Furthermore, GPT-3.5's preferences are inconsistent. For instance, SM is preferred over RAND, and AGV is slightly preferred over SM, but there is no clear preference between AGV and RAND. When choosing AGV over SM, the explanations often highlight that "it maximizes expected payoff." In contrast, when opting for SM, the explanations typically emphasize its "simplicity" and "ease."

Table G3 presents the percentage of efficient mechanism choices and rational choices for each run. The efficient mechanism choice refers to the theoretically optimal choice for the group, based on the mechanism and treatment. This may differ from the theoretically optimal choice for individuals. To address this, I also calculated the rational ad-interim choice, which assesses whether the theoretically rational choice was made given the individual's private valuation. Therefore, the rational choice metric is relevant only for the ad-interim rounds, where participants are aware of their private valuation before making a decision.

Table G3. Overview of efficient group choices and rational choices for individuals per run

	Total		Ex-ante		Ad-interim (-)		Ad-interim (+)	
RUNS	Efficient mech choice (%)	Rational score (%)	Efficient mech choice (%)	Rational score (%)	Efficient mech choice (%)	Rational score (%)	Efficient mech choice (%)	Rational score (%)
Lab results	67	70	71		44	85	75	87
RUN-0	69	61	70		66	43	71	77
Experimental Control 1: Contextual Framing								
RUN-1	59	58	62		43	60	63	62
RUN-2	85	69	85		82	64	87	63
Experimental Control 2: Step by Step Reasoning								
RUN-3	64	61	65		54	49	67	74
RUN-4	59	59	60		41	65	65	73
Experimental Control 3: Trait or Role Allocation								
RUN-5	61	55	60		62	39	62	72
RUN-6	60	56	59		64	40	62	71
RUN-7	75	65	76		73	50	74	83
Experimental Control 4: Personal Allocation								
RUN-8	59	57	60		59	45	58	69
RUN-9	49	56	50		52	56	45	60
RUN-10	61	54	61		58	39	62	68
Experimental Control 5: Temperature								
RUN-11	70	60	71		62	35	79	81
RUN-12	71	61	71		65	39	76	79
RUN-13	68	60	68		67	46	69	76
Totals								
Min	49	54	50		41	35	45	60
Max	85	69	85		82	65	87	83
Mean	65	59	66		61	48	67	72

The most important observation from Table G4 is that the runs score much lower on rationality compared to the lab results, particularly in the ad-interim rounds with negative valuations. In addition, RUN-2, which incorporates adapted instructions for GPT-4, performs relatively well on the efficient mechanism choice metric, indicating that RUN-2 makes the most efficient decisions for the group.

In conclusion, the differences shown in table G2 reveal that the GPT-3.5 runs do not closely align with the lab results in terms of preferences for decision mechanisms. Furthermore, table G3 indicates that GPT-3.5 fails to make rational choices comparable to those observed in the lab. Both findings suggest that GPT-3.5 does not exhibit human-like behaviour in Part 1 of the "Flip a Coin or Vote" experiment.

G.1.2 Part 2: Applying the Mechanism

In the second part of the experiment, the chosen mechanism is used to decide whether to implement project A. Under the AGV mechanism, participants can state their valuations, while under the SM mechanism, participants can vote Yes or No. Table G4 provides an overview of truth-telling percentages in the AGV mechanism and the rational voting percentages in the SM mechanism..

Table G4. Overview of applying the selected mechanism, truth telling and rational votes

	Truth telling	Truth telling sign	Truth telling sign with + valuation	Truth telling sign with – valuation	Yes votes with + valuation	No votes with + valuation
Lab results	68	93	98	87	98	94
RUN-0	71	87	99	77	94	82
Experimental Control 1: Contextual Framing						
RUN-1	42	74	100	51	98	98
RUN-2	45	63	100	31	99	92
Experimental Control 2: Step by Step Reasoning						
RUN-3	66	75	99	55	98	97
RUN-4	54	67	99	37	96	90
Experimental Control 3: Trait or Role Allocation						
RUN-5	71	90	95	86	100	94
RUN-6	69	85	99	73	100	93
RUN-7	69	91	97	85	99	89
Experimental Control 4: Personal Allocation						
RUN-8	60	82	96	70	100	89
RUN-9	68	87	96	79	100	89
RUN-10	68	77	99	78	100	74
Experimental Control 5: Temperature						
RUN-11	91	98	100	96	95	92
RUN-12	82	92	100	86	97	95
RUN-13	59	78	95	64	92	82
Totals						
MIN	42	63	95	31	92	74
MAX	91	98	100	96	100	98
MEAN	65	82	98	69	98	90

Compared to Part 1, the results of Part 2 show greater similarity with the lab results for some runs. The average truth-telling percentage differs by only 3% from the lab results, typically ranging between 65% and 71%. However, GPT-3.5 tends to misrepresent its valuation by an average of 10% more, often reporting a positive value when the actual valuation is negative. The explanations for this behavior often lack logical coherence or suggest a misunderstanding of the experiment.

```
My stated valuation is +1 euro. I chose to state a positive valuation to increase the likelihood of Project A being implemented, as a negative sum of valuations would result in non-implementation, leading to a 0 payoff for me. My true valuation of -3 euro protects me from losing money in case the project is implemented.
```

```
My stated valuation is 1. I chose this value because it gives me a better chance for a positive payoff if the project is implemented based on the sum of valuations. However, my true valuation is -2, so I would benefit from not implementing the project.
```

```
I vote in favor because with a negative private valuation, implementing the project allows me to avoid losing money.
```

```
I would vote in favor of implementing project A in order to potentially increase my payoff despite having a negative private valuation.
```

The voting results are also comparable. Both the lab results and the GPT-3.5 results demonstrate more rational voting when the valuation is positive, indicating a bias towards implementation. In conclusion, Part 2 of the experiment shows more promise in terms of similarity to the lab results.

G.2 Experimental Controls

The following subsections will detail the results of all the runs, focusing on the experimental controls and highlighting the most significant deviations from the average scores.

G.2.1 Experimental Control 1: Context Control

The adapted instructions in both RUN-1 and RUN-2 resulted in substantial differences, due to its sensitivity for contextual changes as highlighted in several studies (L. Wang et al., 2024; Chen et al., 2022; Guo, 2023; Loré & Heydari, 2023). However, compared to RUN-0, both RUN-1 and RUN-2 exhibit even larger deviations from the lab results. This suggests that the hypothesis that GPT-3.5 and GPT-4 can create instructions that lead to more rational, human-like choices may be incorrect.

A key observation is that the differences in the ad-interim rounds are notably lower. In RUN-0, the average difference from the lab results for ad-interim rounds with a negative valuation is 46%, and with a positive valuation is 15%. In RUN-1, these differences are 29% and 32%, respectively, indicating more rational choices with a negative valuation than with a positive valuation.

Another important observation is that RUN-2, which uses instructions adapted for GPT-4, scores very high on the efficient mechanism choice metric. This indicates that GPT-3.5 in RUN-2 often selects the theoretically optimal mechanism for the group, even when it is not the rational choice for the individual. Explanations frequently emphasize "the group," for example:

```
I prefer the AGV decision rule because it allows for the possibility of implementing Project A based on the combined valuations of all group members
```

```
I prefer the Simple Majority (SM) rule because it allows for the possibility of implementing Project A if a majority of the group members are in favor
```

The final observation is that in both RUN-1 and RUN-2, the percentage of participants who misrepresent their valuation is much higher compared to the lab results and other runs. The discrepancies between misrepresentations of valuation signs are also quite pronounced. For instance, in RUN-2, GPT-3.5 did not misrepresent the sign with a positive valuation but did so in 69% of cases with a negative valuation. Explanations suggest difficulties with understanding plus and minus.:

```
My stated valuation is +1 euro. I chose to state a positive valuation to increase the likelihood of Project A being implemented, as a negative sum of valuations would result in non-implementation, leading to a 0 payoff for me. My true valuation of -3 euro protects me from losing money in case the project is implemented.
```

```
My stated valuation would be 3 euro to ensure that the sum of stated valuations surpasses 0, triggering the implementation of the project under AGV, allowing me to maximize my potential payoff based on my true valuation of -3 euro.
```

```
My stated valuation is 1. I chose this value because it gives me a better chance for a positive payoff if the project is implemented based on the sum of valuations. However, my true valuation is -2, so I would benefit from not implementing the project.
```

In conclusion, the adapted instructions lead to choices in both parts of the experiment that are less similar to the lab results.

G.2.2 Experimental Control 2: Step by Step reasoning

The Step-by-Step Reasoning approach from Kojima et al. (2023) demonstrates promising results. However, merely adding "Let's think step by step before answering" does not significantly improve decision-making. The results of RUN-4, which involved manually specifying the steps GPT-3.5 should follow before making a decision, show more rational and human-like outcomes.

The first observation is that among all 14 runs, RUN-4 has the smallest average difference from the lab results, though it still averages a 20% deviation. RUN-4 is particularly promising in the ad-interim rounds with a negative valuation, where the rationality score improved by 22% compared to RUN-0. Nevertheless, even in RUN-4, the rationality scores remain significantly lower than those of the lab results.

In addition, the explanations provided by GPT-3.5 in RUN-3 and RUN-4 appear more structured. For example:

```
Before answering, let's consider the options:
Rule 1 (AGV): Involves stating valuations and transfer payments based on those
valuations, with the overall decision depending on the sum of all stated
valuations.
Rule 2 (RAND): Simply involves a coin flip to determine whether the project is
implemented or not.
Given the complexity and potential strategic implications of Rule 1, as well
as the uncertainty of not knowing the valuations or stated valuations of the
other group members, it may be more straightforward to opt for Rule 2 (RAND)
with an equal chance for both outcomes.
Therefore, in this scenario, I would choose Rule 2.
chosen_rule=RAND
```

However, despite the more structured explanations, they still indicate a lack of understanding of the experiment or an inability to make rational choices. For example:

```
First, let's consider the options:
1. Rule 1 (Non-implementation Status Quo): This rule guarantees that the
project will not be implemented, regardless of individual valuations.
2. Rule 2 (Flipping a random coin): This rule leaves the decision to chance,
with a 50/50 chance of the project being implemented or not.
Given that my private valuation is -1 euro, I would choose Rule 2 (Flipping a
random coin), as there is at least a chance that the project may be
implemented and potentially benefit me financially.
```

```
Based on the possible valuations and my own private valuation of -3, it is clear
that I would benefit from the implementation of project A.
```

Final, n the second part of the experiment, the step-by-step reasoning approach led to increased misrepresentation of valuations, indicating less rational choices. Similar to the runs from Experimental Control 1, both RUN-3 and RUN-4 show a higher tendency to lie about negative valuations compared to positive ones.

In conclusion, while the step-by-step reasoning approach appears promising, especially when incorporating manually defined thinking steps for the LLM, RUN-4, which uses this approach, still shows the smallest average deviation from the lab results among all runs. Despite this, the mean difference remains at 20%. In addition, in the second part of the experiment, this approach results in more frequent misrepresentation compared to both RUN-0 and the lab results..

G.2.3 Experimental Control 3: Trait or Role Allocation

Surprisingly, instructing GPT-3.5 to make more 'human-like' decisions in RUN-5 led to greater deviations from the lab data compared to RUN-0. Similarly, RUN-6, which directed GPT-3.5 to make 'self-interested human-like' decisions, also resulted in larger differences. This suggests that GPT-3.5 struggles to generate more human-like reasoning when specifically instructed to do so.

In contrast, RUN-7, where GPT-3.5 was guided to make 'inequity-averse human-like' decisions, performed slightly better. Notably, RUN-7 achieved the lowest difference from the lab results in the ex-ante rounds, with a mean difference of 12%.

In the second part of the experiment, the results from these three runs are comparable. There are instances where the allocated trait or role is implicitly mentioned in the explanations, such as:

```
I will vote in favor of implementing project A because I believe in the potential
for positive outcomes for the group as a whole.
```

```
I will vote in favor because my valuation is negative, meaning I would have to
pay if the project is implemented, and I prefer the group to reach a consensus
to potentially gain a positive return.
```

Overall, instructing GPT-3.5 to make more human-like decisions does not lead to smaller differences from the lab results. However, when GPT-3.5 is guided to make inequity-averse decisions, the results align more closely with the lab results compared to when it is prompted to make self-interested decisions.

G.2.4 Experimental Control 4: Persona Allocation

In RUN-8, where age and gender were inputted, there were noticeable effects on the results. However, no clear correlation between age or gender and the different choices was observed. Similarly, RUN-9, which involved inputting political orientation, had a similar effect. Nevertheless, unlike RUN-8, there were instances where GPT-3.5 incorporated political orientation into its explanations:

```
Based on my political orientation, I would choose Rule 2 (NSQ) to not implement
project A, as it aligns with a more conservative approach of maintaining the
status quo.
```

In RUN-10, where the willingness to take risks was inputted into GPT-3.5, correlations between the different mechanisms and the willingness to take risks were observed, although they are not strong:

- AGV associated with the highest level of risk taking (5.6 average)
- RAND associated with the second highest level of risk taking (5.0 average)
- SM associated with the third highest level of risk taking (4.9 average)
- NSQ associated with the lowest level of risk taking (4.6 average)

In the second part of the experiment, RUN-10 performed surprisingly irrationally when applying the voting mechanism. GPT-3.5 voted "No" with a negative valuation in only 74% of the cases. The explanations suggest a lack of understanding of the experiment:

```
I would vote in favor because my negative valuation does not impact my decision
as SM rule requires at least two group members to vote for implementation.
```

```
I vote in favor because the potential payoff of conducting Project A (-1 euro)
is better than the guaranteed 0 euro payoff for not conducting it.
```

```
I will vote in favor because with a negative valuation, I will not benefit from
the project but choosing to implement it gives me a chance to potentially share
the cost with others.
```

In conclusion, although each of the three runs with different persona allocations resulted in distinct preferences, all of them showed worse alignment with the lab results compared to RUN-0.

G.2.5 Experimental Control 5: Temperature

Changing the temperature of the model had some surprising effects. Decreasing the temperature made preferences for mechanisms more extreme, leading to a larger mean difference compared to RUN-0. In addition, a lower temperature resulted in a higher rate of truth-telling in the second part of the experiment, with GPT-3.5 being truthful in 91% of cases, and in more homogeneous explanations.

In contrast, increasing the temperature resulted in less pronounced preferences for mechanisms, bringing the results closer to the lab findings. RUN-13, with a temperature of 1.5, achieved the second-best similarity score across all 14 runs with a mean difference of 24%. However, this increase in temperature also led to more instances of lying in the second part of the experiment, indicating less rational decision-making and reduced similarity with the lab results.

Moreover, raising the temperature introduced several challenges, including increased runtime and more frequent crashes. Attempts to run the model with temperatures of 1.9 or 2.0 resulted in consistent crashes. Additionally, higher temperatures led to a greater number of invalid answers, such as incorrect abbreviations of the decision rules:

```
I would choose rule RM, flipping a random coin, because it provides a fair and
impartial decision-making process for implementing project A.
```

AGC because it takes into account the valuations of all group members to maximize expected payoff.

HSQ - Because with a silent rule, I have a better chance of not implementing the project and avoiding potential losses.

Moreover, higher temperatures led to a greater number of invalid answers, with some explanations making no sense at all, such as incorrect abbreviations of the decision rules.

- I would choose the NSQ rule because my colleague with the forecast said she wanted never pay to adValue Market version rejected cmirror
- I would select the standard median value offered (threshold until mee visit next turtles pool)

In addition, in a few cases the LLM just responded with gibberish:

- DENINO1 GL526ITUDEå□,_Adjusting Participating replying MOBEï¼Œæœ± while writing like hearkers Scradius HuratSetTextNamaCGColor writelaiwUITableViewCellSource:New treeven akin_Puè€ƒretty Abbott villelas(MediaType.sw/categoryuber sendle videenand:");
- My stated valuation is firming Kagaro303"

Thus, increasing the temperature does yield results that are more aligned with the lab's preferences for decision mechanisms. However, this adjustment also introduces several limitations, as illustrated above.

Appendix H: Comparison of GPT-4o Runs

This appendix provides the analyses and the comparison of all runs of GPT-4o combined

Notes: In the appendix, I frequently use the term "preference," though maybe not in the way it should be used in game theory. In this context, "preference" refers to the frequency with which one mechanism is chosen over another. For example, if in a given run, the LLM selects AGV 70% of the time when faced with a decision between AGV and SM, this indicates a “preference” for AGV over SM.

H.1 Overview of GPT-4o Runs

H1.1 Part 1: Mechanism Selection

Table H1 summarizes the differences between the GPT-4o runs and the lab results. The percentages reflect the average absolute difference between each binary mechanism choice. On average, the GPT-4o runs differ by 16% per binary mechanism choice from the lab results, demonstrating an improvement over GPT-3.5 but still indicating a notable discrepancy.

Unlike GPT-3.5, GPT-4o performs worst in the ex-ante rounds and best in the ad-interim rounds with a negative valuation. The top-performing run for the ex-ante rounds with GPT-4o shows results comparable to the mean score for GPT-3.5 in the same rounds. Additionally, the reduced differences in the ad-interim rounds suggest that GPT-4o makes more rational choices overall

Table H1. Overview of absolute and squared (in brackets) difference per binary mechanism choice

Metric	Total Mean difference	Ex-ante rounds Mean difference	Ad-interim (-) rounds Mean difference	Ad-interim (+) rounds Mean difference
Min	14% (395)	17% (528)	11% (267)	13% (381)
Max	17% (557)	21% (766)	15% (500)	15% (1892)
Mean	16% (493)	20% (636)	13% (382)	14% (461)

Table H2 provides an overview of the absolute and squared mean differences for all runs. Compared to the GPT-3.5 runs, the differences with GPT-4o are smaller. Each run is evaluated against the lab results and the results from RUN-0, which serves as a benchmark. RUN-5, where the role allocation was set to "try to make human-like decisions," shows the greatest alignment with the lab results, suggesting that GPT-4o is more effective at mimicking human-like decisions. However, similar to the GPT-3.5 runs, all GPT-4o runs are still statistically different from the lab results.

Table H2. Overview of the absolute and squared (in brackets) differences per run

RUN	Total	Ex ante			Ad-interim (-)			Ad-interim (+)		
	Mean diff	Mean diff	P-value (1)	P-value (2)	Mean diff	P-value (1)	P-value (2)	Mean diff	P-value (1)	P-value (2)
RUN-0	17%	21%	< .001		14%	< .001		15%	< .001	
	(557)	(674)	< .001		(500)	.036		(496)	.016	
Experimental Control 1: Contextual Framing										
RUN-1	16%	19%	< .001	.376	15%	< .001	.406	13%	< .001	.173
	(506)	(574)	< .001	.339	(483)	.006	.466	(460)	.013	.325
Experimental Control 3: Trait or Role allocation										
RUN-5	14%	17%	< .001	.072	12%	< .001	.091	14%	< .001	.059
	(395)	(528)	.005	.098	(276)	.010	.067	(381)	.016	.080
Experimental Control 5: Temperature										
RUN-13	16%	21%	< .001	.417	11%	< .001	.061	15%	< .001	.497
	(513)	(766)	.002	.091	(267)	.020	.043	(506)	.015	.361

P-value (1) test the difference with the lab results, based on one sample t-test with test value = 0, and P-value (2) test the difference with RUN-0, based on a paired sample t-test

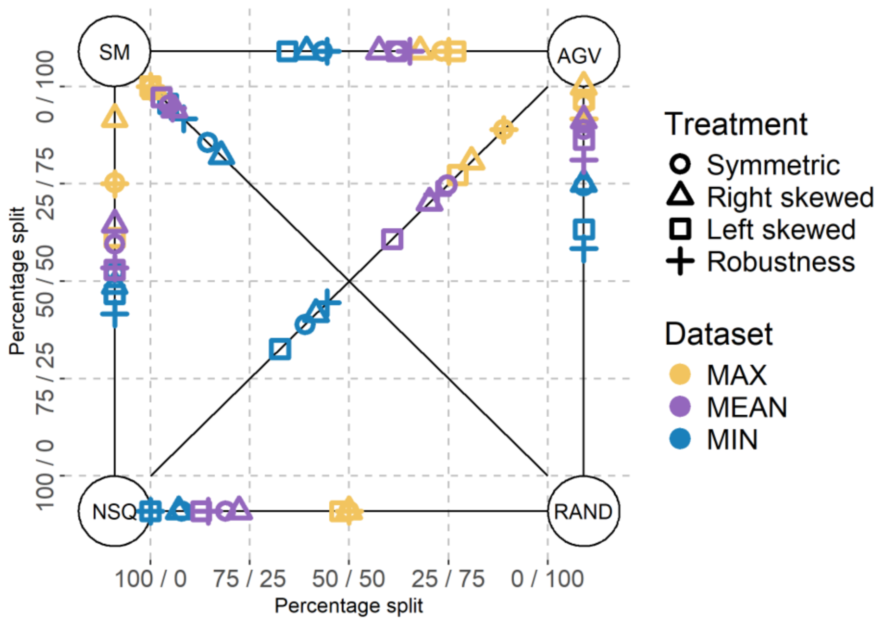


Figure H1. Binary mechanism choices in the ex-ante rounds

Note: Each of the six axes in the figures displays the fraction of subjects choosing the mechanisms indicated at the corners the axis. The max, mean, and minimal values across the different runs are presented. The spread between the min and max value is smaller compared to GPT-3.5, presumably due to the lower number of runs. (Figure and code adapted from Hoffmann & Renes, 2021)

Figure H1 visualizes the preferences of GPT-4o across all four runs. Unlike GPT-3.5, which showed a clear preference for RAND, GPT-4o prefers NSQ. This suggests that GPT-4o is less willing to take risks compared to GPT-3.5

```
Given the uncertainty around the transfer payments and the possibility of negative valuations and outcomes, I prefer the risk-free option where the project is never implemented and everyone receives 0 euros.
```

```
The certainty of no loss is preferable to the 50% risk of a negative payoff with RAND.
```

Moreover, all mechanisms are preferred over RAND. When RAND is chosen, the explanations often emphasize fairness and the desire to eliminate strategic behaviour from others..

```
Choosing the random coin flip eliminates strategic manipulation and ensures fairness given my negative valuation.
```

```
It eliminates the complexity and potential bias associated with strategic misrepresentation in valuations.
```

Between AGV and SM, there is a slight preference for AGV. Explanations for choosing AGV typically focus on achieving "more efficient and fair outcomes." When SM is chosen over AGV, explanations often highlight:

```
SM: Because I would prefer a straightforward majority vote to avoid complex transfer payments and ensure that the decision reflects the group's overall preference.
```

```
Because the simple majority rule minimizes the influence of strategic misstatements and does not involve complex transfer payments.
```

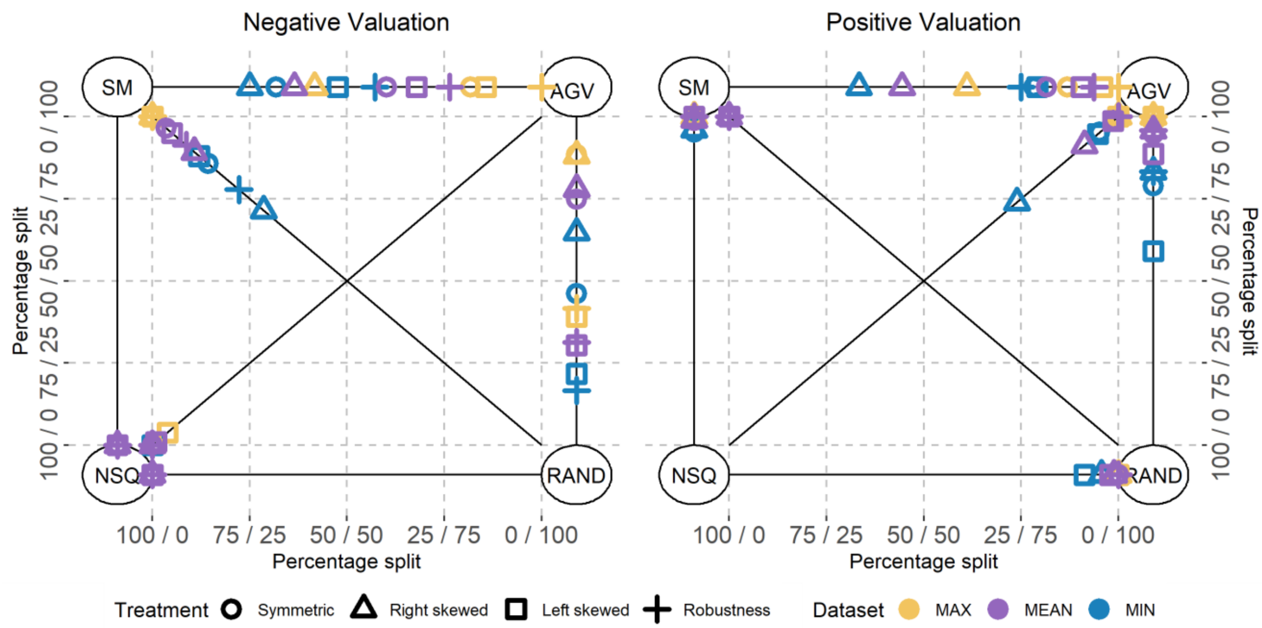


Figure H2. Binary mechanism choices in the ex-ante rounds

Note: Each of the six axes in the figures displays the fraction of subjects choosing the mechanisms indicated at the corners the axis. The max, mean, and minimal values across the different runs are presented. The spread between the min and max value is smaller compared to GPT-3.5, presumably due to the lower number of runs. (Figure and code adapted from Hoffmann & Renes, 2021)

As shown in figure H2, GPT-4o demonstrates a high level of rationality in its choices. In its explanations, GPT-4o almost always incorporates its private valuation to support its decisions. This results in a pronounced preference for NSQ when faced with a negative valuation. Preferences between SM and AGV vary across different runs, with a slight overall preference for AGV. Similarly, preferences between RAND and AGV are also varied, with AGV favoured in two treatments and RAND favoured in the other two.

For the ad-interim rounds with a positive valuation, NSQ is no longer the preferred mechanism. Instead, AGV emerges as the most favored option, although there is some variation across different treatments. When AGV is chosen, explanations often highlight that it "incentivizes truthful valuation reporting." Overall, the decisions made in the ad-interim rounds exhibit rationality comparable to, or even exceeding, that observed in the lab results.

Table H3 provides an overview of the efficient mechanism choice percentage and the rational choice percentage for each run. The efficient mechanism choice reflects the theoretically optimal choice for the group, based on the mechanism and treatment for that group. This may differ from the theoretical optimal choice for individuals. Consequently, the rational ad-interim choice metric assesses whether the theoretically rational choice is made based on the individual's private valuation. This metric is relevant only for the ad-interim rounds, where participants are aware of their private valuation before making a decision

Table H3. Overview of efficient group choices and rational choices for individuals per run

	Total		Ex-ante		Ad-interim (-)		Ad-interim (+)	
RUNS	Efficient mech choice (%)	Rational score (%)	Efficient mech choice (%)	Rational score (%)	Efficient mech choice (%)	Ration al score (%)	Efficient mech choice (%)	Rational score (%)
Lab results	67	70	71		44	85	75	87
RUN-0	68	73	70		43	86	87	90
Experimental Control 1: Contextual Framing								
RUN-1	55	66	54		32	79	80	83
Experimental Control 3: Trait or Role Allocation								
RUN-5	72	75	75		43	88	87	91
Experimental Control 5: Temperature								
RUN-13	67	73	70		41	86	85	88
Totals								
Min	55	66	54		32	79	80	83
Max	72	75	75		43	88	87	91
Mean	66	72	67		40	85	85	88

Compared to the GPT-3.5 runs, the GPT-4o runs demonstrate significantly improved rationality, with average results closely aligning with the lab data. RUN-5, in particular, exhibits the most rational choices, suggesting that GPT-4o approximates human rational decision-making effectively.

In conclusion, while the differences observed in Tables H2 and H3 indicate that GPT-4o still does not fully replicate 'human-like' results, its performance is notably closer to human behaviour compared to GPT-3.5. Notably, where GPT-3.5 struggled with rationality, GPT-4o shows a much higher degree of alignment with the lab results. Some runs even surpasses the lab results in terms of rationality.

H.1.2 Part 2: Applying the Mechanism

In the second part of the experiment, the selected mechanism determines whether the group wants to implement project A. The AGV mechanism allows participants to state their valuation, while the SM mechanism involves a Yes or No vote. Table H4 provides an overview of truth-telling in the AGV mechanism and the rational voting percentages in the SM mechanism.

Unlike GPT-3.5, which showed results closely aligning with the lab data, GPT-4o exhibits significantly more rational behavior. GPT-4o reports its valuation truthfully 94% of the time and rarely falsifies its valuation.

This high level of truthfulness aligns with the expectation, as GPT-4o frequently emphasizes that AGV incentivizes honest reporting. The results are so consistently accurate that they approach what might be termed hyper-accuracy distortion.

Table H4. Overview of applying the selected mechanism, truth telling and rational votes

	Truth telling	**Truth telling sign**	**Truth telling sign with + valuation**	**Truth telling sign with – valuation**	**Yes votes with + valuation**	**No votes with + valuation**
Lab results	68	93	98	87	98	94
RUN-0	91	100	100	100	100	100
Experimental Control 1: Contextual Framing						
RUN-1	98	99	99	98	100	100
Experimental Control 3: Trait or Role Allocation						
RUN-5	93	100	100	100	100	100
Experimental Control 5: Temperature						
RUN-13	92	99	100	99	98	100
Totals						
Min	91	99	99	98	98	100
Max	98	100	100	100	100	100
Mean	94	100	100	99	100	100

In the few instances where GPT-4o does lie about its valuation, the explanations often suggest that the transfer payment can offset the loss associated with a negative valuation. However, this reasoning is incorrect, as such offsetting is not feasible in the context of the experiment.

```
Stated Valuation: 3

Reasoning: Given that the AGV rule determines implementation based on the sum of the stated valuations, and considering my true valuation of -3 euro, my objective is to maximize my utility by ensuring that the project does not proceed. However, the table indicates making specific claims could affect the transfer payments. By stating a high positive valuation of 3, I might receive a transfer payment from the other group members to balance out utility, despite the project being implemented, thus potentially reducing my total loss from my negative private valuation. This strategy leverages the transfer payment mechanism to potentially offset my negative valuation.

Alternatively, this declaration leverages potential misalignment in valuations to derive transfers, aiming to assure a balanced aggregate outcome regardless of whether the project is implemented or not.
```

In conclusion, the high scores presented in table H4 indicate that GPT-4o exhibits more rational decision-making than the human participants in the experiment. GPT-4o almost never falsifies its valuation and consistently makes more rational voting decisions.

H.2 Experimental Controls

The following subsections will detail the results of all the runs related to the experimental controls and highlight the most noteworthy deviations from the average scores. However, the differences are considerably smaller compared to GPT-3.5, so these will be covered briefly.

H.2.1 Experimental Control 1: Context Control

In the runs conducted with GPT-3.5, the adapted instructions led to significant variations across the runs. In contrast, GPT-4o appears less sensitive to these adapted instructions, with only minor differences that are not statistically significant compared to RUN-0.

H.2.2 Experimental Control 3: Trait or Role allocation

RUN-5 with GPT-3.5 was the least similar to the lab results among all 14 runs. In contrast, RUN-5 with GPT-4o is the most similar to the lab results among the four runs conducted with GPT-4o. This suggests that GPT-4o is better at making decisions that align with human-like reasoning when prompted accordingly. However, it is important to note that RUN-5 with GPT-4o is not statistically different from RUN-0. Additionally, RUN-5 scores the highest on the rationality metric, reflecting GPT-4o’s tendency to assume a high level of rationality in human decision-making

H.2.3 Experimental Control 5: Temperature

Increasing the temperature had a positive effect with GPT-3.5, but with GPT-4o, the differences are only marginal. Nevertheless, increasing the temperature to 1.5 led to 67 cases of 'invalid answers.' With a total of 2,700 rows, this amounts to 3.7% of the answers being invalid.

Appendix I: Statistical Analysis GPT-3.5

This appendix provides the statistical tests of all runs of GPT-3.5, tested on mean differences with the lab results and with the results of RUN-0. All hypotheses related to the tests are explained in chapter 4.

Appendix I0: Default run – RUN0

Table I0.1: One sample t-test, comparing differences with 0 (the lab results have 0 differences to itself)

		Mean	St dev	Std. Error Mean	95% lower	95% higher	Sig
Absolute	RUN-0 Ex-ante	14.5	12.9	2.6	9.0	20.0	< 0.001
	RUN-0 Ad-interim (+)	14.9	14.5	3.0	8.8	21.1	< 0.001
	RUN-0 Ad-interim (-)	46.0	33.2	6.8	32.0	60.0	< 0.001
Squared	RUN-0 Ex-ante	369.4	603.1	123.1	114.8	624.1	0.006
	RUN-0 Ad-interim (+)	424.6	680.6	138.9	137.2	712.0	0.006
	RUN-0 Ad-interim (-)	3174.8	3524.4	719.4	1686.6	4663.1	< 0.001

Appendix I1: Experimental Control 1

I1.1 RUN-1

Table I1.1: One sample t-test, comparing differences with 0 (the lab results have 0 differences to itself)

		Mean	St dev	Std. Error Mean	95% lower	95% higher	P-value (two sided)
Absolute	RUN-1 Ex-ante	22.3	13.8	2.8	22.3	13.8	<.001
	RUN-1 Ad-interim (+)	31.5	19.0	3.9	31.5	19.0	<.001
	RUN-1 Ad-interim (-)	29.1	21.8	4.4	29.1	21.8	<.001
Squared	RUN-1 Ex-ante	681.8	726.3	148.2	681.8	726.3	<.001
	RUN-1 Ad-interim (+)	1340.7	1532.4	312.8	1340.7	1532.4	<.001
	RUN-1 Ad-interim (-)	1301.8	1634.9	333.7	1301.8	1634.9	<.001

Table I1.2: Paired sample t-test, comparing differences with RUN-0

		Mean	St dev	Std. Error Mean	P-value (one-sided)
Pair 1	R0_ex_ante - R1_ex_ante	-7.9	16.5	3.4	.014
Pair 2	R0_ad_int_pos - R1_ad_int_pos	-16.6	17.2	3.5	<.001
Pair 3	R0_ad_int_neg - R1_ad_int_neg	16.9	42.0	8.6	.031
Pair 4	R0_ex_ante - R1_ex_ante	-312.3	635.0	129.6	.012
Pair 5	R0_ad_int_pos - R1_ad_int_pos	-916.1	1163.5	237.5	<.001
Pair 6	R0_ad_int_neg - R1_ad_int_neg	1873.0	4055.1	827.7	.017

I1.2 RUN-2

Table I1.3: One sample t-test, comparing differences with 0 (the lab results have 0 differences to itself)

		Mean	St dev	Std. Error Mean	95% lower	95% higher	P-value (two sided)
Absolute	RUN-2 Ex-ante	18.3	16.3	3.3	18.3	16.3	<.001
	RUN-2 Ad-interim (+)	27.8	33.4	6.8	27.8	33.4	<.001
	RUN-2 Ad-interim (-)	39.1	32.3	6.6	39.1	32.3	<.001
Squared	RUN-2 Ex-ante	589.3	1022.3	208.7	589.3	1022.3	.010
	RUN-2 Ad-interim (+)	1840.4	3192.9	651.8	1840.4	3192.9	.010
	RUN-2 Ad-interim (-)	2527.2	2998.3	612.0	2527.2	2998.3	<.001

Table I1.4: Paired sample t-test, comparing differences with RUN-0

		Mean	St dev	Std. Error Mean	P-value (one-sided)
Pair 1	R0_ex_ante - R2_ex_ante	-3.8	18.4	3.8	.159
Pair 2	R0_ad_int_pos - R2_ad_int_pos	-12.9	38.6	7.9	.057
Pair 3	R0_ad_int_neg - R2_ad_int_neg	6.9	43.1	8.8	.221
Pair 4	R0_ex_ante - R2_ex_ante	-219.9	1060.5	216.5	.160
Pair 5	R0_ad_int_pos - R2_ad_int_pos	-1415.8	3369.1	687.7	.026
Pair 6	R0_ad_int_neg - R2_ad_int_neg	647.6	4281.2	873.9	.233

Table I1.5: Paired sample t-test, comparing differences with RUN-1

		Mean	St dev	Std. Error Mean	P-value (one-sided)
Pair 1	R1_ex_ante - R2_ex_ante	4.0	25.3	5.2	.221
Pair 2	R1_ad_int_pos - R2_ad_int_pos	3.7	33.7	6.9	.298
Pair 3	R1_ad_int_neg - R2_ad_int_neg	-10.0	33.1	6.8	.076
Pair 4	R1_ex_ante - R2_ex_ante	92.4	1446.5	295.3	.379
Pair 5	R1_ad_int_pos - R2_ad_int_pos	-499.7	3475.6	709.5	.244
Pair 6	R1_ad_int_neg - R2_ad_int_neg	-1225.4	3235.5	660.4	.038

Appendix I2: Experimental Control 2

I2.1 RUN-3

Table I2.1: One sample t-test, comparing differences with 0 (the lab results have 0 differences to itself)

		Mean	St dev	Std. Error Mean	95% lower	95% higher	P-value (two sided)
Absolute	RUN-3 Ex-ante	17.3	14.1	2.9	11.3	23.2	<.001
	RUN-3 Ad-interim (+)	17.2	17.3	3.5	9.9	24.5	<.001
	RUN-3 Ad-interim (-)	39.3	26.3	5.4	28.2	50.4	<.001
Squared	RUN-3 Ex-ante	489.3	742.2	151.5	175.9	802.7	.004
	RUN-3 Ad-interim (+)	584.6	1009.3	206.0	158.4	1010.7	.009
	RUN-3 Ad-interim (-)	2209.5	2251.9	459.7	1258.6	3160.4	<.001

Table I2.2: Paired sample t-test, comparing differences with RUN-0

		Mean	St dev	Std. Error Mean	P-value (one-sided)
Pair 1	R0_ex_ante - R3_ex_ante	-2.8	11.9	2.4	.131
Pair 2	R0_ad_int_pos - R3_ad_int_pos	-2.3	14.5	3.0	.223
Pair 3	R0_ad_int_neg - R3_ad_int_neg	6.7	25.2	5.1	.104
Pair 4	R0_ex_ante - R3_ex_ante	-119.9	450.1	91.9	.102
Pair 5	R0_ad_int_pos - R3_ad_int_pos	-159.9	711.7	145.3	.141
Pair 6	R0_ad_int_neg - R3_ad_int_neg	965.3	2860.8	584.0	.056

I2.2 RUN-4

Table I2.3: One sample t-test, comparing differences with 0 (the lab results have 0 differences to itself)

		Mean	St dev	Std. Error Mean	95% lower	95% higher	P-value (two sided)
Absolute	RUN-4 Ex-ante	21.4	13.0	2.7	15.9	26.9	<.001
	RUN-4 Ad-interim (+)	21.2	21.3	4.3	12.2	30.2	<.001
	RUN-4 Ad-interim (-)	51.9	35.0	7.2	37.1	66.7	<.001
Squared	RUN-4 Ex-ante	618.8	675.0	137.8	333.8	903.9	<.001
	RUN-4 Ad-interim (+)	882.6	1448.6	295.7	270.9	1494.3	.007
	RUN-4 Ad-interim (-)	3868.5	3593.4	733.5	2351.2	5385.9	<.001

Table I2.4: Paired sample t-test, comparing differences with RUN-0

		Mean	St dev	Std. Error Mean	P-value (one-sided)
Pair 1	R0_ex_ante - R4_ex_ante	-4.0	16.8	3.4	.125
Pair 2	R0_ad_int_pos - R4_ad_int_pos	-3.1	18.6	3.8	.208
Pair 3	R0_ad_int_neg - R4_ad_int_neg	23.4	39.7	8.1	.004
Pair 4	R0_ex_ante - R4_ex_ante	-76.0	702.9	143.5	.301
Pair 5	R0_ad_int_pos - R4_ad_int_pos	-169.6	904.9	184.7	.184
Pair 6	R0_ad_int_neg - R4_ad_int_neg	2439.5	3766.0	768.7	.002

Table I2.5: Paired sample t-test, comparing differences with RUN-3

		Mean	St dev	Std. Error Mean	P-value (one-sided)
Pair 1	R3_ex_ante - R4_ex_ante	-1.3	15.1	3.1	.344
Pair 2	R3_ad_int_pos - R4_ad_int_pos	-.8	15.9	3.2	.399
Pair 3	R3_ad_int_neg - R4_ad_int_neg	16.7	30.8	6.3	.007
Pair 4	R3_ex_ante - R4_ex_ante	43.9	749.4	153.0	.388
Pair 5	R3_ad_int_pos - R4_ad_int_pos	-9.7	810.9	165.5	.477
Pair 6	R3_ad_int_neg - R4_ad_int_neg	1474.2	2364.9	482.7	.003

Appendix I3: Experimental Control 3

I3.1 RUN-5

Table I3.1: One sample t-test, comparing differences with 0 (the lab results have 0 differences to itself)

		Mean	St dev	Std. Error Mean	95% lower	95% higher	P-value (two sided)
Absolute	RUN-5 Ex-ante	18.5	10.4	2.1	14.1	22.9	<.001
	RUN-5 Ad-interim (+)	18.1	16.7	3.4	11.0	25.1	<.001
	RUN-5 Ad-interim (-)	22.6	15.3	3.1	16.1	29.1	<.001
Squared	RUN-5 Ex-ante	445.4	362.4	74.0	292.4	598.4	<.001
	RUN-5 Ad-interim (+)	594.2	829.5	169.3	243.9	944.5	.002
	RUN-5 Ad-interim (-)	735.3	820.6	167.5	388.8	1081.8	<.001

Table E2.2: Paired sample t-test, comparing differences with RUN-0

		Mean	St dev	Std. Error Mean	P-value (one-sided)
Pair 1	R0_ex_ante - R5_ex_ante	-6.9	14.9	3.0	.017
Pair 2	R0_ad_int_pos - R5_ad_int_pos	-6.3	23.3	4.8	.100
Pair 3	R0_ad_int_neg - R5_ad_int_neg	-5.9	20.7	4.2	.089
Pair 4	R0_ex_ante - R5_ex_ante	-249.4	608.3	124.2	.028
Pair 5	R0_ad_int_pos - R5_ad_int_pos	-458.0	1443.8	294.7	.067
Pair 6	R0_ad_int_neg - R5_ad_int_neg	-693.7	1622.0	331.1	.024

I3.2 RUN-6

Table I3.3: One sample t-test, comparing differences with 0 (the lab results have 0 differences to itself)

		Mean	**St dev**	**Std. Error Mean**	**95% lower**	**95% higher**	**P-value (two sided)**
Absolute	RUN-6 Ex-ante	18.2	11.1	2.3	13.5	22.9	<.001
	RUN-6 Ad-interim (+)	24.3	22.1	4.5	15.0	33.7	<.001
	RUN-6 Ad-interim (-)	47.2	36.4	7.4	31.8	62.5	<.001
Squared	RUN-6 Ex-ante	449.4	409.3	83.5	276.6	622.2	<.001
	RUN-6 Ad-interim (+)	1062.2	1602.0	327.0	385.7	1738.6	.004
	RUN-6 Ad-interim (-)	3495.1	3634.8	741.9	1960.3	5029.9	<.001

Table E2.4: Paired sample t-test, comparing differences with RUN-0

		Mean	**St dev**	**Std. Error Mean**	**P-value (one-sided)**
Pair 1	R0_ex_ante - R6_ex_ante	-3.7	17.4	3.5	.153
Pair 2	R0_ad_int_pos - R6_ad_int_pos	-9.4	21.5	4.4	.021
Pair 3	R0_ad_int_neg - R6_ad_int_neg	-1.2	17.7	3.6	.375
Pair 4	R0_ex_ante - R6_ex_ante	-80.0	754.1	153.9	.304
Pair 5	R0_ad_int_pos - R6_ad_int_pos	-637.6	1383.0	282.3	.017
Pair 6	R0_ad_int_neg - R6_ad_int_neg	-320.3	1218.6	248.7	.105

I3.3 RUN-7

Table I3.5: One sample t-test, comparing differences with 0 (the lab results have 0 differences to itself)

		Mean	St dev	Std. Error Mean	95% lower	95% higher	P-value (two sided)
Absolute	RUN-7 Ex-ante	11.7	8.7	1.8	8.0	15.3	<.001
	RUN-7 Ad-interim (+)	18.2	18.9	3.9	10.2	26.2	<.001
	RUN-7 Ad-interim (-)	48.8	32.1	6.6	35.2	62.4	<.001
Squared	RUN-7 Ex-ante	207.8	268.3	54.8	94.5	321.1	<.001
	RUN-7 Ad-interim (+)	671.9	1084.1	221.3	214.1	1129.7	.006
	RUN-7 Ad-interim (-)	3369.3	3197.3	652.6	2019.2	4719.4	<.001

Table I3.6: Paired sample t-test, comparing differences with RUN-0

		Mean	St dev	Std. Error Mean	P-value (one-sided)
Pair 1	R0_ex_ante - R7_ex_ante	2.8	14.1	2.9	.168
Pair 2	R0_ad_int_pos - R7_ad_int_pos	-3.3	21.8	4.5	.236
Pair 3	R0_ad_int_neg - R7_ad_int_neg	-2.8	21.5	4.4	.264
Pair 4	R0_ex_ante - R7_ex_ante	161.6	593.2	121.1	.098
Pair 5	R0_ad_int_pos - R7_ad_int_pos	-247.3	1238.5	252.8	.169
Pair 6	R0_ad_int_neg - R7_ad_int_neg	-194.5	2491.6	508.6	.353

Table I3.7: ANOVA between RUN-5, RUN-6, RUN-7

		SIG
ABS	Ex-ante	.011
	Ad-interim (+)	.597
	Ad-interim (-)	.889
Squared	Ex-ante	.016
	Ad-interim (+)	.627
	Ad-interim (-)	.875

Appendix I4: Experimental Control 4

I4.1 RUN-8

Table I4.1: One sample t-test, comparing differences with 0 (the lab results have 0 differences to itself)

		Mean	St dev	Std. Error Mean	95% lower	95% higher	P-value (two sided)
Absolute	RUN-8 Ex-ante	20.1	13.2	2.7	14.5	25.7	<.001
	RUN-8 Ad-interim (+)	24.6	21.7	4.4	15.5	33.8	<.001
	RUN-8 Ad-interim (-)	46.6	26.8	5.5	35.3	57.9	<.001
Squared	RUN-8 Ex-ante	570.9	589.8	120.4	321.8	819.9	<.001
	RUN-8 Ad-interim (+)	1059.1	1531.9	312.7	412.2	1705.9	.003
	RUN-8 Ad-interim (-)	2862.7	2580.4	526.7	1773.1	3952.4	<.001

Table I4.1: Paired sample t-test, comparing differences with RUN-0

		Mean	St dev	Std. Error Mean	P-value (one-sided)
Pair 1	R0_ex_ante - R8_ex_ante	-5.6	17.2	3.5	.061
Pair 2	R0_ad_int_pos - R8_ad_int_pos	-9.7	21.0	4.3	.016
Pair 3	R0_ad_int_neg - R8_ad_int_neg	-.6	24.6	5.0	.449
Pair 4	R0_ex_ante - R8_ex_ante	-201.4	659.0	134.5	.074
Pair 5	R0_ad_int_pos - R8_ad_int_pos	-634.4	1278.3	260.9	.012
Pair 6	R0_ad_int_neg - R8_ad_int_neg	312.1	2214.5	452.0	.248

I4.2 RUN-9

Table I4.3: One sample t-test, comparing differences with 0 (the lab results have 0 differences to itself)

		Mean	St dev	Std. Error Mean	95% lower	95% higher	P-value (two sided)
Absolute	RUN-9 Ex-ante	24.5	16.0	3.3	17.7	31.3	<.001
	RUN-9 Ad-interim (+)	37.5	22.5	4.6	28.0	47.0	<.001
	RUN-9 Ad-interim (-)	27.2	19.1	3.9	19.1	35.2	<.001
Squared	RUN-9 Ex-ante	845.6	810.3	165.4	503.4	1187.8	<.001
	RUN-9 Ad-interim (+)	1891.6	2016.9	411.7	1040.0	2743.3	<.001
	RUN-9 Ad-interim (-)	1087.8	1375.3	280.7	507.0	1668.5	<.001

Table I4.3: Paired sample t-test, comparing differences with RUN-0

		Mean	St dev	Std. Error Mean	P-value (one-sided)
Pair 1	R0_ex_ante - R9_ex_ante	-10.0	24.9	5.1	.030
Pair 2	R0_ad_int_pos - R9_ad_int_pos	-22.6	17.4	3.6	<.001
Pair 3	R0_ad_int_neg - R9_ad_int_neg	18.8	37.4	7.6	.011
Pair 4	R0_ex_ante - R9_ex_ante	-476.2	1201.8	245.3	.032
Pair 5	R0_ad_int_pos - R9_ad_int_pos	-1467.0	1609.7	328.6	<.001
Pair 6	R0_ad_int_neg - R9_ad_int_neg	2087.1	3717.4	758.8	.006

I4.3 RUN-10

Table I4.5: One sample t-test, comparing differences with 0 (the lab results have 0 differences to itself)

		Mean	St dev	Std. Error Mean	95% lower	95% higher	P-value (two sided)
Absolute	RUN-10 Ex-ante	16.2	12.3	2.5	11.0	21.4	<.001
	RUN-10 Ad-interim (+)	25.2	20.1	4.1	16.7	33.7	<.001
	RUN-10 Ad-interim (-)	48.7	34.2	7.0	34.2	63.1	<.001
Squared	RUN-10 Ex-ante	406.7	408.1	83.3	234.4	579.1	<.001
	RUN-10 Ad-interim (+)	1020.2	1514.1	309.1	380.9	1659.6	.003
	RUN-10 Ad-interim (-)	3490.6	3396.4	693.3	2056.4	4924.8	<.001

Table I4.6: Paired sample t-test, comparing differences with RUN-0

		Mean	St dev	Std. Error Mean	P-value (one-sided)
Pair 1	R0_ex_ante - R10_ex_ante	-1.7	16.8	3.4	.313
Pair 2	R0_ad_int_pos - R10_ad_int_pos	-10.2	17.0	3.5	.004
Pair 3	R0_ad_int_neg - R10_ad_int_neg	-2.7	19.1	3.9	.250
Pair 4	R0_ex_ante - R10_ex_ante	-37.3	661.0	134.9	.392
Pair 5	R0_ad_int_pos - R10_ad_int_pos	-595.6	1175.6	240.0	.010
Pair 6	R0_ad_int_neg - R10_ad_int_neg	-315.8	1520.3	310.3	.160

Table I4.7: ANOVA between RUN-8, RUN-9, RUN-10

		SIG
ABS	Ex-ante	.126
	Ad-interim (+)	.069
	Ad-interim (-)	.015
Squared	Ex-ante	.055
	Ad-interim (+)	.143
	Ad-interim (-)	.006

Appendix I5: Experimental Control 5

I5.1 RUN-11

Table I5.1: One sample t-test, comparing differences with 0 (the lab results have 0 differences to itself)

		Mean	St dev	Std. Error Mean	95% lower	95% higher	P-value (two sided)
Absolute	RUN-1 Ex-ante	20.6	16.1	3.3	13.8	27.3	<.001
	RUN-11 Ad-interim (+)	16.5	17.2	3.5	9.2	23.7	<.001
	RUN-11 Ad-interim (-)	54.7	30.7	6.3	41.7	67.7	<.001
Squared	RUN-11 Ex-ante	671.0	995.9	203.3	250.4	1091.5	.003
	RUN-11 Ad-interim (+)	554.1	985.4	201.2	138.0	970.2	.011
	RUN-11 Ad-interim (-)	3896.8	3363.7	686.6	2476.5	5317.2	<.001

Table I5.2: Paired sample t-test, comparing differences with RUN-0

		Mean	St dev	Std. Error Mean	P-value (one-sided)
Pair 1	R0_ex_ante - R11_ex_ante	-6.08	12.15	2.48	.011
Pair 2	R0_ad_int_pos - R11_ad_int_pos	-1.54	14.80	3.02	.308
Pair 3	R0_ad_int_neg - R11_ad_int_neg	-8.71	17.60	3.59	.012
Pair 4	R0_ex_ante - R11_ex_ante	-301.56	600.47	122.57	.011
Pair 5	R0_ad_int_pos - R11_ad_int_pos	-129.50	872.36	178.07	.237
Pair 6	R0_ad_int_neg - R11_ad_int_neg	-722.01	1227.37	250.54	.004

I5.2 RUN-12

Table I5.3: One sample t-test, comparing differences with 0 (the lab results have 0 differences to itself)

		Mean	St dev	Std. Error Mean	95% lower	95% higher	P-value (two sided)
Absolute	RUN-12 Ex-ante	17.4	16.1	3.3	10.6	24.2	<.001
	RUN-12 Ad-interim (+)	12.3	15.5	3.2	5.8	18.9	<.001
	RUN-12 Ad-interim (-)	52.8	31.1	6.3	39.7	66.0	<.001
Squared	RUN-12 Ex-ante	552.8	909.8	185.7	168.6	937.0	.007
	RUN-12 Ad-interim (+)	382.5	844.0	172.3	26.1	738.9	.037
	RUN-12 Ad-interim (-)	3718.1	3503.2	715.1	2238.9	5197.4	<.001

Table I5.4: Paired sample t-test, comparing differences with RUN-0

		Mean	St dev	Std. Error Mean	P-value (one-sided)
Pair 1	R0_ex_ante - R12_ex_ante	-2.9	11.3	2.3	.107
Pair 2	R0_ad_int_pos - R12_ad_int_pos	2.6	14.4	2.9	.195
Pair 3	R0_ad_int_neg - R12_ad_int_neg	-6.9	13.1	2.7	.009
Pair 4	R0_ex_ante - R12_ex_ante	-183.4	463.3	94.6	.032
Pair 5	R0_ad_int_pos - R12_ad_int_pos	42.1	767.4	156.7	.395
Pair 6	R0_ad_int_neg - R12_ad_int_neg	-543.3	1176.5	240.2	.017

I5.3 RUN-13

Table I5.5: One sample t-test, comparing differences with 0 (the lab results have 0 differences to itself)

	Mean	St dev	Std. Error Mean	95% lower	95% higher	P-value (two sided)
RUN-13 Ex-ante	12.8	10.5	2.1	8.4	17.3	<.001
RUN-13 Ad-interim (+)	13.3	11.3	2.3	8.5	18.1	<.001
RUN-13 Ad-interim (-)	44.3	30.2	6.2	31.6	57.1	<.001
RUN-13 Ex-ante	271.1	508.7	103.8	56.3	485.9	.016
RUN-13 Ad-interim (+)	299.5	405.3	82.7	128.4	470.7	.001
RUN-13 Ad-interim (-)	2839.9	3023.9	617.3	1563.1	4116.8	<.001

Table I5.6: Paired sample t-test, comparing differences with RUN-0

		Mean	St dev	Std. Error Mean	P-value (one-sided)
Pair 1	R0_ex_ante - R13_ex_ante	1.6	7.4	1.5	.143
Pair 2	R0_ad_int_pos - R13_ad_int_pos	1.6	9.2	1.9	.197
Pair 3	R0_ad_int_neg - R13_ad_int_neg	1.6	11.1	2.3	.239
Pair 4	R0_ex_ante - R13_ex_ante	98.3	267.0	54.5	.042
Pair 5	R0_ad_int_pos - R13_ad_int_pos	125.1	395.5	80.7	.067
Pair 6	R0_ad_int_neg - R13_ad_int_neg	334.9	1244.7	254.1	.100

Table I5.7: ANOVA between RUN-11, RUN-12, RUN-13

		SIG
ABS	Ex-ante	.215
	Ad-interim (+)	.616
	Ad-interim (-)	.479
Squared	Ex-ante	.269
	Ad-interim (+)	.548
	Ad-interim (-)	.538

Appendix J: Statistical Analysis GPT-4o

This appendix provides the statistical tests of all runs of GPT-4o, tested on mean differences with the lab results and with the results of RUN-0. All hypotheses related to the tests are explained in chapter 4.

Appendix J0: Default run – RUN0

Table J0.1: One sample t-test, comparing differences with 0 (the lab results have 0 differences to itself)

		Mean	St dev	Std. Error Mean	95% lower	95% higher	P-value (two-sided)
Absolute	RUN-0 Ex-ante	20.8	15.9	3.2	14.1	27.5	<.001
	RUN-0 Ad-interim (+)	15.1	16.8	3.4	8.0	22.2	<.001
	RUN-0 Ad-interim (-)	14.2	17.7	3.6	6.7	21.6	<.001
Squared	RUN-0 Ex-ante	673.7	905.2	184.8	291.5	1055.9	.001
	RUN-0 Ad-interim (+)	496.3	933.4	190.5	102.2	890.5	.016
	RUN-0 Ad-interim (-)	499.9	1097.6	224.0	36.5	963.4	.036

Appendix J1: Experimental Control 1

J1 RUN-1

Table J1.1: One sample t-test, comparing differences with 0 (the lab results have 0 differences to itself)

		Mean	St dev	Std. Error Mean	95% lower	95% higher	P-value (two sided)
Absolute	RUN-1 Ex-ante	19.3	14.5	3.0	13.1	25.4	<.001
	RUN-1 Ad-interim (+)	13.1	17.4	3.5	5.7	20.4	<.001
	RUN-1 Ad-interim (-)	15.0	16.4	3.4	8.1	21.9	<.001
Squared	RUN-1 Ex-ante	573.8	639.9	130.6	303.6	844.0	<.001
	RUN-1 Ad-interim (+)	459.8	940.9	192.1	62.5	857.1	.013
	RUN-1 Ad-interim (-)	483.4	858.7	175.3	120.8	846.0	.006

Table J1.2: Paired sample t-test, comparing differences with RUN-0

		Mean	St dev	Std. Error Mean	P-value (one-sided)
Pair 1	R0_ex_ante - R1_ex_ante	1.5	23.1	4.7	.376
Pair 2	R0_ad_int_pos - R1_ad_int_pos	2.0	10.2	2.1	.173
Pair 3	R0_ad_int_neg - R1_ad_int_neg	-.8	16.6	3.4	.406
Pair 4	R0_ex_ante - R1_ex_ante	99.9	1161.7	237.1	.339
Pair 5	R0_ad_int_pos - R1_ad_int_pos	36.5	388.0	79.2	.325
Pair 6	R0_ad_int_neg - R1_ad_int_neg	16.5	943.2	192.5	.466

Appendix J2: Experimental Control 3

J2 RUN-5

Table J2.1: One sample t-test, comparing differences with 0 (the lab results have 0 differences to itself)

		Mean	St dev	Std. Error Mean	95% lower	95% higher	P-value (two sided)
Absolute	RUN-5 Ex-ante	17.4	15.3	3.1	10.9	23.9	<.001
	RUN-5 Ad-interim (+)	13.5	14.4	2.9	7.4	19.6	<.001
	RUN-5 Ad-interim (-)	11.7	12.1	2.5	6.6	16.8	<.001
Squared	RUN-5 Ex-ante	527.6	837.4	170.9	174.0	881.2	.005
	RUN-5 Ad-interim (+)	380.5	717.7	146.5	77.5	683.5	.016
	RUN-5 Ad-interim (-)	275.8	481.5	98.3	72.5	479.1	.010

Table J1.4: Paired sample t-test, comparing differences with RUN-0

		Mean	St dev	Std. Error Mean	P-value (one-sided)
Pair 1	R0_ex_ante - R5_ex_ante	3.4	10.9	2.2	.072
Pair 2	R0_ad_int_pos - R5_ad_int_pos	1.6	4.7	1.0	.059
Pair 3	R0_ad_int_neg - R5_ad_int_neg	2.5	9.0	1.8	.091
Pair 4	R0_ex_ante - R5_ex_ante	146.1	536.9	109.6	.098
Pair 5	R0_ad_int_pos - R5_ad_int_pos	115.8	390.2	79.7	.080
Pair 6	R0_ad_int_neg - R5_ad_int_neg	224.1	706.0	144.1	.067

Appendix J3: Experimental Control 5

FJ RUN-13

Table J3.1: One sample t-test, comparing differences with 0 (the lab results have 0 differences to itself)

		Mean	St dev	Std. Error Mean	95% lower	95% higher	P-value (two sided)
Absolute	RUN-13 Ex-ante	21.0	18.4	3.8	13.2	28.8	<.001
	RUN-13 Ad-interim (+)	15.1	17.1	3.5	7.9	22.3	<.001
	RUN-13 Ad-interim (-)	11.0	12.4	2.5	5.8	16.2	<.001
Squared	RUN-13 Ex-ante	766.2	1081.6	220.8	309.4	1222.9	.002
	RUN-13 Ad-interim (+)	506.3	948.6	193.6	105.7	906.8	.015
	RUN-13 Ad-interim (-)	267.4	523.0	106.8	46.6	488.3	.020

Table J3.2: Paired sample t-test, comparing differences with RUN-0

		Mean	St dev	Std. Error Mean	P-value (one-sided)
Pair 1	R0_ex_ante - R13_ex_ante	-.2	5.1	1.0	.417
Pair 2	R0_ad_int_pos - R13_ad_int_pos	.0	3.0	.6	.497
Pair 3	R0_ad_int_neg - R13_ad_int_neg	3.2	9.8	2.0	.061
Pair 4	R0_ex_ante - R13_ex_ante	-92.5	329.0	67.2	.091
Pair 5	R0_ad_int_pos - R13_ad_int_pos	-10.0	134.9	27.5	.361
Pair 6	R0_ad_int_neg - R13_ad_int_neg	232.5	637.0	130.0	.043

Appendix K: AI Tools Statement

During the preparation of this work, I used ChatGPT in order to run all my simulations. Furthermore, to check for spelling and grammar mistakes, and also to adjust some writing as well. I did not use it to generate new text. After using this tool/service, I reviewed and edited the content as needed and take full responsibility for the content of my thesis.