UNIVERSITY OF WAIKATO

Hamilton
New Zealand

How does AI affect price discovery and liquidity in asset market experiments?

Carol Luengo, Steven Tucker, Yilong Xu and Frank Scrimgeour

Working Paper in Economics 5/26

September 2026

Corresponding Author

Steven Tucker

Waikato Management School
University of Waikato
Private Bag 3105
Hamilton, 3240
New Zealand

Email: steven.tucker@waikato.ac.nz

Carol Luengo and Frank Scrimgeour

Waikato Management School

University of Waikato

Yilong Xu

Utrecht School of Economics

Utrecht University

Abstract

This paper investigates the causal impact of artificial intelligence (AI) advice on price discovery in a controlled asset-market experiment. Using a call market setting in which the asset's value follows a stochastic geometric random walk, we compare a baseline "No AI" condition against two treatments: "Good AI" (advice aligned with long-term fundamentals) and "Bad AI" (myopic advice anchored to current buyout prices). Our results show that access to AI advice significantly reduces mispricing relative to the baseline. In particular, Good AI market prices started low, then converged and closely followed the fundamentals in the second half of the market. Interestingly, we find no meaningful difference in aggregate mispricing between the Good and Bad AI treatments. Analysis of trader behavior reveals that while over half of the participants follow AI recommendations, a significant portion actively filters out erroneous advice, particularly in the Bad AI treatment. This selective adherence explains why market prices remain resilient to poor advice.

Keywords

Artificial Intelligence

Financial Advice

Asset Market Experiment

Mispricing

Human-AI Interaction

JEL codes

C90

C91

G12

Acknowledgement

We gratefully acknowledge the financial support of the University of Waikato. Carol Luengo acknowledges the support from the scholarship funded by the Agencia Nacional de Investigación y Desarrollo (ANID) / Beca de Doctorado en el Extranjero Becas Chile / 2022 - 72220236.

1. Introduction

Artificial intelligence is rapidly becoming a practical assistant in trading decisions. Retail investors often consult large language models (LLMs) for investment advice (Cheng et al., 2025; eToro, 2025; Lopez-Lira and Tang, 2023; Nicholson-Messmer, 2025), such as price prediction, portfolio views, and interpretations of financial news. When traders can easily obtain AI-generated recommendations that appear to be confident and professional, how does that affect price discovery in markets? Do traders follow the advice of AI? Or do AIs inspire more diverse opinions about asset pricing? Are traders able to distinguish good and bad advice generated by AIs?

To shed some light on these questions, this paper examines the causal impact of AI advice on price discovery and liquidity in controlled asset-market experiments. We focus on a setting where the asset's fundamental value is well-defined but evolves stochastically over time. This environment allows us to test not only whether AI reduces mispricing on average, but also whether the content and alignment of the advice matter for market outcomes. A central behavioral concern is that LLM outputs are typically delivered with high apparent confidence (Zhou et al., 2024) and are highly persuasive (Bai et al., 2025). As a result, participants may fail to discern the actual quality obscured by the model's confident presentation (Hazra and Serra-Garcia, 2025). If so, AI could improve market performance when advice aligns with fundamentals, but impairs it when it is misaligned.

This potential for misalignment is rooted in broader concerns regarding the reliability of generative AI output (Gillespie et al., 2023). First, LLMs can respond incorrectly or "hallucinate" information, expressing confidence in a way that seems correct, without having a basis in accurate facts or data (Zhou et al., 2024), which limits the adoption of LLMs in knowledge-intensive areas such as finance (Wu et al., 2023). Second, effective interaction with generative AI tools depends on the instructions provided by the user, called prompts**,** which define rules, guide processes, and specify the qualities that the generated outputs should possess. Consequently, the quality of the prompts is directly related to the quality of the generated outputs (White et al., 2023). Retail investors may not have the know-how to prompt the AI in a precise manner. This, coupled with the fact that AI can sometimes produce content that appears to be "nearly right", makes it very hard for individuals to judge whether they actually fully understand the AI output or just have a perception of understanding without a solid basis. This potential misperception may foster overconfidence that could prove costly in financial markets.

To our knowledge, ours is the first study to examine whether AI advisors influence market outcomes in a controlled setting. We compare a baseline "No AI" (NAI) setting against two treatment conditions: Good AI and Bad AI. While both treatments allow participants to interact freely with an LLM advisor, they differ in the quality of the underlying model. The trading institution for all treatments was a call market. In the AI conditions, participants first had the option to interact with the advisor before deciding whether to place bids or asks. Because the asset's buyout value follows a geometric random walk, a significant challenge is to distinguish the immediate end-of-period buyout value from the asset's true fundamental value (the expected terminal payoff if held until the terminal period, *T*). This distinction provides a natural laboratory to vary advice quality: the "Good AI" consistently generates recommendations based on long-term fundamental values, whereas the "Bad AI" tends toward myopic advice anchored to the current period's buyout price. This design allows us to isolate not only whether participants adhere to AI recommendations, but whether that adherence is contingent on the advice's objective accuracy.

Our findings show that access to AI advice significantly reduced mispricing relative to the Baseline treatment. However, we find no statistically significant differences in mispricing between the two AI treatments. Additionally, regarding the direction of mispricing, we observed a tendency toward overpricing in the absence of AI and slightly underpricing when AI was present. It is worth noting that the prices track the fundamentals very well after the 7^{th} period (15 trading periods in total) in Good AI. Finally, the analysis of traders' conversations with AI reveals that although a bit over half of the traders exhibited a general tendency to follow AI recommendations, a significant portion of traders also actively filtered out substantial amounts of incorrect advice and chose not to comply. As a result, those who did not follow the advice in Bad (Good) AI treatment tended to trade at higher (lower) prices than the advised price. This, at least in part, explains why market prices under AI conditions are not significantly different from one another.

Our studies make four contributions to the literature. First, the results speak to the effect of communication on asset bubbles. Hargreaves Heap and Zizzo (2011) found that anonymous digital communication appears to have little to no effect on the magnitude of bubbles. Oechssler et al. (2011) note that computer-based messaging might discourage bubble formation. In contrast, Steiger and Pelster (2020) demonstrate that face-to-face interactions exacerbate bubble formation when compared to traditional laboratory settings. Yet these are communications among the traders themselves. Our setting resembles a situation where investors receive financial advice. To this end, Alevy and Price (2017) find that bubbles are significantly reduced when traders are advised by experienced traders who participated in the experiment before. Our results corroborate this finding in a setting in which financial advice is provided by generative AI.

This study also contributes to the emerging literature on augmenting human decision-making with algorithms. This literature has shown that human decision-makers do not interact with algorithms as passive recipients (Caro et al., 2026). Rather, individuals actively adjust their reliance on algorithms, constantly weighing the cognitive costs of verification against the perceived accuracy of AI (Vasconcelos et al., 2023). In practice, investors use AI to filter out bad projects in Initial Coin Offerings (Basu et al., 2021) but rely on personal expertise for good ones. Bianchi and Brière (2026) study how investors interact with a robo-advisor deployed by a large asset management company. The investor has the option to follow the recommendation while retaining full control over the portfolio. The authors find that investors do follow the advice of the robot and become more attentive to deviations from their target allocations, and earn higher returns. In the domain of financial loan evaluations, Wang et al. (2026) show that human-algorithm collaboration can achieve the best results (relative to human- or algorithm-only approaches). This works especially well when humans can call out inconsistencies made by the algorithm. Yang et al. (2026) conducted a field experiment in a real-world financial advisory setting. They find that trust and uptake are lower when financial advice comes solely from AI than when humans are involved. We, too, find that a significant portion of investors do not blindly follow the advice of AI in asset markets, and they filter out bad information actively in the Bad AI treatment.

Third, our study also relates to the literature on algorithm aversion (Davis et al., 2024; Dietvorst et al., 2015). The foundational premise is that individuals tend to discount computer-based advice more heavily than identical human advice, particularly after observing the algorithm make a mistake (Castelo et al., 2019; Dietvorst et al., 2015). In financial contexts, where predictive perfection is impossible, algorithms will inevitably produce errors. When users observe these errors, they may rapidly lose confidence in the algorithm, despite the mathematical reality that the algorithm will still outperform the human over a long-term probabilistic horizon (Dietvorst et al., 2015). Yet, a major difference between generative AI and an algorithm is that the AI can explain its decision (even its mistakes). It is thus interesting to examine whether such an effect carries over in the AI era and in the domain of investment decisions. Logg et al. (2019)'s experiments demonstrate that such aversion is dependent on individuals' own expertise. Algorithm aversion only shows up when they find themselves laypeople frequently adhere more strictly to advice when they believe it originates from an algorithm rather than a human. The results wane when they have expertise on the topic.

Lastly, this paper makes a methodological contribution by proposing an experimental market framework that integrates AI tools into the decision-making environments. This study establishes clear guidelines for setting up the AI-assisted market experiment and designing the interaction between participants and AI systems. The framework is flexible and replicable, and can be adapted in future research to analyze the role of AI advice across different market contexts. Furthermore, since the seminal work of Smith et al. (1988), experimental markets have frequently relied on declining fundamental values to reliably generate price bubbles. Subsequent literature, however, demonstrates that altering this process to a random walk or a constant fundamental value can significantly reduce or eliminate bubble formation (e.g., Huber et al., 2008; Kirchler and Huber, 2007; Kirchler et al., 2012; Lambrecht et al., 2025; Noussair et al., 2022; Stöckl et al., 2015). We contribute to this literature by demonstrating that a call market that uses a random walk with drift can also reliably generate bubbles. In our baseline condition, prices initially start below the fundamental value but eventually rise significantly above it, remaining at elevated levels. This study thus provides a robust alternative framework for investigating bubbles and crashes within controlled laboratory environments.

The rest of the paper is organized as follows. Section 2 lays out the experimental design and procedure. Section 3 lays out the hypotheses. We report the results in Section 4. Discussion and conclusion are presented in Section 5.

2. The Experiment

The experiment consisted of 30 markets conducted at the Waikato Experimental Economics Laboratory (WEEL), located at the University of Waikato, New Zealand, between May 2025 and June 2025. Each market consisted of eight traders, except for one market with seven, resulting in 239 participants in the experiment. Participants were recruited university wide using the ORSEE recruitment system (Greiner, 2015). Some of the participants may have had previous experience in economic experiments, but all participants were inexperienced in asset markets and participated in a single session of this study. The markets were computerized and programmed with the oTree software package (Chen et al., 2016). Trade was facilitated with the experimental currency francs, and final cash holdings were converted into New Zealand dollars at a publicly known exchange rate. Each session lasted between 90 and 120 minutes, depending upon the treatment.[1] Parameters were set for all treatments to achieve expected earnings of 25 NZD per hour with realized average earnings of 49 NZD.

At the beginning of each market, traders were endowed with 10 units of an asset called X and 41,000 francs, and thus the initial cash-to-asset ratio across all markets was fixed at approximately ten.[2] The assets had a finite life of 15 periods during which traders had the opportunity to buy and/or sell assets across this time horizon according to the following constraints. Traders must have sufficient stock (cash) to sell (buy) a unit of the asset, and they cannot submit a purchase order for more units than are available in the market.[3] There were no trading fees nor interest paid on cash holdings.

At the end of the market, the experimenter bought back the assets held in the inventories of each trader. The price paid, and thus fundamental value, for each asset was called the End of Market Buyout Value, which equaled 100 francs plus the sum of fifteen random draws that occurred at the end of each period (Kirchler, 2009; Noussair et al., 2022).[4] The asset's fundamental value (FV_t) in any given period, *t* was calculated as

$$FV_t = BV_{t-1} * (1 + \varepsilon_t)^{(T-(t-1))}$$

where ε was randomly drawn at the end of each period and followed a normal distribution with an average of ε_t =9.80% and a standard deviation of 19.55%.[5] The fundamentals, or what we called the expected End-of-Market Buyout Value for the given period (at the end of period 15), was calculated and displayed on each trader's bidding screen.

[1] Baseline sessions lasted approximately 90 minutes, while AI-assisted market sessions lasted approximately 120 minutes. Reported durations include the instructional period and payment to participants.
[2] We chose a cash-to-asset ratio of ten to create a relatively bubble-prone environment, consistent with prior laboratory evidence that higher initial liquidity raises prices and can intensify bubble dynamics in experimental asset markets (Caginalp, Porter, and Smith, 1998; Haruvy and Noussair, 2006; Noussair and Tucker, 2016; Angerer and Szymczak, 2019).
[3] Available units = total number of participants multiplied by ten units of X minus the units held by the participant.
[4] The random draws were calculated in advance, and the same draws were used in all markets. Further details are provided in Appendix A1.
[5] The mean and variance values of ε_t, were calculated from the S&P 500 index data between the years 1928 and 2023.

The End-of-Period Buyout Value followed a random walk and was also presented on each trader's bidding screen. For any given period t, BV_t follows a process that is described as

$$BV_t = BV_{t-1} * (1 + \varepsilon_t).$$

The closed-book call market trading institution was used across all markets (Smith et al., 2000; Friedman, 1993; Cason and Friedman, 1997). In each period, traders simultaneously decided whether to submit an offer to buy or an offer to sell or not to trade. An offer to buy consists of a maximum quantity they are willing to purchase and a maximum per unit price they are willing to pay for those units. Similarly, an offer to sell consists of a maximum quantity that they are willing to sell and a minimum per unit price that they are willing to sell each of those units. If a trader did not want to trade in a period, they had the option to click a button labelled "I do not want to trade". Once all the traders have made their decisions, the computer creates demand and supply schedules from the submitted offers to buy and sell. A uniform market price is calculated as the lowest price that clears the market. Traders with offers to buy (sell) greater (less) than or equal to the market clearing price made purchases (sales). In the event of excess supply or demand at the market clearing price, successful orders were selected in the order in which they were submitted.

The experiment consisted of three treatments with quality of available AI assistance as the treatment variable: *No AI* (*NAI*), *Good AI* (*GAI*), and *Bad AI* (*BAI*). In the baseline *NAI*, traders did not have access to AI assistance and made decisions using only the information provided. In both AI treatments, participants received the same information as the baseline and were also allowed to consult with the AI up to five times per period. They could select from two default questions or submit their own formulated query. The default questions were "*If I want to sell, what should I do?*" and "*If I want to buy, what should I do?*"

Each AI advisor was programmed to receive the same instructions and data as their trader for each period. More specifically, a given AI advisor only used public information (e.g., experiment instructions, history of market clearing prices, history of buyout values, etc) and trader-specific private information (e.g., asset and cash balances, previous market interaction, previous and current AI queries, etc). At the end of each period, the AI information was automatically updated with new relevant market information just as their assigned trader was updated.

The two AI treatments differed in the quality and alignment of the advice provided. In the *GAI* treatment, the advice was based on the asset's fundamental value (FV). In contrast, the

BAI treatment offered advice aligned with the End of Period Buyout Value (BV). In order to induce this difference in the quality of advice by the AI advisor, we employed different AI models for each treatment. For the GAI, we used the gpt-4.1-2025-04-14 model, while for the BAI treatment, we used the gpt-4o-2024-05-13 model. Pretesting of the models supported consistency with the respective treatment conditions. Subjects were not informed about the quality of the AI models, nor did we mention the specific name of the model.

The timing of events for each session was as follows. (1) Participants were brought into the lab and chose a computer terminal for the session. (2) Instructions were distributed, and participants were provided approximately 15 minutes to read through the instructions on their own.[6] (3) The experimenter summarized the main features of the market and reviewed the market interface via screenshots presented on the overhead. Participants were encouraged to ask questions (to be addressed privately) throughout the experiment. (4) A computerized short quiz[7] focused on the dividend process was administered. If a participant answered a quiz question incorrectly, they were directed to the specific section of the instructions to review the concept and then asked the question again. If a participant answered incorrectly three times in a row, the program prompted them to raise their hand for explanation by the experimenter. (5) Once all participants correctly answered the quiz comprehension questions, they proceeded to a practice round. (6) Upon completion of the practice round, cash balances and asset inventories were reset for the start of the first trading period. The trading market was conducted. (7) After the market completed, an incentivized seven-item Cognitive Reflection Test (CRT-7) was administered (Toplak et al., 2014) in which participants earned $0.5 NZD for each correct answer. (8) Participants were privately paid their cash earnings and then left the experiment.

The quality of AI output

For our analysis, it is critical to understand the quality of advice participants actually received in each treatment to formulate predictions about behavior resulting from interactions with AIs. To this end, we classified the AI's advice at two levels of aggregation: the individual message level and the overall conversation level.

First, at the message level, AI advice was categorized as *consistent*, *inconsistent*, or *other*, depending on its alignment with the reference value of the specific treatment. In the GAI

[6] Instructions provided in the Appendix A2.
[7] Comprehension Quiz provided in the Appendix A3.

treatment, advice was considered *consistent* if it aligned with the Fundamental Value (FV) and *inconsistent* if it aligned with the Buyout Value (BV). In the BAI treatment, these criteria were reversed: alignment with the BV was *consistent*, and alignment with the FV was *inconsistent*. Advice that could not be strictly classified into either category (e.g., general advice lacking reference values, recommendations to use both FV and BV, or suggestions based solely on the current market price) was classified as *other*.

Because participants could submit up to five queries within a single period, we aggregated these message-level classifications to construct a conversation-level measure, which serves as our primary unit of analysis. Prior to aggregation, we excluded interactions unrelated to market decisions, such as small talk, numerical calculations, experimental instructions, or AI responses for which participant compliance could not be verified.

The remaining message-level classifications were then aggregated into a conversation-level classification. The logic for this aggregation is uniform across both treatments, relative to their respective definitions of consistency:

- **Consistent:** All specific advice within the conversation aligns with the treatment's *consistent* reference value.
- **Inconsistent:** All specific advice aligns with the treatment's *inconsistent* reference value.
- **Other:** The conversation contains mixed references (e.g., both FV and BV) or specific values outside the FV/BV framework.

As a sidenote, neutral responses (general statements lacking specific numerical references) do not independently determine the conversation's classification. Instead, they are interpreted alongside the specific advice provided; a conversation containing only neutral and *consistent* advice is classified as *consistent*, while one containing neutral and *inconsistent* advice is classified as *inconsistent*.

Table 1 shows that AI advice differs systematically across treatments. Advice in GAI was mostly consistent, with 92.4% of classifiable conversations aligning with the FV (the correct value of the asset). These results indicate that GAI performed as well as we expected. Conversely, 52.6% of the conversations for BAI were consistent and focused on BV (incorrect value of the asset).[8] This suggests that BAI indeed offers worse advice than the GAI. Yet it does not consistently produce misleading information: 16% of the time it gives advice aligned with FV, and 31.4% of the time it gives advice that is not aligned with either FV or BV. In sharp contrast, good AI gives incorrect advice only 0.4% of the time, consistent with our experimental design. In the remaining 7.2% of the time, it gives irrelevant messages (other).

Table 1. Consistency of AI advice with treatment

Advice alignment	GAI	BAI
FV	92.44%	16.00%
BV	0.40%	52.61%
Other	7.15%	31.38%
Classifiable conversations	489	325

3. Hypothesis

Prior research shows that LLMs such as ChatGPT can provide useful financial advice. Related work in experimental asset markets finds that advice can substitute for experience (Alevy and Price, 2017). We therefore expect that access to good AI advice reduces mispricing and improves market efficiency.

Hypothesis 1: In the GAI treatment, market prices will more closely track fundamental value.

However, earlier studies indicate that individuals have difficulty discerning the accuracy of AI responses (Hazra and Serra-Garcia, 2025). Individuals tend to overweight guidance from seemingly authoritative or confident sources (Price and Stone, 2004) in the context of financial advice. Receiving advice issued with a high degree of confidence by large language models (LLMs) could, therefore, generate a confidence bias among participants, who will show a high level of acceptance of the advice, even if the advice is incorrect or irrelevant.

Hypothesis 2: In the BAI treatment, market prices will deviate further from fundamental value. Access to AI advice will increase mispricing and reduce market efficiency.

The introduction of AI advice increases consensus among traders by providing a common benchmark for expectations. When traders rely on AI advice, their beliefs may become more aligned, allowing them to reach consensus more easily and thus reducing the need to trade. Without AI, traders have more heterogeneous beliefs, increasing the likelihood

[8] Examples AI generated replies aligned with each type (BV, FV and Other) are presented in Appendix A4.

of trading. As established by Harris and Raviv (1993) and Kandel and Pearson (1995), increased heterogeneity in trader beliefs and the differential interpretation of signals directly drive trading volume. Thus, the lower-quality AI inadvertently stimulated liquidity by acting as a shock to belief dispersion. This leads to our third hypothesis.

Hypothesis 3: Turnover will be highest in the NAI treatment, lower in the BAI treatment, and lowest in the GAI treatment.

4. Results

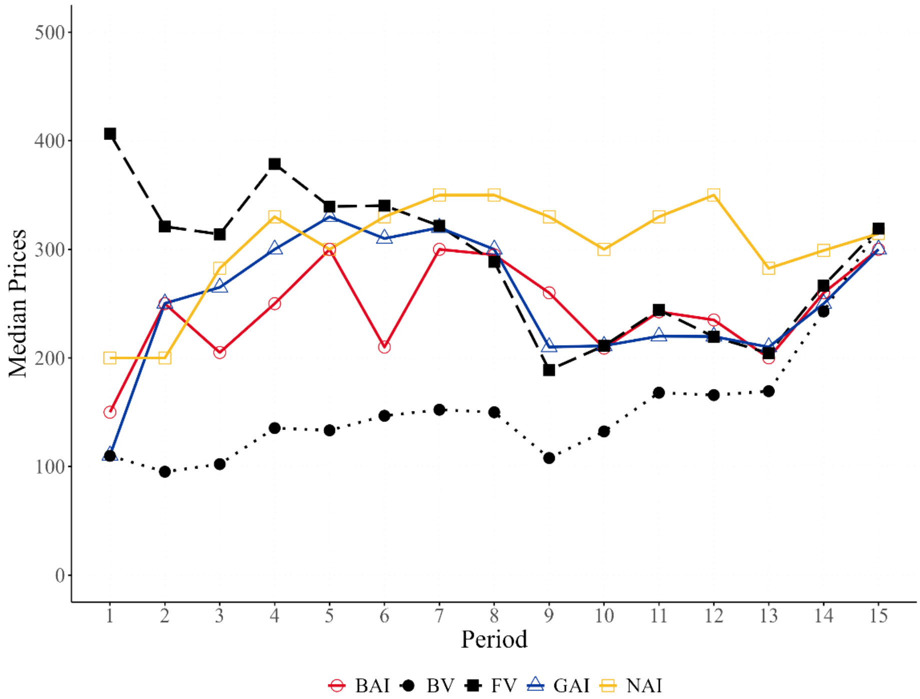
4.1. Mispricing

We examined the impact of AI advice by comparing market prices in the NAI treatment with those in the GAI and BAI treatments. Figure 1 illustrates the median prices observed for each treatment.[9] The vertical axis denotes median prices, while the horizontal axis represents the trading period. For reference, the long-dashed line indicates the asset's Fundamental Value (FV), and the dotted line shows the asset's End-of-The-Period Buyout Value (BV). The buyout value follows a random walk process explained in Section 2. In short, the Fundamental Value represents the expected holding value of the asset in a given period if the trader buys and holds the asset till the end of the market. The End-of-The-Period value represents the value of the asset if the asset were to be bought back at the end of the current period, forgoing all possible future changes in the buyout value.

Under the NAI baseline condition, prices start below the FV but above the BV, then rise above the FV after period 6, and remain at the elevated level of approximately 300 francs. Under the GAI treatment, prices in period 1 are close to BV, converge to the FV by period 5, and remain close to the FV for the remainder of the market. The presence of a good AI makes prices very close to the fundamentals. Similarly, in the BAI treatment, prices also start below FV, fluctuate between BV and FV until period 6, and then remain close to the FV until the final period. Therefore, we do not observe significant differences in terms of mispricing when AI is available, regardless of its quality. Prices are lower in both AI treatments than in the no AI baseline treatment.

[9] When an equilibrium price is not observed within a period, the minimum ask price is used, if available. This choice is consistent with the calculation of the equilibrium price when the bids and asks generate a surplus, i.e. all surplus is allocated to the buyer.

Figure 1. Time series of treatment median prices



To formally assess the magnitude of mispricing in the experiment, we used three bubble measures frequently employed in the literature: Relative Absolute Deviation (RAD), Relative Deviation (RD), and Turnover (Stöckl et al., 2010; Van Boening et al., 1993; Weber et al., 2018). The first measure, RAD, evaluates how closely prices align with fundamental values, capturing the average degree of mispricing in the market. It is defined as $RAD = \{\sum_t |P_t - FV_t| / (\sum_t (FV_t)/T)\}/T,$ where T represents the total number of periods in a market session, FV_t denotes the fundamental value in period t, and P_t is the average price in period t. The second measure, RD, indicates whether prices are on average above (RD > 0) or below (RD < 0) the fundamental values. Therefore, it is a measure of the average over or under valuation and is defined as $RD = \{\sum_t (P_t - FV_t) / (\sum_t (FV_t)/T)\}/T$. Turnover is the total number of transactions in a market session, normalized by the total units of asset available in the market, and calculated as, $Turnover = (\sum_t q_t)/TSU$, where TSU denotes the total stock of units. In addition to these conventional measures, we also include the Geometric Absolute Deviation (GAD) and Geometric Deviation (GD), introduced by Powell (2016). The interpretations of GAD and GD are the same as those of RAD and RD, and these measures satisfy numeraire independence. Their respective formulas are given as $GAD = \exp\left(\frac{1}{T}\sum_t \left|\ln\left(\frac{P_t}{FV_t}\right)\right|\right) - 1$ and $GD = (\prod_t P_t / FV_t)^{1/T} - 1$.

As shown in Table 2, in the NAI baseline condition, the average degree of absolute mispricing reaches 245% relative to the fundamental values (RAD), whereas the same statistics decrease to 33.8% and 30.4% with AI advice, for the GAI and BAI treatments, respectively. The GAD statistics show a similar pattern. Furthermore, only in NAI do we observe overpricing as indicated by positive RD and GD, while both AI treatments have negative RD and GD values due to underpricing. Specifically, in the NAI treatment, RD is positive and 216.1% above the FV on average. In contrast, the RD values in the GAI and BAI treatments indicate that prices are 3.9% and 13.8% below the FV, respectively. For Turnover, we see that each stock is traded on average 0.5 times without AI, 0.4 times with good AI, and 0.6 times with bad AI.[10]

Table 2. Average mispricing measures

Treatment	RAD	RD	Turnover	GAD	GD
NAI	2.451	2.161	0.543	1.029	0.389
	(5.006)	(5.005)	(0.345)	(0.851)	(0.894)
GAI	0.338	-0.039	0.428	0.544	-0.106
	(0.256)	(0.323)	(0.114)	(0.382)	(0.308)
BAI	0.304	-0.138	0.614	0.471	-0.184
	(0.095)	(0.138)	(0.21)	(0.217)	(0.148)
AI	0.321	-0.089	0.521	0.507	-0.145
	(0.189)	(0.247)	(0.190)	(0.305)	(0.239)

Note: The standard deviations are indicated in parentheses.

We next conduct Mann-Whitney U tests to formally examine the impact of AI advice on asset market prices. In particular, we compare the NAI baseline with each AI treatment, and we also compare the price differences between the two AI treatments. Table 3 summarizes the test results. For RAD, we see, not surprisingly, that the degree of mispricing is significantly higher in NAI than in GAI and BAI. Yet, there is no difference between GAI and BAI, regardless of the quality of AI output. When focusing on the proportion of mispricing relative to the fundamentals, GAD results point in the same direction, but the differences in GAD are not significant, even if we pool the AI treatments. The significant difference in RAD, but not in GAD, suggests the presence of more extreme price-deviation events in NAI than in AI treatments.

Similarly, the RD difference between NAI and GAI is statistically significant at the 5% level, whereas the difference between NAI and BAI is statistically significant at the 1% level. GD points in the same direction, and the differences between No AI and Bad AI treatments are

[10] Note that the maximum Turnover possible is 1.0 due to the call market trading institution.

significant, as is the case when AI sessions are pooled. Overall, the results so far demonstrate that prices are significantly higher in markets where AI advice is absent than in markets where traders are advised by AI, regardless of quality.

Table 3. *Mann-Whitney* U Test *p values*

Measure	Comparison	P value
RAD	NAI vs GAI	0.043**
RAD	NAI vs BAI	0.005***
RAD	GAI vs BAI	1.000
RD	NAI vs GAI	0.023**
RD	NAI vs BAI	0.000***
RD	GAI vs BAI	0.796
Turnover	NAI vs GAI	0.650
Turnover	NAI vs BAI	0.241
Turnover	GAI vs BAI	0.031**
GAD	NAI vs GAI	0.166
GAD	NAI vs BAI	0.052*
GAD	GAI vs BAI	0.853
GD	NAI vs GAI	0.075*
GD	NAI vs BAI	0.012**
GD	GAI vs BAI	0.739

Note: ***,**,* indicate significant difference at 1%, 5%, and 10% level, Mann-Whitney U test.

We summarize our results so far below.

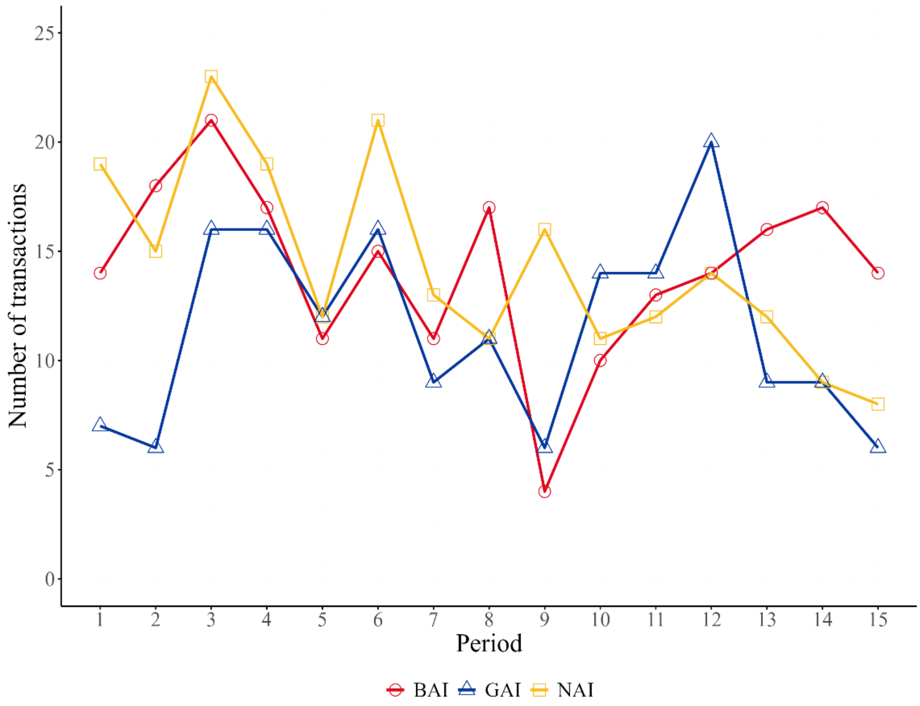
Result 1: Prices in AI-assisted markets, regardless of AI quality, are lower than in the baseline without AI.

Result 2: There is no difference between the good and bad AI markets in terms of mispricing.

Interestingly, Turnover is higher under the BAI condition than under the GAI condition, suggesting that participants are more active in making transactions when guided by BAI's bad recommendations. This difference is statistically significant at the 5% level. As will be discussed in the following section, the advice provided in the BAI treatment was considerably mixed in both quality (recommendations to follow FV or BV) and usefulness. Thus, one possible explanation for the significantly higher turnover in BAI is that when individuals received the “bad” advice, the heterogeneity in the quality and usefulness of advice created heterogeneity in beliefs. The literature has established that trading volume is positively associated with the degree of disagreements among investors (Harris and Raviv, 1993; Kandel and Pearson, 1995). Figure 2 displays the number of transactions per period for each treatment.

Result 3: The significant difference in Turnover between the GAI and BAI treatments indicates that market liquidity is lower in the GAI treatment. Hypothesis 3 is partially supported.

Figure 2. Number of transactions per period and treatment



4.2. Traders' responses to AI Advice

In the GAI and BAI treatments, participants could interact with the AI during each trading period. We analyze these interactions to assess participants' responses to advice, as well as their potential role in market outcomes.

Table 4 reports several interesting observations regarding conversations by treatment. A conversation is defined as a period during which a participant engages with AI. A participant may ask one or more questions within a given conversation. We see that participants in the GAI treatment engaged more actively in conversations than those in the BAI treatment. The average number of conversations per participant is 6.56 in GAI, but 5.24 in BAI (Mann-Whitney U test, p-value < 0.1). Note that in both treatments, they could have engaged in more conversations with AI, but chose to use only about 40% of the opportunities.[11] Another interesting observation is that about 30% of traders did not engage with the AI during the first

[11] There are 15 periods in the market, and thus 15 opportunities to engage in a conversation with AI. Thus, on average, participants engaged with AI in 6 of the 15 periods (40%). In most cases, individuals' usage decreases over time.

trading period. Furthermore, conversations decrease over time (Spearman's rho = -0.92, p-value < 0.01).

Table 4. Summary statistics of conversations across treatments

Category	GAI	BAI
Total participants	80	79
Percentage of participants with at least one conversation	93.75%	94.93%
Total questions	780	586
Total conversations	525	414
Average conversations per participant	6.56	5.24
Percentage of conversations out of the total possible per treatment	43.75%	34.94%
Percentage of participants who started their first conversation during the first period.	71.25%	72.15%
Classifiable questions	700	431
Classifiable conversations	489	325

Note: The total of possible conversations per treatment is obtained by multiplying the maximum number of possible conversations per participant (15, one per period) by the total number of participants (N).

To analyze participants' responses to the advice, we reviewed all messages a participant received within a given period (i.e., a conversation) and compared them with the decision they made during the same period. This is an advantage of the call market over a more dynamic double auction because we can clearly link their decision to a conversion in a given period in a call market. We then classified these decisions into four categories, those who directly followed the advice (Follow), those who did not follow the advice by choosing a contradicting price (Not Follow), those who did not follow the advice by choosing to make a different type of offer (Other Offer), for example, they inquired about an offer to buy but decided to make an offer to sell, and finally, those who decided not to trade even though AI advised a trading strategy (No Trade). To perform this classification, we reviewed all advice messages received by a participant over a given period and compared them with the decision made during that period.

Table 5 presents this classification by treatment type. The first column indicates the category under which the conversation responses were classified (FV, BV, or Other). The second column shows the participant's response to these conversations. Finally, the third and fourth columns show the percentages associated with each category by treatment.

Table 5. Participant responses to AI advice by treatment and advice alignment

Advice alignment	Participant response	GAI	BAI
FV	Follow	74.44%	13.23%
	Not Follow	9.41%	1.54%
	Other Offer	4.91%	1.23%
	No Trade	3.68%	0.00%
	Sum (FV Advice)	**92.44%**	**16.00%**
BV	Follow	0.20%	28.92%
	Not Follow	0.00%	16.62%
	Other Offer	0.20%	3.38%
	No Trade	0.00%	3.69%
	Sum (BV Advice)	**0.40%**	**52.61%**
Other	Follow	4.70%	25.23%
	Not Follow	1.02%	3.69%
	Other Offer	1.43%	0.92%
	No Trade	0.00%	1.54%
	Sum (Other Advice)	**7.15%**	**31.38%**
Classifiable conversations		489	325

Note: The data includes all periods. The sum of the percentages for FV, BV, and Other in each category equals 100%.

We find that, regardless of the type of advice (FV, BV, or Other), traders tend to follow the AI advice with the following rates significantly larger than 50% (Mann-Whitney U test, $p<0.001$). Furthermore, the following rates are significantly higher when the advice was aligned with FV than when it was aligned with BV (Mann-Whitney U test, $p < 0.001$)[12]. Lastly, we observe that the "Not follow" rate is higher when the advice is aligned with BV than when it is aligned with FV, showing that traders questioned the advice given by AI.

Result 4: In both treatments, participants showed a general propensity to follow AI advice, and the follow rates were higher for FV-aligned advice than for BV-aligned advice.

While Result 4 demonstrates a general propensity for participants to follow AI recommendations, it remains unclear whether participants are consciously evaluating the accuracy of the advice to make informed decisions, or if they are merely exhibiting blind reliance on the AI regardless of its quality. Because the BAI treatment exposes traders to a substantial proportion of incorrect, BV-aligned advice, disentangling true value recognition

[12] Follow rates were calculated at the participant level, considering only "Follow" or "Not follow" responses. Analyses of overall and treatment-level follow rates included all advice types (FV, BV, and Other), while analyses aggregated by advice type (FV and BV) excluded responses categorized as "Other offer" or "No trade."

from heuristic compliance is essential. To address this, we shift our unit of analysis from individual conversations to the participants themselves. By aggregating each trader's decisions over the course of the experiment, we can better assess their ability to discern the quality of advice.

Specifically, we examine participant-level behavior based on the advice received and the participant's response into three categories: *accuracy-aligned, accuracy-misaligned*, and *other*. Behavior is considered a*ccuracy-aligned* if it follows the advice associated with the FV and not that associated with the BV, while *accuracy-misaligned* behavior corresponds to decisions that do not follow advice associated with the FV but do follow that associated with the BV. Decisions that cannot be assigned to either category are classified as *other*. For each participant, we calculated the proportion of decisions in each category and reported participant-level averages. This analysis is presented in Table 6. The first column indicates the treatment, the second column includes the three participant response categories, and the third column shows the average percentage of participants for each category.

Table 6 shows that, in the GAI treatment, participants exhibited, on average, a higher rate of accuracy-aligned behavior (69.2%), which was significantly greater than the rate of accuracy-misaligned behavior (11.39%) (Wilcoxon signed rank test, p-value < 0.01). The rate of accuracy-aligned behavior in GAI was also higher than that observed in the BAI treatment (31.13%) (Mann-Whitney U test, p-value < 0.01). In contrast, in the BAI treatment, the rate of accuracy-aligned behavior did not differ significantly from the rate of accuracy-misaligned behavior (28.47%).

Notably, these behavioral rates reveal a significant departure from the actual distribution of AI advice. In the BAI treatment, the 28.47% accuracy-misaligned rate is much lower than the 52.61% proportion of BV-aligned (incorrect) advice generated by the AI. Conversely, while the AI only provided FV-aligned (correct) advice 16.00% of the time, participants still achieved an accuracy-aligned behavior rate of 31.13%. This two-sided discrepancy strongly suggests that participants do not blindly follow the AI. Instead, they actively filter out a substantial portion of incorrect recommendations and manage to capture the scarce good advice or make correct decisions independently, even when the AI fails to guide them properly. One possibility is that conversations with AI encourage some traders to think more deeply about the fundamental value of the asset. In addition, we found a positive relationship between the number of correct answers in the CRT and the percentage of accuracy-aligned decisions (Spearman's rho = 0.17, p-value < 0.1).

Table 6. Mean Proportions of Participant Response Categories in GAI and BAI Treatments

Treatment	Category	Mean
GAI	Accuracy-aligned	69.20%
	Accuracy-misaligned	11.39%
	Other	19.41%
BAI	Accuracy-aligned	31.13%
	Accuracy-misaligned	28.47%
	Other	40.39%

Finally, to assess the extent to which the market-level mispricing results are due to isolated cases, we look into extreme bids and asks that deviated greatly from the AI advice. We refer to the observation of a buyer or seller in a transaction as a participant-transaction observation. Furthermore, we isolate cases of extreme non-compliance by flagging submitted offers that deviate sharply from the reference value promoted by the AI (FV in GAI and BV in BAI). Table 7 shows the sample composition and its breakdown by AI interaction.

In the GAI treatment, we reviewed in detail the offers that deviated by 75% or more from FV. We observed that these offers were made by a small number of participants, most of whom consulted AI. Notably, one participant was responsible for nine of these offers. When analyzing all the advice participants received up to the point of making an offer, we noted that only one participant received advice not fully aligned with the FV. Therefore, in most cases, participants who posted extreme offers received advice aligned with the FV but did not follow it. This can be explained by participants' comments about the usefulness of the advice, with some indicating a lack of understanding. In other cases, the participants' initial behavior suggests a lack of understanding that was corrected in subsequent periods after continued interaction with AI. It should be noted that the participant who made the most offers in this category initially requested guidance, received a recommendation consistent with the FV, but indicated that they did not understand the advice and decided not to consult AI again.

Table 7. Sample composition and participant-transaction observations by treatment

Category	GAI	BAI
Total participants	80	79
Participants make ⩾1 transaction	74	74
Total transactions	171	212
Participant-transaction observations	263	326
Observations with offers deviating from the reference value[1]	142(56)	213(67)
Asked AI in the same period	33	71
Asked AI in a previous period	96	115
No prior AI interaction	13	27
Observations with extreme deviations (>90%) from the reference value	17(6)	87(43)
Asked AI in the same period		39(26)
Asked AI in a previous period		36(25)
No prior AI interaction		12(6)
Observations with extreme deviations (>75%) from the reference value	25(11)	
Asked AI in the same period	6(5)	
Asked AI in a previous period	17(7)	
No prior AI interaction	2(2)	

[1] The reference value is FV in GAI and BV in BAI. The values represent participant-transaction observations. The number of participants generating these observations is included in parentheses.

In the BAI treatment, we examined offers that deviated from the BV by more than 90%.[13] This generated a considerably larger number of observations than in GAI, even though a wider deviation range was considered. These offers were mostly made by participants who had previously consulted AI and were primarily buy offers, with only 5 being sell offers. Analysis of the advice received showed that all the sell offers received advice consistent with the BV, which was not followed. In contrast, the advice received by participants making their buy offers was diverse. Approximately a quarter of the offers received advice only associated with the BV, while the majority received advice associated with the "other" category, either by not including references, including references to both FV and BV, or presenting contradictory advice. Finally, a smaller fraction received advice aligned with the FV, which was followed.

Overall, the evidence indicates that extreme deviations from the reference value were more frequent and more widely distributed among participants in the BAI treatment group than in the GAI group. While in GAI these deviations were concentrated in a small group of individuals and were mainly associated with a lack of adherence to or understanding of advice consistent with the FV, in BAI they were considerably more frequent and involved a larger group of

[13] We decided to choose a more extreme threshold because in BAI, we observe more frequent deviations (also to a greater extent). The analysis remains qualitatively the same if we also chose 75% here as the reference point.

participants, as shown in Table 7. These results explain why prices in BAI are not significantly different from those in GAI.

5. Discussion and conclusion

In this paper, we report a controlled asset market experiment to study the causal impact of AI advice on price discovery and liquidity. By introducing AI advisors with varying levels of accuracy into a market environment, we aim to study whether traders follow good or bad advice without exercising critical thinking. We observed severe overpricing in the baseline condition without AI. With the presence of an AI advisor, regardless of its quality, the market is initially slightly underpriced and appears to successfully mitigate mispricing by the midpoint of the market. The similarity of AI treatments emerges despite stark differences in the quality of the advice provided. In the GAI treatment, the advice predominantly anchored participants to the Fundamental Value (FV). Conversely, the BAI provided a noisy mix of advice, heavily weighted toward the incorrect Buyout Value (BV) and irrelevant information (categorized as Other).

The lack of divergence in aggregate mispricing stems from participants' heterogeneous adherence behaviors. While traders generally followed AI recommendations, their compliance was notably higher when the advice aligned with the true FV. Consequently, participants in the BAI treatment actively filtered out a portion of the poor advice, preventing the market from fully degrading to the BV anchor. Notably, extreme deviations from reference values in our participant-transaction data were predominantly driven by traders explicitly ignoring the AI (e.g., because of a lack of understanding), rather than blindly following poor recommendations. The discussion with AI might have encouraged the traders to consider the fundamentals of the asset more, instead of solely focusing on the price momentum. Moreover, the repeated reiterations of AI on either the fundamental value or the end-of-period value of the asset provide another anchor that draws attention away from the price momentum, and thus attenuates bubbles. Our findings demonstrate the necessity of examining not just the output accuracy of these AI systems, but the human-AI interaction paradigm that translates AI outputs into executed trades (Caro et al., 2026). While mispricing magnitudes were similar, AI quality strongly affects market liquidity. Turnover was systematically lower in the GAI treatment compared to the BAI treatment. This aligns with foundational theoretical frameworks regarding trading volume. The consistent, FV-aligned advice in the GAI treatment harmonized investor expectations, thereby reducing the motive to trade. In contrast, the contradictory and varied advice generated by the BAI system fostered dispersed expectations among participants. As established by Harris and Raviv (1993) and Kandel and Pearson (1995), increased heterogeneity in trader beliefs and the differential interpretation of signals directly drive trading volume. Thus, the lower-quality AI inadvertently stimulated trade by acting as a shock to belief dispersion.

Our findings contribute to the literature on financial advice. Previous experimental work highlights that access to advice can act as a substitute for endogenous expertise, mitigating price deviations and enhancing market efficiency (Alevy and Price, 2017). Our baseline versus AI-based advice results strongly corroborate this mechanism. However, our results also expose the vulnerabilities associated with AI reliance. The high tendency of adherence to AI advice, even when it deviated from fundamentals in the BAI treatment, echoes findings on confidence bias, wherein individuals overweight guidance from seemingly authoritative or confident sources (Price and Stone, 2004) in the context of financial advice. The difficulty participants had in independently assessing the AI's accuracy highlights a critical friction point in human-AI interaction within complex financial markets, or any complex environment in general.

The design of our experiment also makes methodological contributions. First, since the seminal work of Smith et al. (1988), experimental markets have frequently employed deterministically declining fundamental values, which reliably generate price bubbles. However, subsequent literature (e.g., Huber et al., 2008; Kirchler and Huber, 2007; Kirchler et al., 2012; Lambrecht et al., 2025; Noussair et al., 2022; Stöckl et al., 2015) demonstrates that altering this process to a random walk or to a constant fundamental value can significantly reduce or eliminate bubble formation. We show in a call market that the random walk (with a drift) process employed in our study can also reliably generate bubbles. In the baseline (NAI) condition, prices started below the FV but eventually rose significantly above it, remaining at elevated levels. This provides an alternative framework to study bubbles and crashes in controlled lab environments. Furthermore, the market experiment framework can easily be expanded by incorporating more advanced AI models or a more complex trading institution, such as an open-book double-auction mechanism.

Our paper is a first step in understanding how AI affects trading and investment decisions in a market setting. As AI advice becomes ubiquitous in our lives, a deeper understanding of how human-AI interaction affects humans' decisions is warranted. Several promising avenues for future research. First, while our study varied the underlying accuracy of the AI, future experiments could manipulate the AI's confidence or tone of the advice to isolate how presentation format affects investor compliance. Second, it would be valuable to expand our call-market framework to a more competitive, dynamic trading environment, such as an open-book double-auction market, and observe how AI advisors influence human strategy with real-time market input. Finally, investigating these dynamics with professional traders, rather than retail proxies, may help clarify whether domain expertise leads to better questions and can more reliably assess the quality of the AI output. The AI tool may lead to very different outcomes when it is in the hands of people who know how to use it best.

References

Alevy, J. E., and Price, M. K. (2017). Advice in the marketplace: A laboratory study. Experimental Economics, 20(1), 156-180. https://doi.org/10.1007/s10683-016-9480-5

Angerer, M., and Szymczak, W. (2019). The impact of endogenous and exogenous cash inflows in experimental asset markets. Journal of Economic Behavior & Organization, 166, 216-238. https://doi.org/10.1016/j.jebo.2019.08.009

Bai, H., Voelkel, J. G., Muldowney, S., Eichstaedt, J. C., and Willer, R. (2025). LLM-generated messages can persuade humans on policy issues. Nature Communications, 16(1), 6037. https://doi.org/10.1038/s41467-025-61345-5

Basu, S., Garimella, A., Han, W., and Dennis, A. R. (2021). Human decision making in AI augmented systems: Evidence from the initial coin offering market. Proceedings of the 54th Annual Hawaii International Conference on System Sciences, HICSS 2021, 176-185. https://doi.org/10.24251/HICSS.2021.020

Bianchi, M., and Brière, M. (2026). Human-robot interactions in investment decisions. Management Science, 72(1), 14-31. https://doi.org/10.1287/mnsc.2022.03886

Caginalp, G., Porter, D., and Smith, V. L. (1998). Initial cash/asset ratio and asset prices: An experimental study. Proceedings of the National Academy of Sciences of the United States of America, 95(2), 756-761. https://doi.org/10.1073/pnas.95.2.756

Caro, F., Colliard, J.-E., Katok, E., Ockenfels, A., Stier-Moses, N., Tucker, C., and Wu, D. J. (2026). Introduction to the special issue on the human-algorithm connection. Management Science, 72(1), 1-13. https://doi.org/10.1287/mnsc.2023.intro.v72.n1

Cason, T. N., and Friedman, D. (1997). Price formation in single call markets. Econometrica, 65(2), 311-345. https://doi.org/10.2307/2171895

Castelo, N., Bos, M. W., and Lehmann, D. R. (2019). Task-dependent algorithm aversion. Journal of Marketing Research, 56(5), 809-825. https://doi.org/10.1177/0022243719851788

Chen, D. L., Schonger, M., and Wickens, C. (2016). oTree: An open-source platform for laboratory, online, and field experiments. Journal of Behavioral and Experimental Finance, 9, 88-97. https://doi.org/10.1016/j.jbef.2015.12.001

Cheng, Q., Lin, P., and Zhao, Y. (2025). Does generative AI facilitate investor trading? Early evidence from ChatGPT outages. Journal of Accounting and Economics, 80(2), 101821. https://doi.org/10.1016/j.jacceco.2025.101821

Davis, A. M., Mankad, S., Corbett, C. J., and Katok, E. (2024). OM Forum: The best of both worlds: Machine learning and behavioral science in operations management. Manufacturing & Service Operations Management, 26(5), 1605-1621. https://doi.org/10.1287/msom.2022.0553

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114-126. https://doi.org/10.1037/xge0000033

eToro. (2025). US retail investors flock to AI tools, with usage surging 75% in one year. https://www.etoro.com/en-us/news-and-analysis/latest-news/press-release/us-retail-investors-flock-to-ai-tools-with-usage-surging-75-in-one-year/

Friedman, D. (1993). How trading institutions affect financial market performance: Some laboratory evidence. Economic Inquiry, 31(3), 410-435. https://doi.org/10.1111/j.1465-7295.1993.tb01302.x

Gillespie, N., Lockey, S., Curtis, C., Pool, J., and Ali, A. (2023). Trust in artificial intelligence: A global study. The University of Queensland and KPMG Australia. https://doi.org/10.14264/00d3c94

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. Journal of the Economic Science Association, 1(1), 114-125. https://doi.org/10.1007/s40881-015-0004-4

Hargreaves Heap, S. P., and Zizzo, D. J. (2011). Emotions and chat in a financial markets experiment. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.1783462

Harris, M., and Raviv, A. (1993). Differences of opinion make a horse race. The Review of Financial Studies, 6(3), 473-506. https://doi.org/10.1093/rfs/6.3.473

Haruvy, E., and Noussair, C. N. (2006). The effect of short selling on bubbles and crashes in experimental spot asset markets. The Journal of Finance, 61(3), 1119-1157. https://doi.org/10.1111/j.1540-6261.2006.00868.x

Hazra, S., and Serra-Garcia, M. (2025). Understanding trust in AI as an information source: Cross-country evidence (CESifo Working Paper No. 11954). https://doi.org/10.2139/ssrn.5327285

Huber, J., Kirchler, M., and Sutter, M. (2008). Is more information always better? Experimental financial markets with cumulative information. Journal of Economic Behavior & Organization, 65(1), 86-104. https://doi.org/10.1016/j.jebo.2005.05.012

Kandel, E., and Pearson, N. D. (1995). Differential interpretation of public signals and trade in speculative markets. Journal of Political Economy, 103(4), 831-872. https://doi.org/10.1086/262005

Kirchler, M. (2009). Underreaction to fundamental information and asymmetry in mispricing between bullish and bearish markets: An experimental study. Journal of Economic Dynamics and Control, 33(2), 491-506. https://doi.org/10.1016/j.jedc.2008.08.002

Kirchler, M., and Huber, J. (2007). Fat tails and volatility clustering in experimental asset markets. Journal of Economic Dynamics and Control, 31(6), 1844-1874. https://doi.org/10.1016/j.jedc.2007.01.009

Kirchler, M., Huber, J., and Stöckl, T. (2012). Thar she bursts: Reducing confusion reduces bubbles. American Economic Review, 102(2), 865-883. https://doi.org/10.1257/aer.102.2.865

Lambrecht, M., Sofianos, A., and Xu, Y. (2025). Does mining fuel bubbles? An experimental study on cryptocurrency markets. Management Science, 71(3), 1865-1888. https://doi.org/10.1287/mnsc.2022.01238

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90-103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lopez-Lira, A., and Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. arXiv. https://doi.org/10.48550/arXiv.2304.07619

Nicholson-Messmer, E. (2025). Clients are turning to ChatGPT to double-check: “Is my advisor right?” Yahoo Finance. https://finance.yahoo.com/news/clients-turning-chatgpt-double-check-194427141.html

Noussair, C. N., and Tucker, S. (2016). Cash inflows and bubbles in asset markets with constant fundamental values. Economic Inquiry, 54(3), 1596-1606. https://doi.org/10.1111/ecin.12320

Noussair, C., Tucker, S., and Ryan, M. (2022). The effect of favorable and unfavorable information on asset prices. In S. Füllbrunn and E. Haruvy (Eds.), Handbook of experimental finance (pp. 194-212). Edward Elgar Publishing. https://doi.org/10.4337/9781800372337.00024

Oechssler, J., Schmidt, C., and Schnedler, W. (2011). On the ingredients for bubble formation: Informed traders and communication. Journal of Economic Dynamics and Control, 35(11), 1831-1851. https://doi.org/10.1016/j.jedc.2011.05.009

Powell, O. (2016). Numeraire independence and the measurement of mispricing in experimental asset markets. Journal of Behavioral and Experimental Finance, 9, 56-62. https://doi.org/10.1016/j.jbef.2015.11.002

Price, P. C., and Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. Journal of Behavioral Decision Making, 17(1), 39-57. https://doi.org/10.1002/bdm.460

Smith, V. L., Van Boening, M., and Wellford, C. P. (2000). Dividend timing and behavior in laboratory asset markets. Economic Theory, 16(3), 567-583. https://doi.org/10.1007/PL00020943

Smith, V. L., Suchanek, G. L., and Williams, A. W. (1988). Bubbles, crashes, and endogenous expectations in experimental spot asset markets. Econometrica, 56(5), 1119-1151. https://doi.org/10.2307/1911361

Steiger, S., and Pelster, M. (2020). Social interactions and asset pricing bubbles. Journal of Economic Behavior & Organization, 179, 503-522. https://doi.org/10.1016/j.jebo.2020.09.020

Stöckl, T., Huber, J., and Kirchler, M. (2010). Bubble measures in experimental asset markets. Experimental Economics, 13(3), 284-298. https://doi.org/10.1007/s10683-010-9241-9

Stöckl, T., Huber, J., and Kirchler, M. (2015). Multi-period experimental asset markets with distinct fundamental value regimes. Experimental Economics, 18(2), 314-334. https://doi.org/10.1007/s10683-014-9404-1

Toplak, M. E., West, R. F., and Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. Thinking & Reasoning, 20(2), 147-168. https://doi.org/10.1080/13546783.2013.844729

Van Boening, M. V., Williams, A. W., and LaMaster, S. (1993). Price bubbles and crashes in experimental call markets. Economics Letters, 41(2), 179-185. https://doi.org/10.1016/0165-1765(93)90194-H

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., and Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), Article 129. https://doi.org/10.1145/3579605

Wang, H., Zhang, Y., and Lu, T. (2026). The power of disagreement: A field experiment to investigate human-algorithm collaboration in loan evaluations. Management Science, 72(1), 96-118. https://doi.org/10.1287/mnsc.2022.03844

Weber, M., Duffy, J., and Schram, A. (2018). An experimental study of bond market pricing. The Journal of Finance, 73(4), 1857-1892. https://doi.org/10.1111/jofi.12695

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv. https://doi.org/10.48550/arXiv.2302.11382

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q. L., and Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. IEEE/CAA Journal of Automatica Sinica, 10(5), 1122-1136. https://doi.org/10.1109/JAS.2023.123618

Yang, C., Bauer, K., Li, X., and Hinz, O. (2026). My advisor, her AI, and me: Evidence from a field experiment on human-AI collaboration and investment decisions. Management Science, 72(1), 242-264. https://doi.org/10.1287/mnsc.2022.03918

Zhou, Y., Kim, K., and Xue, L. (2024). How machine-generated ratings and social exposure affect human reviewers: Evidence from initial coin offerings. Proceedings of the 57th Hawaii International Conference on System Sciences, 4353-4362. https://doi.org/10.24251/HICSS.2024.524

Appendix A

A 1. Path

For path selection, 100,000 price path simulations were generated using a random walk model with parameters estimated from historical S&P 500 data (1928-2023). Each simulation considered 15 periods and an initial price of 100. First, annual returns were simulated with a normal distribution, and then price paths were generated by accumulating those returns. Subsequently, statistics were calculated for each path, and the median of each statistic was calculated. For each path, the absolute distance between each statistic and the median was calculated. Each path was then assigned a ranking for each statistic, based on its closeness to the median. An aggregate ranking was then constructed that included the average return, cumulative return, volatility, skewness, kurtosis, and maximum drawdown for each path, choosing the path with the lowest aggregate ranking.

A 2. Experiment Instructions

1. General Instructions

This is an experiment in the economics of market decision-making. The instructions are simple and if you follow them carefully and make good decisions, you might earn a considerable amount of money, which will be paid to you in cash at the end of the experiment. The experiment will consist of fifteen trading periods in which you will have the opportunity to buy and sell in a market. The currency used in the market is francs. All trading and earnings will be in terms of francs.

________ francs = 1 NZ dollar

Your francs will be converted to dollars at this rate, and you will be paid in dollars when you leave the lab today. The more francs you earn, the more dollars you earn.

In each period, you can make an offer to buy or sell units of an asset called X, or you can decide not to trade. Your inventory of X carries over from one trading period to the next, meaning that your cash and inventory of X at the beginning of period 2 will remain the same as at the end of period 1.

Buyout Value

At the end of the *last period*, the experimenter will buy back any assets you hold in your inventory. The amount the experimenter pays for each asset at the end of the experiment is called the Buyout Value.

The Buyout Value is a random process that evolves over time. At the start of the market, the buyout value is equal to 100 francs. At the end of each period, the buyout value will randomly change. The amount of change follows a normal distribution with an average of 9.80% and a standard deviation of 19.55%. This means that the buyout value will increase by 9.80% in every period on average, but not with certainty. We explain this random process below in more detail.

End of Period Buyout Value

A random change in Buyout Value occurs at the end of each period. Therefore, during a given period, you do not know with certainty what the Buyout Value will be at the end of that period.

For example, the End of Period Buyout Value at the end of period 1 is expected to be 109.8 francs. That is, the 100 francs at the start of the period plus the expected increase by 9.8 (= 100 * 9.80%).

NOTE: the end of period change in the Buyout Value will not always be exactly the expected value because the variations can be larger or smaller. More specifically, the magnitude of these fluctuations is measured by the standard deviation, which in this case is 19.55%. This implies that the realized end of period buyout value could be above or below the average of 109.8 francs.

For example, at the end of the first period,

- there's approximately a 68% chance that the buyout value after one period will be between 80.45 and 119.55 francs, which is about one standard deviation around the mean.
- there's approximately a 95% chance that the buyout value after one period will be between 60.9 and 139.1 francs, which is about two standard deviations around the mean.

As you can see, it is possible for the End of Period Buyout Value to increase or decrease each period. However, on average, the End of Period Buyout Value will increase by 9.80% each period. That is, there is an expected upward trend in buyout value across the 15 periods of the market.

End of Market Expected Buyout Value

Even though you can see how the Buyout Value evolves over time, you cannot sell the asset to the experimenter before the end of the last period. The End of Period Buyout Value is there to inform you about the process under which the buyout value reaches the final payout amount. This is called the End of Market Buyout Value.

At any point in time during the market, it is possible to calculate the End of Market Expected Buyout Value. In other words, due to the random process, we cannot know for sure what the buyout value will be at the end of the market, but we can calculate what it will be on average. This value is calculated assuming the change in Buyout Value is a positive 9.80% per period. Below is a discussion and examples of how this is calculated:

First Period Calculation:

- **Starting Point:** The initial Buyout Value is 100 francs.
- **Expected Growth:** Assuming all subsequent growth to be 9.80% per period.
- **Calculation:** We project this 9.80% growth over the remaining 15 periods.
  - Formula: $100\times(1+0.098)^{15} = 406.5$ francs.

Second Period Calculation:

- **New Starting Point:** Suppose the Buyout Value at the end of the first period realized to be 110 francs, higher than the expected due to fluctuations.
- **Expected Growth:** still assuming all subsequent growth to be 9.80% per period.
- **Calculation:** Now, we calculate the expected value from this new starting point over the remaining 14 periods.
    - Formula: $110 \times (1+0.098)^{14} = 407.2$ francs.

Third Period Calculation:

- **New Starting Point:** Suppose the Buyout Value at the end of the second period realized to be 118 francs, lower than the expected due to fluctuations.
- **Expected Growth:** still assuming all subsequent growth to be 9.80% per period.
- **Calculation:** Now, we calculate the expected value from this new starting point over the remaining 13 periods.
    - Formula: $118 \times (1+0.098)^{13} = 397.8$ francs.

The End of Market Expected Buyout Value will be calculated and presented on the bidding screen throughout the market.

Recall that the experimenter will pay each subject the End of Market Buyout Value for each unit of the asset held in their inventory at the end of the experiment (end of period 15). Therefore, the End of Market Expected Buyout Value calculates this expected payment for each asset at any point in time during the market.

To express more simply, the End of Market Expected Buyout Value estimates how much on average the experimenter will pay you for each unit of the asset you hold in your inventory at the end of the market (end of period 15).

Artificial Intelligence (AI) Advisor

During each period, you will have access to an AI that you may use as a financial advisor to assist with your decisions if you choose. In each period, we provide two sample questions to give you an idea of how you might use it. The AI has access to the experiment instructions as well as information for the period that you can see on the bidding screen, such as your inventory (cash, X units), End of Period Expected Buyout Value, End of Market Expected Buyout Value, etc. Additionally, at the end of the period the AI advisor will have access to your inventory at the end of the period and the market price of X.

2. Your Earnings

At the start of the experiment, you will receive 41,000 francs in your cash inventory. Your earnings in the experiment equal your cash inventory at the end of period 15 plus any cash you receive from buying back any units of X you have at the end of the experiment.

NOTE: As discussed in the previous section, the End of Market Expected Buyout Value provides an estimate of the cash you will receive from buying back any units of X. So, at any point in time during the market, you are provided with the expected end of market earnings from holding a unit of X at the end of period 15.

All money spent on purchases is subtracted from your cash inventory.

All money received from sales is added to your cash inventory.

Example of earnings at the end of period 15: If you have 6 units of X at the end of period 15 and the buyout value at the end of period 15 is 523, then your earnings from buying back at the end of period 15 equal 6 units x 523 francs = 3,138 francs, plus any cash you have in your inventory.

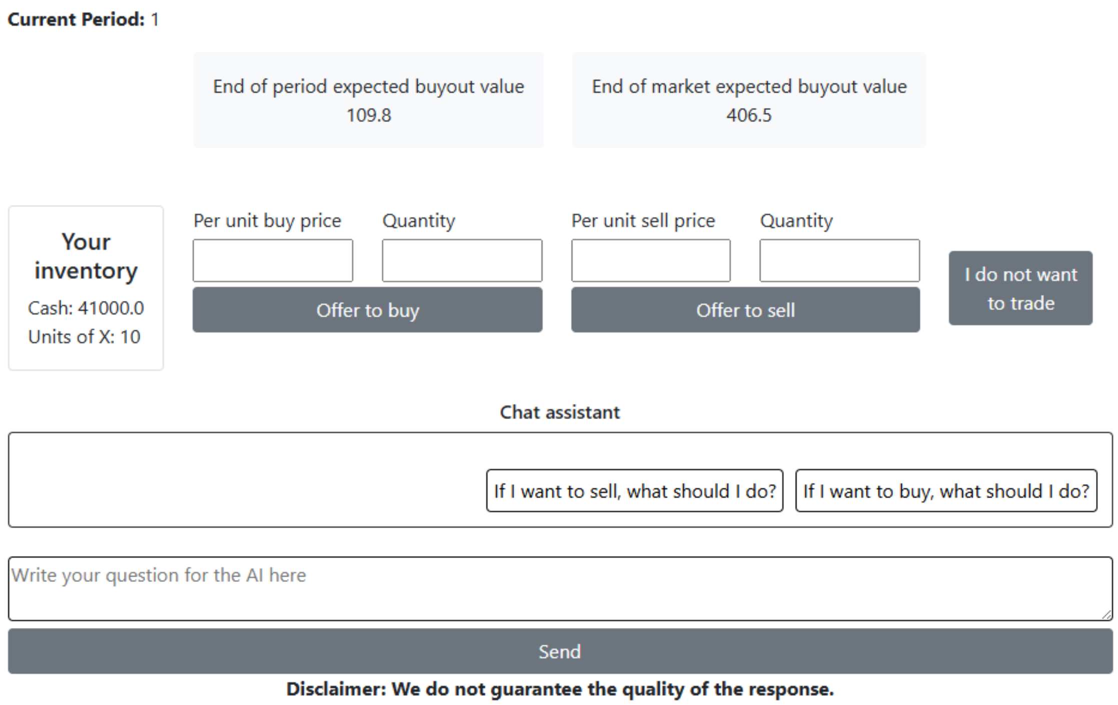
3. Market Rules and Trading

At the beginning of the experiment, you will have an initial inventory of 10 units of X and 41,000 francs. The experiment will consist of 15 periods. In each period, you will see a computer screen like the one below. You can use the interface to make an offer to buy or sell units of X or make no offers (decide not to trade in that period). On your computer screen, you can see the Cash and the quantity of Units of X that you have available.

At the start of each trading period, if you want to buy units of X, you can submit a buy offer. Your buy offer indicates the number of units of X that you would like to buy and the highest price that you are willing to pay FOR EACH UNIT. Conversely, if you want to sell units of X, you can submit a sell offer.

Your sell offer indicates the number of units of X that you are offering to sell and the lowest price that you are willing to accept FOR EACH UNIT. In each period, you can either buy or sell or do not trade. You cannot buy and sell in a given period. **The price you specify in your order is a price per unit, at which you offer to buy or sell each unit of X.**

Example of the Trading Screen for Period 1



Your offer to sell is limited by your inventory of X, and your offer to buy is limited by your cash holdings and the maximum price that you are willing to pay multiplied by the quantity of X you want to buy. The period ends when all market participants have made a decision. An example of the bidding screen is provided above.

At the end of the period, the computer program will organize the buy and sell orders and use them to determine the trading price at which units of X are bought and sold. All transactions in a given period will be made at the same trading price. This will generally be a price at which the quantity of units of X with sell order prices at or below this trading price is equal to the quantity of units of X with buy order prices at or above this trading price. People who submit buy orders at prices above the trading price are buying, and those who submit sell orders at prices below the trading price make sales.

Examples of how the market works.

Example 1. Suppose that in period 2, there are four traders in the market and:

- Trader 1 submits a buy offer at 60

- Trader 2 submits a buy offer at 20
- Trader 3 submits a sell offer at 10
- Trader 4 submits a sell offer at 40

At any price above 40, more units are offered for sale than for purchase. At any price below 20, more units are offered for purchase than for sale. At any price between 20.1 and 39.9, an equal number of units are offered for purchase and sale. The trading price is the lowest price at which an equal number of units are offered for purchase and sale. In this example, the price is 20.1. Trader 1 makes a purchase from Trader 3 at a price of 20.1.

Example 2. Suppose that in period 2, four traders participate in the market and:

- Trader 1 submits a buy offer at 360
- Trader 2 submits a buy offer at 320
- Trader 3 submits a sell offer at 310
- Trader 4 submits a sell offer at 300

The trading price is the lowest price at which there is an equal number of units offered for purchase and sale. In this example, the price is 310. Traders 1 and 2 make a purchase from Traders 3 and 4 at a price of 310.

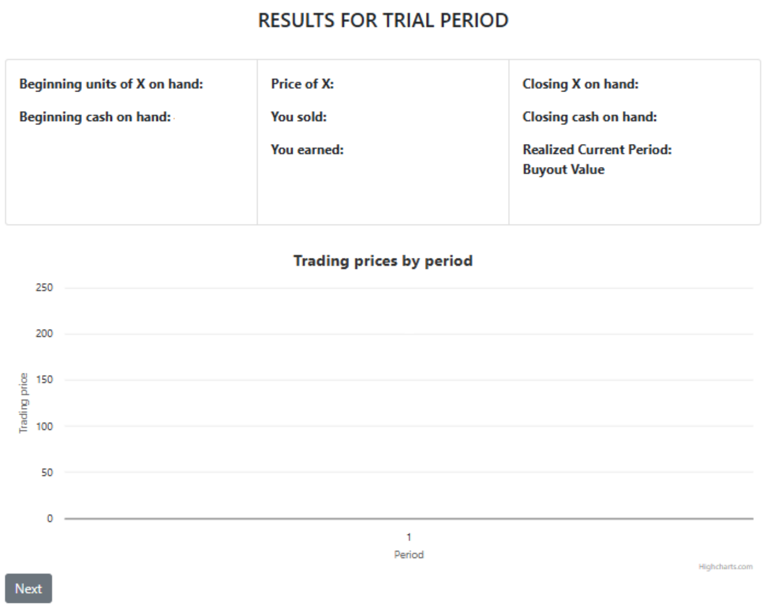
4. Opinions of AI

After you make your market decisions for the period, we will ask you your opinion on the quality/usefulness of the AI assistance. If you did not use AI in the period, then this question will not be available to you.

5. Summary of Market Outcomes

At the end of each period, you will see a summary screen (an example of the summary screen is shown below).

Example of a Summary screen



6. Summary of the instructions

This experiment involves market decision-making over 15 trading periods, using francs as the currency. Participants can buy or sell an asset called X or choose not to trade. The inventory of X and cash carries over each period. At the end of the experiment, the experimenter will buy back any remaining assets at the End of Market Buyout Value.

Buyout Value

- **Initial Value:** 100 francs.
- **Evolution:** Changes randomly each period, with an average increase of 9.80% and a standard deviation of 19.55%.
- **End of Period:** The Buyout Value is expected to increase but can fluctuate due to randomness. Participants will not know the exact Buyout Value during a given period until it concludes.

End of Period Expected Buyout Value

- Calculated based on a consistent 9.80% growth assumption.
- **Example Calculations:**
    - **Period 1:** $100\times(1+0.098) = 109.8$ francs, calculated from an initial 100 francs with an anticipated increase of 9.8% (9.8 francs).

End of Market Expected Buyout Value

- Calculated based on a consistent 9.80% growth assumption.
- **Example Calculations:**
    - **Period 1:** $100\times(1+0.098)^{15} = 406.5$ francs.
    - **Period 2:** If realized value is 110 francs, then $110\times(1+0.098)^{14} = 407.2$ francs.
    - **Period 3:** If realized value is 118 francs, then $118\times(1+0.098)^{13} = 397.8$ francs.

AI Assistant

- Participants can consult ChatGPT for advice during each period.
- ChatGPT has access to the experiment instructions and current period data, including inventory and expected values.

Earnings

- **Initial Cash:** 41,000 francs.
- **Final Earnings:** Cash at the end of period 15 plus buyout value of remaining X units.

- **Example:** 6 units of X at a buyout value of 523 francs results in 3,138 francs plus remaining cash.

Market Rules and Trading

- **Initial Inventory:** 10 units of X and 41,000 francs.
- **Trading Interface:** Participants can submit buy or sell offers, specifying the number of units and price per unit. Or choose not to trade.
- **Trading Price:** Determined by matching buy and sell orders at the lowest price where quantities match.

Trading Examples

- **Example 1:** Buy offers at 60 and 20, sell offers at 10 and 40. Trading price is 20.1.
- **Example 2:** Buy offers at 360 and 320, sell offers at 310 and 300. Trading price is 310.
- At the end of each period, participants will see a summary of their trading activity and current inventory.

A 3. Comprehension Quiz

Question 1: Suppose that you purchase a unit of X in period 6.

a. What is the average change in the buyout value of a unit of X in period 6? Please enter a number to the first decimal place (without a percentage symbol). For example, if the value is 6.8%, enter 6.8. _9.8_
b. What is the value of the standard deviation of the random change? Please enter a number to the second decimal place (without a percentage symbol). For example, if the value is 6.80%, enter 6.80._19.55_

Question 2: Suppose that you purchase a unit of X in period 15.

a. What is the average change in the buyout value of a unit of X in period 15? Please enter a number to the first decimal place (without a percentage symbol). For example, if the value is 6.8%, enter 6.8. _9.8_
b. What is the value of the standard deviation of the random change? Please enter a number to the second decimal place (without a percentage symbol). For example, if the value is 6.80%, enter 6.80._19.55_

Question 3: Suppose that in period 2 four traders participate in the market and:

- Trader 1 submits an offer to buy at 400
- Trader 2 submits an offer to buy at 120
- Trader 3 submits an offer to sell at 310
- Trader 4 submits an offer to sell at 300

a. The trading price is the lowest price at which there is an equal number of units offered for purchase and for sale. In this example that price is _300_.

b. At the trading price, trader (s) _1_ make a purchase from trader (s) _4_.

Question 4: What would be your profit at the end of period 15 if you have 7 units, the price trading is 60, the buyout value is 70 and you have 1,300 francs? _1790_.

A 4. Examples of responses by AI category criteria

Each message received by AI was evaluated to determine its alignment with either the BV, FV or Other. The table below presents examples of messages aligned with each of the three types. For each type, a single message reply from AI is presented. For example, the reply message associated with FV is classified as such because it provides a purchasing strategy specifically related to the fundamental value for that period, i.e. 313.9 francs.

Type	AI Reply Message
FV	• **Reference the End of Market Expected Buyout Value:** The expected value per unit at the end of the experiment is 313.9 francs. • **Consider Previous Market Activity:** No trades occurred last period, possibly due to low buy offers or high sell offers. • **Increase Your Buy Offer**: To improve your chances, consider offering a price higher than your previous offer, but still below 313.9 francs for potential profit. • **Suggested Offer Range**: A buy offer between 200 and 270 francs per unit balances value and likelihood of a successful trade. • **Monitor Seller Behavior**: If sellers are not accepting lower prices, moving your offer closer to the expected value may help. • **Stay Within Budget**: Ensure your total offer (price × quantity) does not exceed your available cash (41,000 francs).
BV	• **Review Your Inventory**: You have 10 units of X and 41,000 francs in cash. • **Determine Your Buying Price**: Decide the highest price per unit of X you are willing to pay. Consider the End of Period Expected Buyout Value of 109.8 francs as a reference. • **Enter Buy Offer**: On the trading screen, enter the number of units of X you want to buy and the price per unit you are willing to pay. • **Submit Offer**: Click the "Offer to buy" button to submit your buy offer. • **Wait for Market Clearing**: The market will match buy and sell offers to determine the trading price. If your offer is matched, you will purchase units at the trading price. • **Monitor Results**: At the end of the period, check the summary screen to see if your buy offer was successful and review your updated inventory and cash balance. • **Adjust Strategy**: Based on the results, adjust your strategy for the next period if necessary.
Other	• **Review Your Cash:** You have 41,000 francs available for purchasing units of X. • **Determine Buy Price:** Decide the highest price per unit of X you are willing to pay. • **Enter Quantity:** Specify the number of units of X you want to buy. • **Submit Buy Offer:** Enter the price per unit and the quantity in the buy fields on your trading screen. • **Consider Expected Values:** The End of Market Expected Buyout Value is 406.5 francs, and the End of Period Expected Buyout Value is 109.8 francs. Use these values to guide your decision. • **Ensure Sufficient Cash:** Make sure the total cost (price per unit * quantity) does not exceed your available cash. • **Monitor Market:** Keep an eye on the market to see if your buy offer is accepted.