Human and Artificial Agents in Social Decision-Making: Experimental, Computational, and Neurobiological Evidence

Thesis by
Thomas Henning

In Partial Fulfillment of the Requirements for the
Degree of
Social Decision Neuroscience



CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2027
Defended June 25, 2026

© 2027

Thomas Henning
ORCID: 0009-0005-5423-9142

All rights reserved

ACKNOWLEDGEMENTS

I feel deeply grateful to have had the chance to do my graduate studies at Caltech. I never imagined the amount of knowledge and experiences this school would give me, I am an entirely changed person due to the incredible people I've had the chance to study and interact with.

To Antonio Rangel, I cannot overstate the positive influence you've had on my PhD and my life. Taking a student with little scientific experience under your wing, and managing my eccentricities, headstrongness, and ambition with such care and grace. You made me a better scholar, leader, and person. I will never forget the numerous long and detailed discussions about science, politics, and life in general. There was never a moment when I doubted that you had my back, and I move on to the next phase of my life without any doubt that this support would continue; the confidence that gives me can truly not be overstated. I could not have asked for a better advisor/mentor.

To Colin Camerer, You were the first person I ever spoke to at Caltech, and I remain grateful for the role that conversation played in bringing me here. I have also valued the moments we spent brainstorming ideas and thinking through research questions together.

To John O'Doherty, Thank you for your guidance over these last few years, and for grounding you gave me in neuroimaging. You made sure I went into every project with my eyes open. That balance of realism and encouragement is rare, and it made a real difference in how I learned to approach science. I am also grateful for our conversations about academia, funding, and the profession more broadly; they helped me understand not only how research gets done, but how the academic world works. Thank you for your generosity, honesty, and steady support, and for many great jokes and laughs.

To Mike Alvarez, I am deeply grateful for all your encouragement over the years. I will always appreciate your guidance, and you willingness to find time to talk with me and offer thoughtful advice. The sports science class with you and Coach Eslinger was one of the most interesting and exciting courses I have ever taken, and it opened me up to the world of sports analytics and business. I have rarely met a

scholar as generous and gracious with his time as you, and I am grateful that I could always come to you for advice.

Thank you, Anastasia Buyalskaya; you have truly been a godsend for me during this PhD. I am so grateful for getting to know you over the last few years and for being able to work on such an interesting chapter of work. Your creativity and general approach to science is something I strive to emulate.

There are more mentors who shaped my PhD than I could possibly name here, but I want to specifically thank Brenden Eum, Marcos Gallo, Deshawn Sambrano, Zeynep Enkavi, Pantelis Vafidis, Carina Hausladen, Nina Rouhani, and Mihai Toma. Your guidance, support, and example helped carry me through this process, and I can safely say that without you, I would not be defending my PhD.

To Other Professors and Mentors: Michael Pierce, Michael Crow, Greg Samanez-Larkin, Sam McClure, without you, this journey would never have started. Uma Karamarkar and Ryan Webb, for the consistent support, you always made me feel like I belonged. Bil Clemons, David Van Valen, John Dabiri, Adam Wierman, Eric Mazumader, Omer Tamuz, Ralph Adolphs, Dean Mobbs, Charlie Sprenger, Kirby Nielson, Marina Agarnov, Alec Smith, Tomomi Tanaka.

To my coauthors, Thank you for sticking with me through the wild ideas, the failed and revived experiments, and the highs and lows that come with doing scholarship. Ross Spoon, Siddhartha Ojha, Thierno Diallo, Mihai Toma, Alec Smith, Cady Ngo, and Jiatong Han, Danny Ebanks, Chigozirim Ifebi, I am grateful for your collaboration, patience, and willingness to think through hard problems with me.

Thank you to a number of extraordinary mentees who helped with various research projects: Stephen Ebaseh-Onofa, Thierno Diallo, Cady Ngo, Chigozirim Ifebi; the passion and optimism you all brought to science was often the energy I needed to keep pushing through.

To the various Caltech staff: Liz Schneider, Mike Tyzka, Barbara Estrade, Laurel Auchampaugh, Diana Bohler, Sarah Sherwood; literally nothing would've gotten done without you. Both your logistics but also the moral support and friendly discussions you always shared with me.

To the other members of SDN, it was a joy going through grad school together. Brenden, Marcos, Sanghyun, Weilun, Aniek, Qianying, Wenning, Sneha, Seokyoung, Noah, Yi-Chuang, Anya, Amogh, and Ethan, I will miss these years with you

all, the jokes, the dinners, and the drinks. Hard to imagine a better crew to be in the "foxhole" with these past few year.

To Wenning and Isabel, thank you for being my ride or die.

To Jiatong; you taught me more about the world than you may ever know. Those years in the lab will never be forgotten.

To Katelyn and Paul, thank you for supporting my crazy ideas.

To Cam Jackson, thank you for teaching me to travel the world. You opened my eyes to the wonders of so many cultures and experiences and profoundly changed my life.

To my BSEC cabinet members and friends, thank you for being the support that I often desperately needed; Sarah, Evan, Jasmine, Ange, Susana, Jordan, Gilbert, Josh, Judah, and Brenna.

Additionally, I would like to thank my friends and family from the Caltech and LA community. When I moved here, I could never have imagined the richness of life that you'd bestow onto me. The time spent with you all has truly been the best days of my life.

Of course none of this work could be done without the generous funding contribution from the Chen Institute and the Linde Institute.

To my family: my Mom, Dad, Step-Mom, Step-Dad, Grandparents, and Sister, thank you for enduring the endless phone calls, rants, spirals, and never once withholding support. This PhD is as much a testament to your love as anything else.

To my younger brother Michael, your groundedness was an anchor for me to rely on, whether you realized it or not. Hope I made you proud.

ABSTRACT

Artificial intelligence is changing both the environments in which humans make decisions and the methods scientists use to study decisions. This dissertation investigates decision-making in complex, uncertain, and socially embedded settings by combining experimental economics, psychophysiology, machine learning, large language models, online text analysis, and neuroimaging. Across five chapters, I study how humans form beliefs, respond emotionally, communicate socially, and distinguish real from synthetic information, while also examining how artificial intelligence can serve as both a scientific tool and a comparison point for human behavior.

Chapter 2 provides an introduction through a research agenda for using AI agents in experimental social science. Large language model agents can act as interactive partners, synthetic participants, and scientific assistants, enabling new forms of piloting, hypothesis generation, cross-cultural simulation, and large-scale data analysis. Rather than replacing human subjects or scientists, the chapter argues that AI systems are most valuable when used as complements that expand the range, speed, and scale of social scientific inquiry.

Chapter 3 uses AI and ML techniques to study the psychobiological correlates of experimental asset bubbles. In experimental markets with a known constant fundamental value, participants traded a risky asset while providing forecasts, completing risk-elicitation tasks, and, for in-lab participants, wearing biometric sensors. Traditional market variables, especially lagged returns and participant forecasts; substantially predict future returns, period-by-period tonic electrodermal activity does not add incremental predictive power for one-period-ahead returns. We do find evidence that smoothed physiological arousal, more akin to "moods", is associated with longer-horizon returns and crash dynamics. Additional exploratory results find that lower arousal variability is associated with higher trading earnings, and elevated arousal predicts de-risking behavior. The chapter also shows that neither higher monetary stakes nor prior market experience eliminates bubbles, and that participant forecasts systematically violate rational-expectations benchmarks.

Chapter 4 compares out-of-the-box large language model agents in agent only markets with the performance of human traders in human only markets; within the same experimental finance paradigm described in Chapter 3. Unlike humans, LLM agents

generally price assets near fundamental value and display only muted propensity for bubble formation. This pattern persists across homogeneous single-model markets, heterogeneous "battle royale" markets, dividend shocks, and repeated-exposure treatments. Analyses of LLM-generated strategy text indicate lower variance, reduced bias, and stronger reliance on fundamentals relative to humans. These results suggest that out-of-the-box LLM agents may fail to reproduce core features of human-driven market behavior, especially large emergent bubbles.

Chapter 5 uses LLM-based text analysis to study social incentives in online financial communities. In this chapter we analyze posts from *r/wallstreetbets* in which users disclose realized trading gains or losses, the chapter shows that loss disclosures receive disproportionate engagement, especially when written in a self-deprecating tone. The magnitude of self-reported losses also increases over time, consistent with the possibility that platform-level engagement mechanisms reward and amplify negative financial outcomes. These results suggest that investors may derive non-pecuniary utility from publicly sharing losses, and that social rewards can shape financial communication and potentially risk-taking.

Chapter 6 examines whether the human brain encodes real and AI-generated faces differently. The results provide exploratory evidence that real and AI-generated faces evoke distinguishable neural response patterns in several visual, parietal, and frontal regions.

In the chapter appendix, we attempt to use these representation differences to test whether neural responses contain information about image authenticity beyond what is captured by overt behavioral judgments. We do this by using fMRI multivariate decoding methods. In some configurations, neural features modestly improve classification relative to behavioral measures alone. However, these effects are sensitive to analytic choices, and generalization across subjects remains limited. The chapter, therefore, provides proof of concept that neural activity contains information relevant to synthetic-media perception, while emphasizing the methodological challenges that remain before neuroimaging could contribute to practical deepfake detection.

Together, these studies demonstrate that human decision-making is shaped by the interaction of cognition, emotion, social incentives, and perceptual sensitivity to authenticity. They also show how artificial intelligence can be useful not only as a tool for measurement and analysis, but also as a subject for study in itself. By integrating behavioral science, machine learning, and neuroscience, this dissertation

advances a framework for studying decision-making in an increasingly AI-mediated world.

PUBLISHED CONTENT AND CONTRIBUTIONS

Henning, T. and C. F. Camerer (2025). "Synthetic AI Agents for Experimental Social Science". In: *Proceedings of AI for Accelerated Research Symposium* 3. T.H. developed the central idea and wrote the manuscript., pp. 77–84.

Henning, T., S. M. Ojha, et al. (2025). "LLM Agents Do Not Replicate Human Market Traders: Evidence from Experimental Finance". In: *Advances in Financial AI @ ICLR 2025*. T.H. developed the research question, designed and implemented the experimental market framework, analyzed the data, and led the writing of the manuscript. DOI: 10.48550/arXiv.2502.15800.

TABLE OF CONTENTS

Acknowledgements . . . iii
Abstract . . . vi
Published Content and Contributions . . . ix
Table of Contents . . . ix
List of Illustrations . . . xiii
List of Tables . . . xxv
Chapter I: Introduction . . . 1
1.1 Motivation . . . 1
1.2 Overview of Chapters . . . 2
1.3 Integrative Framework . . . 4
Chapter II: Synthetic AI agents for experimental social science . . . 6
Abstract . . . 7
2.1 Introduction . . . 8
2.2 AI-human interactions . . . 8
2.3 AI agents as substitutes for human subjects . . . 9
2.4 What AI-behavioral science lies ahead? . . . 11
2.5 AI tools for analyzing and generating social data . . . 12
2.6 Conclusions and future directions . . . 12
Chapter III: Psychobiological Correlates of Experimental Asset Bubbles . . . 17
Abstract . . . 18
3.1 Introduction . . . 18
3.2 Brief Literature Review . . . 20
3.3 Experimental Methods . . . 23
3.4 Price and forecast data . . . 27
3.5 Predicting Market Phenomena . . . 31
3.6 Associations of Arousal with Trading Behavior . . . 38
3.7 Models of Forecasting . . . 42
3.8 Conclusion . . . 53
3.9 Extended Literature Review . . . 58
3.10 High vs Low Earner Behavior in experimental markets . . . 60
3.11 Subject Differences In-Lab vs. Prolific . . . 63
3.12 Design Details . . . 66
3.13 Bio-metric measure details . . . 69
3.14 Experimental Procedure . . . 71
3.15 Individual Market and Forecast Plots . . . 72
3.16 Experienced Participants Selection Effects . . . 76
3.17 Evaluating individual forecast consistency . . . 80
3.18 Model Features . . . 82
3.19 Pre-registered Analysis . . . 86

3.20 Standard Session Regressions . . . 89
3.21 Mathematical Explanation of the ANN Model . . . 89
3.22 Mathematical Explanation of the SVR Model . . . 93
3.23 DOSE Preference Measures . . . 101
3.24 EDA^{ton} activity by Earnings Tercile . . . 102
3.25 Debriefing and Language Modeling . . . 105
3.26 Time Varying Risk Premia . . . 108
3.27 Physiological Arousal and Trading Behavior . . . 115
3.28 Summary Stats . . . 121
3.29 Forecast Rationality . . . 122
3.30 Instructions . . . 133
3.31 Quiz . . . 136
3.32 Surveys . . . 138
Chapter IV: LLM Agents Do Not Replicate Human Market Traders: Evidence from Experimental Finance . . . 156
Abstract . . . 157
4.1 Introduction . . . 158
4.2 Background . . . 159
4.3 Experiment . . . 161
4.4 Single-Model LLM Market Performance . . . 162
4.5 Battle Royale: Comparing LLM Market Performance . . . 163
4.6 Trading Strategy Analysis . . . 165
4.7 Market Treatments . . . 172
4.8 Rationality Properties of Individual LLM Forecasts . . . 174
4.9 Potential Limitations and Broader Impacts . . . 175
4.10 Discussion and Conclusions . . . 177
4.11 Ethics Statement . . . 178
4.12 Reproducibility Statement . . . 178
4.13 LLM Disclosure . . . 178
4.A Additional Experiment Details . . . 179
4.B Fundamental Value Calculation . . . 182
4.C Additional Treatments/Robustness . . . 183
4.D Additional Analysis . . . 187
4.E Rationality Properties of Individual LLM Forecasts . . . 192
4.6 Human Experiment Details . . . 201
4.7 Prompts . . . 207
Chapter V: Neural Responses during Detection of Real vs AI-generated Faces 225
5.1 Introduction . . . 226
5.2 Materials and Methods . . . 228
5.3 Results . . . 236
5.4 Conclusion . . . 253
5.A Decoding Methods . . . 256
5.B Decoding Results . . . 262
5.C PC Curve Plots, Within Subject Decoding . . . 276
Chapter VI: General Discussion . . . 287

6.1 Summary of Contributions 287
6.2 Limitations and Future Directions 290
6.3 Concluding Remarks 291

LIST OF ILLUSTRATIONS

Number *Page*

2.1 **AI agent vs. human market prices across 30 periods.** The six left panels (top and bottom rows) show price paths for six different LLMs prompted by human instructions, compared to the constant fundamental value (14, dotted line). The rightmost plot ("Mean human prices") shows data from human experiments averaged over 19 sessions. None of the LLMs reproduces the magnitude or trajectory of the human bubble–crash pattern, though several exhibit similar tendencies. This invites agent prompt optimization (APO) tuning of LLMs to better match human data (ongoing research). 10

3.1 (a) Average prices and forecasts across 19 Standard markets. The thick line shows the average price path. Each circular point on the graph represents the average current period forecast (just before the current price is announced) averaged across participants. The thin lines extending from the circular point are the "tongues" of average future forecasts in period t for each of the 2-, 5-, and 10-period ahead prices. Market-specific graphs can be found in section 3.15 in the Appendix. (b) All Standard markets are shown here as thin light gray lines. The average standard price over time is shown as the bold line. Fundamental value is shown as a dashed line at F=14. 28

3.2 Price path and volume of a typical market along with commonly used bubble metrics. **Rel Abs Dev** is a volume-weighted average of how far the price deviates from fundamental value, **Average Bias** is a simple average of price deviations. **Price Amplitude**, shows the difference between the market's highest and lowest prices. This gives an idea of the range of market prices observed. **Duration** is the longest number of periods in the market during which the price increases. 28

3.3 (a) Average prices and forecasts across four High-Stakes markets. The thick line shows the average price path. Each circular point on the graph represents the current period forecast (just before the current price is announced) averaged across participants. The thin lines extending from the circular point are the "tongue" of future forecasts in period t for each of the 2-, 5-, and 10-period ahead prices. Market-specific graphs can be found in section 3.15 in the Appendix. (b) All Experienced markets are shown here as thin light gray lines. The average standard price over time is shown as the bold line. Fundamental value is shown as a light dashed line at F=14. . . . 29

3.4 (a) Average prices and forecasts across five Experienced markets. The thick line shows the average price path. Each circular point on the graph represents the current period forecast (just before the current price is announced) averaged across participants. The thin lines extending from the circular point are the "tongue" of future forecasts in period t for each of the 2-, 5-, and 10-period ahead prices. Market-specific graphs can be found in section 3.15 in the Appendix. (b) All Experienced markets are shown here as thin light gray lines. The average standard price over time is shown as the bold line. Fundamental value is shown as a light dashed line at F=14. . . . 30

3.5 Mean out-of-sample SVR return predictions. Shaded regions show 95% confidence intervals from 1,000 bootstrap resamples of the training data, evaluated on a fixed held-out test set. The majority of realized returns lie within the bootstrap confidence bands across all held-out market groups, suggesting the model captures common market dynamics and provides well-calibrated out-of-sample uncertainty. For figure clarity, we choose not to plot the confidence interval of our held-out samples; for interested readers, this can be seen in Figure 3.29 in the Appendix. 36

3.6 **(a)** Tonic EDA by earnings tercile around the market peak. Panel A shows mean EDA_t^{ton} with 90% confidence intervals for High and Low tercile participants across rounds relative to the market peak ($\tau = 0$). The x-axis is cropped to only include τ bins for which there are 10 or more observations to avoid showing extremely large CI bands. The plot in panel (b) is cropped to match. **(b)** OLS quadratic regressions of mean tonic EDA (EDA_t^{ton}) on squared distance from the market peak (τ^2) with robust standard errors (Equation (3.2)), estimated separately for High and Low earning terciles. Points are group means; curves are fitted predictions from $EDA_t^{ton} = \alpha + \beta\tau + \gamma\tau^2$. Fitted curves were optimized and those optimal points are indicated by large orange triangular markers. Detailed values can be seen in Table 3.5 40

3.7 Mean forecast error by period for each of the four forecast horizons. As the forecast horizon increases, errors grow because projecting several periods ahead introduces greater uncertainty, causing small expectation errors to compound over longer intervals. 95 percent confidence intervals are shown by shaded regions. As time passes and the bubble begins to peak (and crash), the forecast error becomes more and more negative, indicating that participants are slow to update their price expectation. 45

3.8 Scatter plot of forecast rationality βs; the means of which are presented in Table 3.6. Figure shows that non-significance is distributed around the sample population instead of clustered around a few participants. The vertical axis is presented as a `symlog` scale, which is a logarithmic scale adapted to handle negative numbers and is linear in the light gray band between -0.1 and 0.1. This is why the trend line appears kinked. Joint significance counts can be found in Table 3.7 . 47

3.9 Histograms of Parameter Estimates for Different Expectation Models.

Notes: Different colors represent different levels of statistical significance. Light gray indicates that the estimate is not significantly different from zero ($p > 0.05$), while dark gray indicates that the estimate is significantly different from zero ($p < 0.05$). All models are estimated using OLS regression, with standard errors that are robust to heteroskedasticity. 49

3.10 Double Adaptive Model: Actual vs. Fitted Forecasts.
Notes: In the appendix, we check robustness and compare realized forecast levels to the model's fitted predictions under different representations: Figure 3.44 forecasts expressed as a percentage of lagged price; Figure 3.45 high- vs. low-earners; and Figure 3.46 pre- vs. post-crash periods. The dashed line indicates the 45-degree line where fitted values equal observed values. 51
3.11 Average shares held over 30 periods for participants grouped by market-earnings tercile (low, medium, high). The high-earning group accumulated shares early in the market and sold them off later. The lowest-earning group increases their holdings late in the market, particularly after the average peak period. 61
3.12 Mean forecast error for the highest and lowest earning terciles. Shaded regions represent 95 percent confidence intervals. 62
3.13 Plot of mean bonuses by site. Standard markets only. 95 percent confidence intervals are shown. 64
3.14 Distribution of the difference in means of the market bonus by site (μ = \$0.099; σ = 0.226; df=19,999). Market bonuses were shuffled within each market. We then calculated the difference in the means of market bonuses by site. The observed difference of \$0.763 is shown by a red dashed line. The p-value of this comparison was calculated by calculating $1 - \text{cdf}(\mu, \sigma, \text{df}) = 0.002$ (labeled Student's t). We also computed the p-value by calculating the proportion of the bootstrapped observations that are greater than the observed difference (labeled "observed"). 66
3.15 Order submission page (20 seconds). Participants are asked to submit the price and volume for their limit orders using the grid. 66
3.16 Forecast page. (30 seconds) Participants are asked to submit their estimation for the current market period and periods 2, 5, and 10 ahead. 67
3.17 Round Results page. (10 seconds) Participants see the actual realized price for that period, the total number of shares traded, and the number of shares they traded. 67

3.18 Risk Elicitation Page. (20 seconds). Participants are shown four pairs of a risky lottery, visually displayed as a pie chart, and a less risky lottery, and asked to choose between them. After each selection, the probabilities involved are adjusted according to the DOSE protocol (Chapman et al., 2024) so that the next choice offers the most information about the participants' risk preferences. 68

3.19 A comparison of market measures by market type. *Note.* 95 percent confidence intervals shown. Because there are only four High-Stakes markets and five Experienced, the tests are too weakly powered to detect significant differences. Notice the standard error bars are larger for High-Stakes and Experienced statistics compared to the standard. Attempted (Executed) Trades refer to the average number of trades attempted (or executed) per participant in each market. 72

3.21 19 Standard markets. The lines represent the dynamics of price levels for 30 periods. Each point represents the current period forecast, and subsequent lines represent average price forecasts for the 0-, 2-, 5-, and 10-period ahead periods. Note: There was an error in Market (2024-04-04); participants were only endowed with an average of two shares of stock rather than a set number of four. The qualitative performance was similar to that of our markets; therefore, we combined it into our analysis. 74

3.22 Four High-Stakes markets. The lines represent the dynamics of price levels for 30 periods. Each point represents the current period forecast, and subsequent lines represent average price forecasts for the 0-, 2-, 5-, and 10-period ahead periods. 75

3.23 Five Experienced markets. The lines represent the dynamics of price levels for 30 periods. Each point represents the current period forecast, and subsequent lines represent average price forecasts for the 0-, 2-, 5-, and 10-period ahead periods. 76

3.24 Market Bonus Comparison One-timers vs. Repeaters (Bootstrap N=10,000) The mean market bonus for one-time *prolific.com* participants (μ_o = \$3.09) was compared to the first market bonus earned by repeat participants (μ_r = \$3.47). Market bonuses were shuffled within each market. We then calculated the difference in the means of market bonuses, grouped by one-timers and repeaters. We repeated this 10,000 times to generate a distribution (μ = \$0.18; σ = 0.26; df = 9,999). The observed difference of \$0.39 is shown by a red dashed line. The p-value of this comparison was calculated by calculating $1 - \text{cdf}(\mu, \sigma, \text{df}) = 0.206$. We also computed the p-value by calculating the proportion of the bootstrapped observations that are greater than the observed difference (p=0.210). 80

3.25 Plot of individual regression coefficients for net orders on an indicator for forecasting returns above the 5% risk-free rate. Each point is a participant-level coefficient from a regression of net orders on an indicator equal to one when the participant's forecasted return exceeds the 5% risk-free rate. Stars represent participants with positive coefficients, while triangles represent participants with negative coefficients. The coefficient measures how much net buying increases (or decreases) in high–expected-return states relative to low- or sub-risk-free-return states. More positive coefficients indicate a higher degree of internal consistency between beliefs and trading behavior. . 82

3.26 Scree plot showing the explained variation of each principle component of the three forecast measures used the main regression (Table 3.2). Forecast variables are f_t^t, f_{t-2}^t, f_{t-5}^t. Exact explained percentage of each component can be found beside the data point. 89

3.27 Correlational matrix of regressor in future returns predictions OLS (Table 3.2). The image shows a high level of correlation among the three forecast measures; indicating the possibility of multicollinearity. The first principle component of the three forecasts are included to avoid this in the regressions provided in the main text. 92

3.28 **SVR with EDA^{ton} Model Predictions vs Actual Returns.** The SVR mean prediction curve represents the model's out-of-sample forecasts averaged across all (four; [market: 1, 2, 6, 12]) held-out test markets for each trading period. Each SVR was trained on the remaining markets (15), cross-validated on training folds for tuning, and then evaluated once on the four held-out markets. Confidence intervals were generated using a nonparametric bootstrap applied to the training data: for each of 1000 iterations, the training data were resampled with replacement, a new SVR was fit to the bootstrapped sample, and its performance was evaluated on the fixed held-out test markets. 96

3.29 Mean out-of-sample SVR return predictions. Shaded purple regions show 95% confidence intervals from 1,000 bootstrap resamples of the training data, evaluated on a fixed held-out test set. Shaded gray regions show 95% confidence intervals from the standard deviation of the held-out test data. The majority of realized returns lie within the bootstrap confidence bands across all held-out market groups, suggesting the model captures common market dynamics and provides well-calibrated out-of-sample uncertainty. 98

3.30 Example DOSE decision tree for generating estimates of ρ. 103

3.31 Average EDA^{ton} over 30 periods for participants grouped by market-earnings tercile (low, medium, high). 90% confidence intervals shown. 104

3.32 Average EDA^{ton} over 30 periods for participants grouped by market-earnings median (low, high). 95% confidence intervals shown. 104

3.33 Mean Returns by EDA^{ton} median split. EDA^{ton} activity predicts future returns of the risky asset. Periods within each market were split into high and low activity based on a five-period moving average of EDA^{ton} activity. The bars show that the mean five-period forward return is higher for above-median EDA^{ton} activity. 106

3.34 **Word cloud of trading strategy** Larger words indicate higher use frequency in described strategies. 107

3.35 Strategy Group Assignments by Earning Median. Counts of strategy group assignment by earnings median. The count of all strategies assigned to a group is shown above the bar clusters. The strategy group `No Strategy` consists of those who responded with “No", “None", or the like. `No / Trend` mostly stated that they did not follow a strategy, but also indicated that they followed the market trend or guessed. `Buy Low / Sell High` are those that express that classic strategy. `Low-Risk` strategies focused on accruing interest and selling shares as soon as possible. `Buy Early` strategies involve acquiring the stock at an early stage with the intention of selling it later. 108

3.36 PCA of sentence embeddings with K-Means clustering. After vectorizing all strategies, we ran a k-means clustering algorithm and used principal component analysis (PCA) to reduce the vector space to two dimensions. The first principal component accounts for 13.7 percent of the variance, and the second component accounts for 7.1 percent. . 109

3.37 Log Return R_t vs ρ Across Periods. ρ refers to the risk-aversion coefficient estimated for each subject with the risk elicitation task (higher values indicate greater risk aversion). The parameter is estimated using the procedure described in (Chapman et al., 2024). There is no statistically significant relationship between ρ and R_t, suggesting that time-varying risk premias are not the drivers of bubble-like returns. 110

3.38 Log Return R_t vs μ Across Periods. μ refers to the choice consistency coefficient estimated for each subject during the risk elicitation task (higher values indicate greater consistency of choice). The parameter is estimated using the procedure described in (Chapman et al., 2024). 111

3.39 Histograms of forecast errors for current round and 2 periods ahead. 122

3.40 Distribution of Forecast Error Grouped by Period 124

3.41 Median forecast error by period for each horizon. Forecast error is defined as $E^h_{i,t} = P_{t+h} - \tilde{f}^{t+h}_t$ similar to Eq 3.3. Here $\tilde{f}$ indicates the median forecast of all participants in the period t for horizon h. A positive forecast error indicates an underestimate while a negative error indicates an overestimate. 125

3.42 Histograms of Parameter Estimates for Different Expectation Models in the Experienced Markets
Notes: Different colors represent different levels of statistical significance. Light gray indicates that the estimate is not significantly different from zero ($p > 0.05$), while dark gray indicates that the estimate is significantly different from zero ($p < 0.05$). All models are estimated using OLS regression, with standard errors that are robust to heteroskedasticity. 127
3.43 Histograms of Parameter Estimates for Different Expectation Models in High-Stakes Markets
Notes: Different colors represent different levels of statistical significance. Light gray indicates that the estimate is not significantly different from zero ($p > 0.05$), while dark gray indicates that the estimate is significantly different from zero ($p < 0.05$). All models are estimated using OLS regression, with standard errors that are robust to heteroskedasticity. 128
3.44 Double Adaptive Model: Actual vs. Fitted Forecasts: Forecasts expressed as a percentage of lagged price 129
3.45 Double Adaptive Model: Actual vs. Fitted Forecasts: High vs Low Earners 129
3.46 Double Adaptive Model: Actual vs. Fitted Forecasts: Pre- vs. Post-crash periods 130
4.1 Mono-Agent Price Charts compared to Humans 164
4.2 Battle Royale Market Performance with Human Comparison 165
4.3 Word cloud representation of the most commonly used words in the strategies for human and LLM agents. 167
4.4 Percentage of human and LLM strategies assigned to one of two topics. 169
4.5 Distributions of sentiment for human and LLM agents 170
4.6 Sentiment scores are plotted against the mispricing of the market compared to the fundamental value of \$14 171
4.7 Mono-Agent Market Price Charts Dividend Shock Doubling 173
4.8 Mono-Agent Market Price Charts Dividend Shock Halving 174
4.9 Mean Forecast Error by Round 175
4.10 Order Submission Page 179
4.11 Forecast Page 180
4.12 Round Results Page 180

4.13 Experienced Market Trajectories for different LLMs. 184
4.14 Risk-Elicitation Priming Market Trajectories for different LLMs. . . . 186
4.15 Human Market Trajectories 187
4.16 Gemini 1.5 Pro Markets 188
4.17 Gemini 1.5 Pro Portfolio Variance 188
4.18 Mistral-Large 24.11 Portfolio Markets 189
4.19 Mistral-Large 24.11 Portfolio Variance 189
4.20 Claude-3.5-Sonnet Markets 190
4.21 Claude-3.5-Sonnet Portfolio Variance 190
4.22 GPT-3.5 Markets 191
4.23 GPT-3.5 Portfolio Variance 191
4.24 GPT-4o Markets 192
4.25 GPT-4o Portfolio Variance 192
4.26 Comparison of forecast errors across rounds for human participants and AI models 194
4.27 Additional Forecast Error Distribution Graph 198
4.28 The percentage of no autocorrelation bias of each model at each horizon 199
4.29 The results of the autocorrelation test. 200
4.30 Results of the correlation between forecast error and forecast 201
5.1 A) Prompt → Image Prompt: "Frontal view of a real young dominican woman with long brown hair, wearing a formal suit, sitting inside an interviewing office" B) Real Image → Prompt → Image Prompt: smiling asian woman sitting at desk and writing on notebook, in the style of light teal and black, uniformly staged images, representational, contact printing, nonrepresentational. C) Real Image D) Real Image → Prompt → Image Prompt: a young man with long hair is standing by a fence, in the style of afro-caribbean influence, light indigo and dark amber, strong facial expression, organic formations, dinocore, quadratura, confessional. 229
5.2 Sample task screen for the online behavioral task. 231
5.3 Accuracy and reaction time distributions for the images shown in the online task, shown separately for real and AI-generated stimuli. . . . 231
5.4 Examples from the fMRI stimulus set. (A) Hard AI image, (B) Hard real image, (C) Easy AI image, and (D) Easy real image. 232
5.5 fMRI Task. 233

5.6 Online vs Scanner Behavioral Accuracy image comparison, online vs scanner. The variance is wider across scanner subjects because the online images were directly selected to fall within a set accuracy range. 238
5.7 Online vs Scanner reaction time image comparison, online vs scanner. The variance is wider across scanner subjects because the online images were directly selected to fall within a set accuracy range. . . . 239
5.8 Accuracy and reaction time distributions for the images shown in the fMRI task, shown separately for real and AI-generated stimuli. 239
5.9 Cluster masks from GLM 1, thresholded at voxelwise $p < 0.005$ with cluster extent $k > 50$ voxels. Panels labeled C1–C6 correspond to cluster IDs in Table 5.3. 243
5.10 Cluster masks from GLM 2, thresholded at voxelwise $p < 0.005$ with cluster extent $k > 50$ voxels. Panels labeled C1–C9 correspond to cluster IDs in Table 5.4. 248
5.11 Cluster masks from GLM 3, thresholded at voxelwise $p < 0.005$ with cluster extent $k > 50$ voxels. Panels labeled C1–C9 correspond to cluster IDs in Table 5.5. 252
5.12 Neuro-only Data Classification AUC scores by number of Principal Components. **Inset:** Explained variance by Principal Component. . . 263
5.13 Neuro+Behavior Data Classification AUC scores by number of Principal Components. **Inset:** Explained variance by Principal Component. 264
5.14 Coronal view of masks generated from GLM 1 on a) Easy Trials, b) Hard trials, c) Overlap of Easy Mask and Hard Mask. 265
5.15 Per-bin decoding performance (AUC) compared to behavioral accuracy. Bold indicates Model AUC higher than behavioral accuracy. . . 266
5.16 Overlap heatmap of regions predictive in multiple decoding robustness checks. The cluster in the Precuneus/Retrosplenial area was consistently implicated across permutations (8/10) | (8/9) of the robustness checks that didn't fail. 268
5.17 Group-level average peak decoding accuracy map from whole-brain searchlight analysis. Only voxels exceeding 58% decoding accuracy in at least 4 subjects and belonging to clusters > 40 voxels are shown. 270

5.18 Schematic of the latent-variable rating task. Each participant rated three randomly assigned attributes (on a 0–100 scale) for all 200 images. Mean attribute scores served as inputs to subsequent behavioral and fMRI analyses. 272
5.19 AUC scores by Principal Components, comparison between Neuro-Only, Neuro+Behavior, and Behavioral Only Models. 273
5.20 AUC scores by Explained Variance, horizontal reference line indicates behavioral score. 273
5.21 Neuro + Behavior AUC scores by Variance Explained, horizontal reference line indicates behavioral score. 274
5.22 Training AUC scores by Principal Components, horizontal reference line indicates behavioral score; represents the AUC in the training folds. 275
5.23 PC decoding curves for the 50 Mask / 50 Train within-subject split. Each panel shows the decoding performance for each mask as principal component increases. Horizontal line represents average behavior accuracy across subjects. 277
5.24 PC decoding curves for the 30 Mask / 70 train within-subject split. Each panel shows the decoding performance for each mask as principal component increases. Horizontal line represents average behavior accuracy across subjects. 278
5.25 PC decoding curves for the 50 Mask / 50 Train within-subject split. Each panel shows the decoding performance for each mask as principal component increases. Each mask iteration is combined with behavioral data to show the combined predictive power of Neuro and Behavior. Horizontal line represents average behavior accuracy across subjects. 279
5.26 PC decoding curves for the 30 Mask / 70 Train within-subject split. Each panel shows decoding performance across principal components for each mask using combined neural and behavioral features. Each mask iteration is combined with behavioral data to show the combined predictive power of Neuro and Behavior. Horizontal line represents average behavior accuracy across subjects. 280

LIST OF TABLES

Number *Page*

3.1 Summary of Regression Notation 32
3.2 Predicting Next-Period Log Returns R_t (returns $\times 100$) 33
3.3 Average SVR model performance across held-out market validation groups. Mean values are computed by averaging the scores across all held-out samples. Ranges report the minimum and maximum score observed across the held-out market groups. 37
3.4 SVR feature importance weights. Mean importance weights are reported across held-out market validation groups. Ranges indicate the minimum and maximum importance weight observed across held-out groups. 37
3.5 Quadratic Regressions of Tonic EDA Around the Market Peak 41
3.6 Normality and Rationality Acceptance Rates (%) Across Horizons . . 45
3.7 Counts of participants by joint significance of the zero auto-correlation β (AC, rows) and the zero error-forecast correlation β (EF, columns). 47
3.8 Summary Statistics of Parameter Estimates and Fit by Forecasting Model 51
3.9 Expectation Model Parameters Across Markets 52
3.10 Market & Forecast Bonus Explained by Forecast Parameters and Controls 54
3.11 High-Stakes and Experienced markets: Predicting Future Returns R_t (returns $\times 100$) 57
3.12 Forecast Error Comparison 62
3.13 Summary statistics comparing survey responses from participants based on site. 63
3.14 Summary statistics comparing market performance measures from participants based on site (Lab L or Prolific P. 64
3.15 Mean bonuses by group with p-values and 95 percent confidence intervals. 65
3.16 Comparing means of measures of participants who repeated the experiment with the measures of their peers during their first market session. 78
3.17 Mean trades by Experienced Markets. 79

3.18 Predicting Next-Period Log Returns R_t with Backward-Looking Forecasts (returns ×100) . . . 86
3.19 Predicting Next-Period Log Returns R_t with High-Stakes Sessions (returns ×100) . . . 87
3.20 Predicting Next-Period Log Returns R_t in Experienced Sessions (returns ×100) . . . 88
3.21 Predicting Next-Period Log Returns R_t with Backward- vs Forward-Looking Forecasts (returns ×100) . . . 90
3.22 Lagged Returns (x100) and Forecast Measures in Standard Sessions . 91
3.23 Support Vector Regression (SVR) w/o EDA Feature Importance (Linear Kernel) . . . 96
3.24 Performance Metrics for Selected Models (with and without EDA^{ton}) 97
3.25 Importance values are normalized absolute coefficients from the linear SVR model (sum to 1). Higher values indicate a stronger influence on predicted returns. Using the same random split as above held out markets: [1,2,6,12]. . . . 97
3.26 Predicting next-period returns (R_t) using LASSO. . . . 100
3.27 Chow Test for Differences in Quadratic EDA Profiles by Earnings Tercile . . . 100
3.28 Final model performance metrics. Other Machine Learning Models . 101
3.29 Ranges of market earnings by tercile. . . . 103
3.30 EDA Peak is correlated, at the individual-trial level, with risk-aversion ρ 111
3.31 EDA Derisk Logistic Regressions . . . 116
3.32 Summary statistics. . . . 119
3.33 Individual Executed Trades and Bonuses for Experienced Participants 120
3.34 Market level summary statistics. . . . 121
3.35 Summary Statistics of Forecasts and Forecast Errors . . . 122
3.36 Summary Statistics of Parameter Estimates and Fit by Forecasting Model for Experienced Markets . . . 123
3.37 Summary Statistics of Parameter Estimates and Fit by Forecasting Model for High-Stakes Markets . . . 123
3.38 OLS and Fixed-Effects Regression Results: Market Bonus vs. Forecast Bonus . . . 126
3.39 OLS and Fixed-Effects Regression Results: Market Bonus vs. Forecast Bonus - Experienced Markets . . . 131

3.40 OLS and Fixed-Effects Regression Results: Market Bonus vs. Forecast Bonus - High-Stakes Markets . . . 132
4.1 Market performance metrics . . . 163
4.2 Battle Royale Agent Performance . . . 164
4.3 Most popular words used by human participants and by the LLM Agents . . . 167
4.4 Distribution (%) of speculative, fundamental, and generic text in agent strategy statements. The behavior column reflects observed market outcomes. . . . 168
4.5 Language–based strategy fingerprints (plans+insights). . . . 169
4.6 Fixed effects regression of sentiment score on round number and level of mispricing. Standard errors are clustered on LLM and shown in parentheses. Sentiment was multiplied by 100 to reduce the significant digits in the estimates. . . . 171
4.7 Forecast Accuracy and Rationality Properties of Humans and LLMs. 176
4.8 Approximate token usage across models. . . . 182
4.9 Rationality test results . . . 195
4.10 Descriptive statistics for Price and Forecasts by source. . . . 197
4.12 Descriptive statistics and forecast errors for each model and humans. . 202
5.1 Behavioral Accuracy for Real and AI conditions. . . . 237
5.2 Reaction times (ms) by condition. . . . 238
5.3 Significant clusters from the Real vs. AI contrast (GLM 1). Threshold: p-value < 0.005, $k > 50$ voxels. . . . 243
5.4 Significant clusters from the Perceived Real vs. Perceived AI contrast (GLM 2). Threshold: p-value < 0.005, $k > 50$ voxels. . . . 247
5.5 Significant clusters from the Correct vs. Incorrect contrast. Threshold: p-value < 0.005, $k > 50$ voxels. . . . 252
5.6 Cluster tables for Easy, Hard, and Overlap. Easy/Hard entries show peak coordinates grouped by cluster (thresholded at $p < 0.005$, $k > 50$ voxels). Overlap shows the intersection computed voxelwise ($|t| > 3$) with component size. . . . 266
5.7 SVC ROC across 50 Mask/50 Train, Neuro-Data . . . 267
5.8 Overlap clusters with peak MNI coordinates, peak overlap counts, and Harvard–Oxford ROI labels. . . . 268
5.9 SVC classification performance across iterations. Accuracy remains at chance level with no significant decoding. . . . 269

5.10 Significant clusters from whole-brain searchlight decoding (Threshold: 58%, Cluster size > 40 voxels) 270
5.11 Significant clusters from Neuro + Behavior searchlight decoding (Threshold: 58%, Cluster size > 40 voxels) 270
5.12 Latent perceptual attributes adopted from Miller et al., 2023 272
5.13 Per-bin decoding performance compared to behavioral accuracy. . . . 276

Chapter 1

INTRODUCTION

How do humans make decisions when outcomes are uncertain, information is incomplete, and the environment is shaped by the choices of others? This question sits at the intersection of economics, psychology, and neuroscience, and it has taken on new urgency as artificial intelligence systems increasingly participate in, and transform, the settings in which human decisions unfold. This dissertation investigates decision-making in complex environments through the lens of behavioral science, combining experimental economics, psychophysiology, computational modeling, neuroimaging, and large-scale text analysis. Across four studies, it examines how cognition, emotion, and social incentives interact to shape choices in financial markets, online communities, and perceptual tasks, and how artificial intelligence both diverges from and illuminates human behavior.

1.1 Motivation

Two developments motivate this work. The first is the persistent gap between canonical models of rational decision-making and observed human behavior. Despite decades of theoretical refinement, models based on expected utility maximization and rational expectations fail to account for many robust empirical phenomena: speculative bubbles that inflate and crash despite transparent fundamental values, systematic forecast biases that worsen at longer horizons, and risk-taking behavior shaped by social rewards rather than expected payoffs. These anomalies are not merely curiosities; they have consequences for asset prices, financial stability, and welfare. Understanding their origins requires tools that go beyond revealed preference, including physiological measurement, belief elicitation, and analysis of the social contexts in which decisions are embedded.

The second development is the rapid rise of large language models and generative AI. LLMs are increasingly proposed as substitutes for human participants in experiments, as autonomous agents in economic environments, and as analytical tools for processing social data at scale. Yet the extent to which these systems replicate, approximate, or fundamentally diverge from human decision-making remains poorly understood. If AI agents can reproduce human market behavior, they could accelerate experimental research and stress-test economic theories. If they can-

not, understanding why not reveals something important about what makes human decision-making distinctive. This dissertation contributes to both questions.

1.2 Overview of Chapters

The dissertation is organized around five self-contained studies, each addressing a different facet of decision-making in uncertain environments. While each chapter makes independent contributions, they are connected by a common interest in the interplay between human cognition and artificial intelligence, and by shared experimental infrastructure, particularly the experimental asset market platform used in Chapters 3 and 4.

Chapter 2: Synthetic AI Agents for Experimental Social Science. AI will transform social sciences. LLM agents can converse, act autonomously, and arguably reason. This position paper provides three complementary AI–human research uses. First, as interactive partners, AI agents let scholars learn more about human–AI social interactions. Second, finely-tuned AI models can substitute for human subjects, allowing rapid piloting and hypothesis generation, cross-cultural "synthetic cultural agents", and "out-of-this-world" (OOTW) speculations about experiments that can't be conducted with humans. Third, AI agents can co-science by coding and summarizing data at scales much faster than humans with minimal error (in some cases). This position paper outline elements of a research agenda in which AI complements human subjects and scientists.

Chapter 3: Psychobiological Correlates of Experimental Asset Bubbles. This chapter is the empirical centerpiece of the dissertation. It reports results from 28 experimental markets (19 standard, 4 high-stakes, and 5 experienced) in which 475 participants traded a risky asset with a known constant fundamental value while physiological arousal was recorded via wearable biosensors. The study makes several contributions. First, it shows that standard financial variables, particularly lagged returns and near-term forecasts, predict next-period returns with substantial accuracy (out-of-sample $R^2 \approx 0.67$), while period-by-period measures of tonic electrodermal activity do not add incremental predictive power. Second, it demonstrates that smoothed arousal, capturing "market mood" rather than momentary emotion, is significantly associated with future returns and crash probability. Third, it documents that lower arousal variability is associated with higher earnings, and that elevated arousal predicts de-risking behavior, suggesting that physiological sig-

nals operate through behavioral channels rather than direct price effects. Fourth, it applies three classical tests of forecast rationality and finds systematic violations that grow with horizon, accompanied by structured heterogeneity in expectation formation across adaptive, extrapolative, and double-adaptive models. Fifth, it finds that neither higher monetary stakes nor prior market experience eliminate bubbles, though experienced participants peak earlier. Trial-by-trial risk preferences are not associated with returns, ruling out time-varying risk premia as a driver of mispricing in this setting.

Chapter 4: LLM Agents Do Not Replicate Human Market Traders. In this study, we observe Large Language Models (LLMs) trading in a similar experimental-finance paradigm to Chapter 3, where prices are determined endogenously through agent interaction alone. Using this well-established asset-trading design, we run homogeneous (mono-agent) markets with single-model LLM agents and heterogeneous "Battle Royale" markets with multiple LLM models. Our findings reveal that LLMs generally exhibit a "textbook-rational" approach, pricing the asset near its fundamental value and showing only a muted tendency toward bubble formation. Additional treatments, including dividend shocks and repeated-exposure ("experienced") runs, show that these dynamics persist across experimental settings. Analyses of LLM-generated strategy text indicate low variance, limited bias, and strong reliance on fundamentals. These results characterize how autonomous LLM agents behave in endogenous markets, an increasingly important question as such agents are deployed in real economic environments, and indicate that key behavioral features of speculative markets, such as large emergent bubbles, do not arise spontaneously from agent interaction alone.

Chapter 5: Neural Responses during Detection of Real vs AI-generated Faces. The final empirical chapter examines whether neural responses distinguish real from AI-generated face images during a classification task. The chapter centers on how image origin, perceived realness, and classification accuracy are encoded in whole-brain BOLD responses. These analyses show exploratory evidence that objective image origin is associated with measurable but anatomically distributed differences in neural activity: real faces elicit stronger responses in early visual cortex, whereas AI-generated faces elicit stronger responses in regions including lingual gyrus, precuneus, cingulate cortex, and frontal pole. A perceived-realness contrast partially converges with this pattern, particularly in early visual cortex

and frontal pole, suggesting that subjective judgments track some of the same stimulus properties that distinguish real from AI-generated images. In contrast, the correct-versus-incorrect contrast produces a much broader pattern, more consistent with task difficulty, uncertainty, effort, or response-related processing than with a focal neural signature of successful AI detection. Multivariate decoding analyses provide a secondary test of whether these distributed responses contain image-level information about authenticity, but the decoding results are modest, analytically sensitive, and unreliable across participant splits. Overall, the chapter provides cautious evidence that neural responses carry information relevant to synthetic face perception, while emphasizing that this information is weak, distributed, and far from sufficient for practical neuroimaging-based deepfake detection.

1.3 Integrative Framework

While the chapters span distinct empirical domains- laboratory asset markets, LLM agent behavior, online social platforms, and neuroimaging, they share several intellectual commitments that lend coherence to the dissertation as a whole.

First, each chapter takes seriously the idea that decision-making is shaped by factors that standard economic models either abstract away or cannot measure. Physiological arousal, belief heterogeneity, social norms, and perceptual sensitivity to authenticity all play documented roles in the decisions studied here, and none is captured by models that assume a representative rational agent.

Second, the dissertation uses artificial intelligence as both a tool and a foil. LLMs serve as as experimental participants in Chapter 4, and as generators of the stimuli whose neural perception is studied in Chapter 5. In each case, the comparison between human and AI behavior is informative: AI agents are more rational but less interesting as market participants; AI-generated faces are increasingly realistic but neurally distinguishable; AI-based text analysis reveals social structures that traditional methods cannot measure at scale.‘

Third, the work is methodologically pluralist by design. It combines incentivized laboratory experiments, wearable biosensors, machine learning, natural language processing, fMRI neuroimaging, and large-scale observational data analysis. This breadth reflects a conviction that complex questions about human decision-making are best addressed by converging evidence from multiple methods, each with different strengths and limitations, rather than by any single paradigm in isolation.

The General Discussion Chapter 6 synthesizes findings across chapters, identifies

cross-cutting themes, and discusses limitations and directions for future research.

Chapter 2

SYNTHETIC AI AGENTS FOR EXPERIMENTAL SOCIAL SCIENCE

Thomas Henning and Colin F. Camerer

ABSTRACT

AI will transform social sciences. LLM agents can converse, act autonomously, and arguably reason. This position paper provides three AI-human complementary research uses. First, as interactive partners, AI agents let scholars learn more about human–AI social. Second, finely-tuned AI models can substitutes for human subjects, allowing rapid piloting and hypothesis generation, cross-cultural "synthetic cultural agents", and "out-of-this-world" (OOTW) speculations about experiments that can't be conducted with humans. Third, AI agents can co-science by coding and summarizing data at scales much faster than humans with minimal error (in some cases). This Position paper outline elements of a research agenda in which AI complements human subjects and scientists.

2.1 Introduction

We are in a golden age of social science (Buyalskaya, Gallo, and Camerer, 2021). This golden age is created by a torrent of cheaper and better data, new open science sharing norms, and computational power. AI tools will accelerate this boom.

While social science will be accelerated by AI, social science can repay the favor by economic analysis of the impact of AI on the workplace, inferring 'preferences' that implicitly motivate AI agents, and using causal inference to understand how AI changes make a difference (Jackson et al., 2025). Theory can also be used to understand mathematical limits to value creation in human-AI collaboration (Peng, Garg, and Kleinberg, 2025).

We focus on three applications: (1) Mixed human–AI interactions; (2) using AI agents as substitutes for human experiments; and (3) using AI agents to analyze qualitative social data at scale and generate new quantitative data. The last two are examples of "AI co-science"– using AI to improve the quality and pace of scientific discovery and knowledge accumulation.

Our goal is to educate readers about how AI agents can complement social science methods and accelerate progress. Anthis et al. (2025) also argue for this position (and include a list of 53 studies in AI-Behavioral Science).

2.2 AI-human interactions

People already interact with AI in many ways: We watch movies because they were recommended by a Netflix algorithm, use LLMs to improve essay writing, talk to customer service chatbots, and increasingly face "personalized pricing" (called "first-degree price discrimination" in economics). Evidence is accumulating about what people like and dislike about such AI interactions. People dislike bargaining with AI (Erlei et al., 2022), cheat more often when reporting results to a computer (Treiman, Ho, and Kool, 2023), feel less guilty about exploiting AI counterparts (Karpus et al., 2021), and underreport how much they use AI (Ling and Imas, 2025).

The study of human-computer interaction (HCI) is a powerful way for social scientists to probe these phenomena. AI agents can be deployed as highly controlled "confederates" in experiments, allowing researchers to systematically vary a counterpart's attributes (e.g., presenting it as a human or an AI, or giving it an approachable persona vs. an abstract one) limiting perceptual noise (Maehigashi, Tsumura, and Yamada, 2022). For example, one could test how people's willingness to trust a teammate changes based on its personality traits, while enabling them to actually

"interact" with that teammate in a relatively unstructured way (helping to hide the fact that it isn't a human).

Recent reviews highlight several factors that can be manipulated in human–AI interaction studies, including emotional framing and cultural cues (Horton, 2023; Baskaran et al., 2023; S. P. Kim, Lee, and S. Y. Park, 2023). By leveraging AI agents with controlled behaviors, researchers can isolate human biases and preferences in mixed human–AI settings with greater precision than ever before. This approach will improve AI design (by creating agents that humans find more trustworthy or fair) and advance social science by revealing how we treat non-human actors in social and economic exchanges.

Another frontier for HCI is using AI in clinical practices such as diagnosis and mental health therapy. The FDA has currently approved 1247 AI devices; that number will grow rapidly. Meta-analysis of 83 studies found that AI diagnoses were as accurate as most physicians, but worse than the most expert Takita et al., 2025.

2.3 AI agents as substitutes for human subjects

A dreamy use case for AI-Behavioral Science is to create artificial agents who, properly trained on human data, can act sufficiently human-like to predict accurately in never-seen OOD tasks. This would enable rapid piloting and hypothesizing, in which human subject experiments come at the end.

There are many successful examples. Hewitt et al. (2023) showed that LLM (GPT-4) responses, to social science surveys of representative US human samples in 70 experiments, matched previous human data with correlation r=.91. Binz et al. (2025) introduced a Centaur model that reproduced many findings about human cognition. J. S. Park et al. (2024) created an LLM 1000-agent sample that matched the results of the General Social Survey with 85% precision.

Despite these apparent successes, there is intense debate about the current and future value of AI subjects. An example of this useful debate began with Dillion et al. (2023) showing that LLMs could reproduce human moral judgments. Then Schröder et al. (2025) showed that in perturbations of those scenarios, LLMs were more sensitive to small prompt variations than humans. They pessimistically concluded that "...LLMs [are] useful but fundamentally unreliable tools."

Our Position is that the pessimism expressed by Schröder et al., 2025 is a myopic mistake. The cycle of improvement after AI flaws are evident has been incredibly fast (months or years, not decades). AI images no longer show wonky fingers,

debiasing for fairness is a hot topic, blackbox interpretability is being improved by xAI, and LLM hallucination and sycophancy are being conquered. The checklist of AI flaws which led Schröder et al. (2025) to call methods “unreliable" is actually a to-do list for active researchers, who are eager to improve models quickly and will do so.

Describing a specific AI co-science experimental design will illustrate some of our ideas. Our group put humans in a simplified experimental financial market. In the market they trade assets that pay a financial dividend in each of T periods, or can invest experimental cash into a risk-free bond. The market is designed so there is a constant fundamental ("intrinsic") value. In these markets, humans often trade at accelerating prices in a “bubble" (about 3x the fundamental), which crashes toward a terminal period T.

Henning et al. (2025) created AI agent-based trading agents in the same trading environment (with no steered prompting) and compared them with human subjects. Unlike human traders, AI agents priced assets close to their fundamental value (Figure 2.1). These findings suggest that these AI agents behave more like textbook rational agents than humans do.

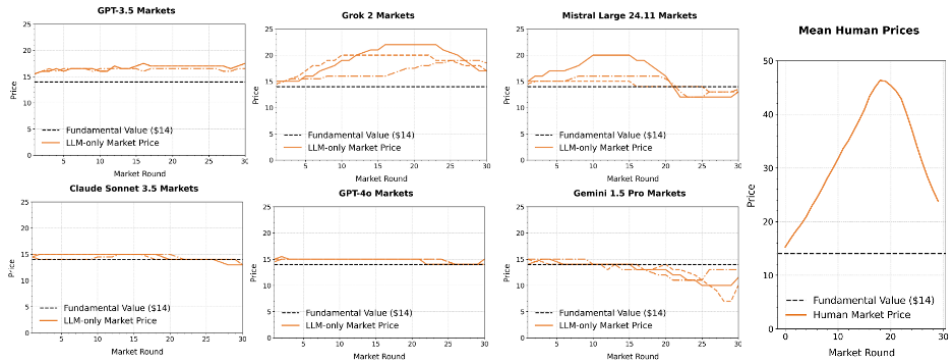


Figure 2.1: **AI agent vs. human market prices across 30 periods.** The six left panels (top and bottom rows) show price paths for six different LLMs prompted by human instructions, compared to the constant fundamental value (14, dotted line). The rightmost plot (“Mean human prices”) shows data from human experiments averaged over 19 sessions. None of the LLMs reproduces the magnitude or trajectory of the human bubble–crash pattern, though several exhibit similar tendencies. This invites agent prompt optimization (APO) tuning of LLMs to better match human data (ongoing research).

The next step in this research is to steer the AI agents, through prompt optimization, to look more human-like. **xie2025** showed the promise of this agenda. They trained AI agents across a wide range of canonical games, such as the prisoner’s

dilemma, dictator and public goods games. They were able to steer AI agents through language prompting to approximate overall human data. The language prompts represent “behavioral codes" which can be considered candidate mental states corresponding to what people prefer and choose.

2.4 What AI-behavioral science lies ahead?

Experiments on human subjects are not easy. They must be recruited to a physical lab, or connected and filtered online. IRB rules means humans have to self-select to volunteer and can’t be put into many lifelike situations. Humans get tired, and drop out of gold-standar longitudinal studies that require repeated contact. AI subjects could overcome *many* of these limits. Furthermore, social science vastly oversamples people from WEIRD (Western, educated, industrialized, rich, and democratic) societies (Henrich, Heine, and Norenzayan, 2010). Synthetic AI agents could be used to study non-WEIRD societies that are often hard to sample (such as small-scale societies in poor countries with limited internet.

Synthetic cultural agents (SCAs) can be fine-tuned using web scraping and retrieval-augmented generation (RAG). Gonzalez-Bonorino, Pantoja, and Capra (2024) created SCAs for 6 small-scale societies for which there was some human experimental data about economic games. The SCAs displayed cross-cultural qualitative patterns resembling human behavior. SCA predictions can be used as pilot data to decide which cultures are worth spending money to travel to and study up close.

AI agents could be used to conduct what we call "out-of-this-world" (OOTW) experiments (an acronym extending OOD). Some human subjects and conditions cannot ethically or feasibly be done with humans. Can we learn about how ancient Athenians trade, bargain, and speak politically? What about how traders create and crash bubbles over a 1,000-period trading horizon? Do groups of 250 AI agents create threshold public goods (such as political revolutions) differently than groups of 10, on which we have experimental human data, or groups of 1M, on which we have no controlled data? These are experiments that range from impossible to impractical with humans. All of them could be done using AI agents (Mei et al., 2024).

Of course in many experimental domains LLMs can be bad predictors of human behavior if they are not well-prompted (Gao et al., 2025). However, theory-guided prompting in the same games studied by Gao et al. (2025) can produce very good predictions, including OOD in games in a large parametrized set of novel designs

(Manning and Horton, 2025). A feature of this debate is the speed of resolution. Gao et al. (2025) was published on June 13, 2025. Manning and Horton (2025) was on arXiv less than 3 months later.

2.5 AI tools for analyzing and generating social data

AI agents also show promise as tools for processing qualitative data, coding open-ended responses, and generating new textual data. In computational social science (CSS), Ziems et al. (2023) found that zero-shot AI agents (without fine-tuning) can achieve near-human agreement on classification tasks, and excel at free-form coding.

Researchers studying social movements could use an AI agent to classify millions of tweets by protest topics and summarize long transcripts of public hearings or analyze the aggressiveness of wartime communications between adversaries. AI agents also enable rapid quantification of qualitative data by embedding text into high-dimensional representations, which can then be analyzed with conventional statistical techniques, or by simply providing numerical ratings in lieu of human labelers. However, the authors caution that AI agent output can be non-deterministic and highly sensitive to the phrasing of the prompt, necessitating careful documentation. Though, this is often true in humans as well.

It is crucial to recognize that AI agents' open-ended survey coding will not always be analogous to humans, Heyde et al. (2025) compared different AI agents for classifying free-text survey responses, only a fine-tuned model achieved satisfactory performance. Off-the-shelf AI agents misclassified responses across categories, and the distribution of codes varied widely. While AI agents may reduce labor costs, researchers should be concerned about the validity of their performance and should use humans to validate. This echoes a broader theme: AI tools are best used to augment, rather than fully replace, human analysis.

2.6 Conclusions and future directions

AI agents will transform behavioral science. They allow scholars to create and explore AI-human interactions, create synthetic AI agents to pilot-test design substitute for human participants OOD or over impossible OOTW historical and numerical scales, and use AI tools for analysis of text and do other co-science, such as detecting fraudulent science or predicting replicability (Altmejd et al., 2019).

There are certainly limits to using AI agents in social science right now. AI behavior is often different from that of humans (sometimes "hyper-rational"), and these

differences may be quirkily domain-specific- e.g., LLMs evaluate risks like humans but discount future rewards hugely more than humans do (Ross, Y. Kim, and Lo, 2024). AI agents are also trained on a corpus that reflects dominant cultures and may reflect harmful biases. Asking LLMs to compete for success (as many applications will do) creates inadvertent misalignment (El and Zou, 2025).

In co-science AI, a recent large-scale study compared human-only, AI-assisted, and AI-led teams reproducing quantitative social science results. AI-only approaches were not capable of reliably reproducing complex scientific analyses without substantial human oversight (Brodeur, Valenta, Marcoci, et al., 2025).

Our position is that research on AI-Behavioral Science benefits from the conversational cycle between AI-derived insights and criticisms that those insights provoke. An adversarial process is the best way to make progress, as long as it's based on new evidence. In political science, for example, evidence that LLMs can reproduce human election polling (Argyle, Busby, Fulda, et al., 2023) was criticized (Bisbee et al., 2024), leading to a thoughtful assessment of LLM pros and cons one year later (Argyle, Busby, Gubler, et al., 2025).

There is a lot more to learn about meta-science questions: How does the presence of AI agents affect human subjects' beliefs, trust, and cooperation? Can AI be used to pre-test experimental designs and predict outcomes? Ethicists and policymakers must also help ensure that AI benefits are equitably distributed and that potential harms are quickly identified and limited.

References

Altmejd, Adam et al. (2019). "Predicting the replicability of social science lab experiments". In: *PloS one* 14.12. URL: https://doi.org/10.1371/journal.pone.0225826.

Anthis, Jacy R. et al. (2025). "LLM Social Simulations Are a Promising Research Method". In: *arXiv preprint* arXiv:2504.02234. Version v2, revised 5 June 2025. URL: https://arxiv.org/abs/2504.02234.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, et al. (2023). "Out of One, Many: Using Language Models to Simulate Human Samples". In: *Political Analysis* 31.3, pp. 337–351. DOI: 10.1017/pan.2023.2.

Argyle, Lisa P., Ethan C. Busby, Joshua R. Gubler, et al. (2025). *Arti-"fickle" Intelligence: Using LLMs as a Tool for Inference in the Political and Social Sciences*. arXiv: 2504.03822 [cs.CY]. URL: https://arxiv.org/abs/2504.03822.

Baskaran, A. et al. (2023). "A Review of Emotions in Human–Conversational Agent Interaction". In: *International Journal of Human–Computer Studies* 178, p. 103040. DOI: https://doi.org/10.1609/aaaiss.v1i1.27477. URL: https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://ojs.aaai.org/index.php/AAAI-SS/article/view/27477&ved=2ahUKEwjJk5uHlq2VAxUIPUQIHUBiEj8QFnoECBgQAQ&usg=AOvVaw1iSxW9zSVykJ5E7yLg0E0A.

Binz, Marcel et al. (2025). "A foundation model to predict and capture human cognition". In: *Nature* 644.8078, pp. 1002–1009. ISSN: 1476-4687. DOI: 10.1038/s41586-025-09215-4. URL: https://doi.org/10.1038/s41586-025-09215-4.

Bisbee, James et al. (2024). "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models". In: *Political Analysis* 32.4, pp. 401–416. DOI: 10.1017/pan.2024.5.

Brodeur, Abel, David Valenta, Alexandru Marcoci, et al. (2025). *Comparing Human-Only, AI-Assisted, and AI-Led Teams on Assessing Research Reproducibility in Quantitative Social Science*. Tech. rep. DP 195. Institute for Replication. URL: https://www.econstor.eu/bitstream/10419/308508/1/I4R-DP195.pdf.

Buyalskaya, Anastasia, Marcos Gallo, and Colin F. Camerer (2021). "The Golden Age of Social Science". In: *Proceedings of the National Academy of Sciences* 118.5, e2002923118. DOI: 10.1073/pnas.2002923118.

Dillion, Danica et al. (2023). "Can AI language models replace human participants?" In: *Trends in Cognitive Sciences* 27.7, pp. 597–600. ISSN: 1364-6613. DOI: 10.1016/j.tics.2023.04.008.

El, Batu and James Zou (2025). "Moloch's Bargain: Emergent Misalignment When LLMs Compete for Audiences". In: *arXiv preprint* arXiv:2510.06105. Preprint, v1, 7 Oct 2025. URL: https://arxiv.org/abs/2510.06105.

Erlei, Alexander et al. (2022). "For What It's Worth: Humans Overwrite Their Economic Self-Interest to Avoid Bargaining with AI Systems". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–18. DOI: 10.1145/3491102.3517734.

Gao, Yuan et al. (2025). "Take caution in using LLMs as human surrogates". In: *Proceedings of the National Academy of Sciences*. DOI: 10.1073/pnas.2501660122. URL: https://www.pnas.org/doi/10.1073/pnas.2501660122.

Gonzalez-Bonorino, Augusto, Emilio Pantoja, and C. Monica Capra (2024). "LLMs Model Non-WEIRD Populations: Experiments with Synthetic Cultural Agents". SSRN working paper.

Henning, Thomas et al. (2025). "LLM Trading: Analysis of LLM Agent Behavior in Experimental Asset Markets". In: *arXiv preprint arXiv:2502.15800*. Advances in Financial AI Workshop at ICLR 2025. URL: https://arxiv.org/abs/2502.15800.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan (2010). "The Weirdest People in the World?" In: *Behavioral and Brain Sciences* 33.2–3, pp. 61–83. DOI: 10.1017/S0140525X0999152X.

Hewitt, Luke et al. (2023). "Predicting Results of Social Science Experiments Using Large Language Models". In: *Working Paper*. *equal contribution. URL: https://arxiv.org/abs/2306.08641.

Heyde, Leah von der et al. (2025). "United in Diversity? Contextual Biases in LLM-Based Predictions of the 2024 European Parliament Elections". In: *arXiv preprint arXiv:2409.09045*.

Horton, John J. (2023). "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" In: *arXiv preprint arXiv:2301.07543*. Version 1, January 18. DOI: 10.48550/ARXIV.2301.07543.

Jackson, Matthew O. et al. (Aug. 2025). *AI Behavioral Science*. Tech. rep. SSRN. DOI: 10.2139/ssrn.5395006. URL: https://ssrn.com/abstract=5395006.

Karpus, Jurgis et al. (2021). "Algorithm Exploitation: Humans Are Keen to Exploit Benevolent AI". In: *iScience* 24, p. 102679. DOI: 10.1016/j.isci.2021.102679.

Kim, Sung Park, Jin Lee, and S. Y. Park (2023). "Anthropomorphic Response: Understanding Interactions with AI Agents". In: *Computers in Human Behavior* 144, p. 107693. DOI: 10.1016/j.chb.2023.107693. URL: https://doi.org/10.1016/j.chb.2023.107693.

Ling, Yier and Alex Imas (May 2025). *Underreporting of AI use: The role of social desirability bias*. Tech. rep. SSRN 5232910. Revised May 18, 2025. SSRN. DOI: 10.2139/ssrn.5232910. URL: https://ssrn.com/abstract=5232910.

Maehigashi, Akihiro, Takahiro Tsumura, and Seiji Yamada (2022). "Experimental Investigation of Trust in Anthropomorphic Agents as Task Partners". In: *Proceedings of the 10th International Conference on Human–Agent Interaction (HAI 2022)*. ACM, pp. 302–305. DOI: https://doi.org/10.1016/j.chb.2022.107512. URL: https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.sciencedirect.com/science/article/pii/S0747563222003326.

Manning, Benjamin and John J. Horton (2025). "General Social Agents". In: *Available at SSRN*. DOI: 10.2139/ssrn.5521881. URL: https://ssrn.com/abstract=5521881.

Mei, Qiaozhu et al. (Feb. 2024). "A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans". In: *Proceedings of the National Academy of Sciences* 121.9, e2313925121. DOI: 10.1073/pnas.2313925121. URL: https://www.pnas.org/doi/10.1073/pnas.2313925121.

Park, Joon Sung et al. (2024). *Generative Agent Simulations of 1,000 People*. arXiv: 2411.10109 [cs.AI]. URL: https://arxiv.org/abs/2411.10109.

Peng, Kenny, Nikhil Garg, and Jon Kleinberg (2025). "A No Free Lunch Theorem for Human–AI Collaboration". In: *Proceedings of the 39th AAAI Conference on Artificial Intelligence*.

Ross, Jillian, Yoon Kim, and Andrew W. Lo (Aug. 2024). "LLM economicus? Mapping the Behavioral Biases of LLMs via Utility Theory". In: arXiv:2408.02784. arXiv:2408.02784 [cs]. DOI: 10.48550/arXiv.2408.02784. URL: http://arxiv.org/abs/2408.02784.

Schröder, Sarah et al. (2025). *Large Language Models Do Not Simulate Human Psychology*. Tech. rep. arXiv: 2508.06950 [cs.AI]. URL: https://arxiv.org/abs/2508.06950.

Takita, H. et al. (2025). "A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians". In: *npj Digital Medicine*. DOI: 10.1038/s41746-025-01543-z. URL: https://www.nature.com/articles/s41746-025-01543-z.

Treiman, Lauren S., Chien-Ju Ho, and Wouter Kool (2023). "Humans Forgo Reward to Instill Fairness into AI". In: *Proceedings of the 11th AAAI Conference on Human Computation and Crowdsourcing (HCOMP '23)*, pp. 152–162. DOI: 10.1609/hcomp.v11i1.27556.

Ziems, Caleb et al. (2023). "Can Large Language Models Transform Computational Social Science?" In: *arXiv preprint arXiv:2305.03514*.

Chapter 3

PSYCHOBIOLOGICAL CORRELATES OF EXPERIMENTAL ASSET BUBBLES

Thomas Henning, Ross Spoon, Mihai Toma, Jiatong Han, Cady Ngo, Alec Smith, and Colin F. Camerer

ABSTRACT

Asset market bubbles are episodes in which prices increase and reach persistently higher levels than their intrinsic fundamental value, ultimately “crashing" and re-aligning with the fundamental value. Experimental settings provide a controlled environment that allows for a better understanding of the mechanics of bubble formation and crashes.

3.1 Introduction

Behavioral finance has shown many ways in which investor and market behavior can deviate from rationality principles such as market efficiency. Financial bubbles are one such class of deviations that are often large and impactful, but also controversial. Bubbles are extended episodes in which the prices of an asset rise well above its fundamental value, ending in a price crash (e.g., C. Camerer, 1989) Modern bubbles include the 2000 dot-com bubble and the 2008 housing market collapse (Barberis, 2013).

Despite many theories and data, there is no consensus explanation for why bubbles form and crash. Analysts often do not agree on whether an asset is currently in a bubble until prices crash (Fama, 1970; Barberis, Greenwood, et al., 2018; Greenwood, Shleifer, and You, 2019). And bubbles in naturally-occurring markets are difficult to detect and study because fundamental values are rarely known (Lee, Shleifer, and Thaler, 1991). Experiments are therefore a useful tool to study bubbles since they allow explicit control of all the determinants of the market, including an asset’s fundamental value. In field data, price fluctuations that look like bubbles and crashes could be caused by many factors, such as changes in information and beliefs about others, to discount rates and risk preferences, as well as institutional details (such as short-selling constraints and uncertainty about new technology). Most of these complicating factors are not present in the experiments. So if bubbles are common, that suggests a psychological basis which our measures can help us understand.

This paper reports experiments using a 30-period design in which the fundamental value of the asset (derived from dividends and interest rates that were carefully controlled) is constant. Bubbles are evident if prices rise well above the fundamental value.

The design combined several measures which have not been combined together in

previous experimental designs.

- Emotional arousal: In each session, electrodermal response (EDA) was recorded from 3-5 subjects in an on-campus lab.
- Belief forecasts: In each period, just before the trading price is announced, subjects give incentivized forecasts for the current price and three future prices. We analyze whether these forecasts are rational and what models fit them.
- Risk preference: Participants' risk preferences are measured from a simple binary choice over financial risks (unrelated to the main trading task) every period. We use these data to test theories that connect returns to time-variation in risk premia.

Together, this novel set of measures allow us to separate three channels that are often difficult to distinguish in naturally occurring markets. Belief forecasts measure what traders expect prices to do, risk elicitations measure how willing they are to bear financial risk, and psychophysiological recordings measure both momentary and slower-moving arousal as markets rise and fall. Because these variables are collected repeatedly (and from all traders) within the same controlled asset market sessions, we can ask whether bubbles and crashes are better explained by beliefs, risk preferences, physiological states, or standard market observables such as lagged returns and order flow.

There are two levels of results: aggregate market phenomena, and individual-level results.

At the market-level, bubbles regularly form and typically crash to the fundamental value at the end, even with high-stakes and trader experience. Bubbles vary across sessions and the average peak price is around three times the fundamental value. Returns are highly predictable (using regression or machine learning), either from lagged returns or from traders' forecasts. Short-term arousal, measured by EDA, has little predictive power. However, 5-period moving averages of arousal are predictive of 5-period average returns.

Contrary to theories that time-varying risk premia are associated with returns, trial-by-trial risk preferences are not at all correlated with returns or phases of the bubble. We also observe that similar bubbles occur even with higher financial stakes and

with trader experience (repeating a market session a second or third time), although experience does cause bubbles to crash earlier.

The individual level is more interesting. The highest earning traders have more stable EDA profiles, and are highest near the price peak. Lower earning traders have more volatile EDA profiles, and their arousal is lowest near the price peak. This pattern is consistent with some other evidence linking emotional variation to trading performance (Lo and Repin, 2002; Coates and Herbert, 2008; A. Smith et al., 2014). Arousal is also correlated with trial-by-trial risk preference, and with “derisking" by selling assets.

Finally, forecasted returns consistently underpredict actual returns, during both bubbles and crashes: Traders are consistently surprised by the prices that they themselves create by trading (even in the high-stakes and experienced-trader conditions). Forecasts mostly fail simple rationality tests because forecasts use too little available information. However, forecasts are accurately predicted by simple models involving adaptation and extrapolation.

3.2 Brief Literature Review

Several experimental market designs have been used to study bubbles. By far the most popular is the declining fundamental value design introduced in 1988 (V. L. Smith, Suchanek, and Williams, 1988) (Henceforth, SSW). In the SSW design, participants can buy or sell a tradable asset with a known time horizon T (typically 15 periods). Holders of the asset get a stochastic dividend with expected value D_t in each discrete period. There is no redemption value after period T. As a result, the rational price path of the asset is a declining step function $P_t^* = (16 - t)D_t$ for period t.

Despite this simple fundamental value structure, empirical prices often start far below this level, rise to a bubble price well above P_t^*, and then eventually crash back towards the rational final-period price of 0 in the later market periods. Subsequent work confirmed the robustness of these bubbles to scale: they emerged even in markets with participant counts exceeding 200 (Williams, 2008), and were observed consistently across computerized classroom environments as well (Williams and Walker, 1993).

The SSW design do not resemble equities which fluctuate in value; instead, they are more similar to finite resources (gold mines or oil fields) or consumer durables which depreciate, such as cars and houses. Furthermore, since prices almost always

start far below the initial fundamental value, the subsequent process of upward equilibration toward the fundamental may be what jumpstarts beliefs that prices will increases and creates a bubble, in the form of overshooting (V. L. Smith, 2010; Kirchler, J. Huber, and Stöckl, 2012). The SSW bubbles are sometimes fragile to small changes, such as how the assets are described (Kirchler, J. Huber, and Stöckl, 2012), or showing the fundamental value in a chart rather than a table (J. Huber and Kirchler, 2012). A deeper review of the first 25 years of experimental asset market research documents that numerous such effects can be found in the SSW design (Palan, 2013).

Another popular paradigm is called “learning-to-forecast" C. Hommes et al. (2005) and C. Hommes et al. (2008). Participants do not set prices by trading. Instead, they repeatedly forecast the future price of a risky asset and the realized market price is then generated mechanically from the average forecast of the group through a pricing rule in which higher forecasts create higher prices.[1] Payoffs depend solely on forecast accuracy. Large bubbles and crashes reliably emerge in this paradigm. A lot is known about how heterogeneous behavior (C. H. Hommes, 2011; Bao, C. Hommes, Sonnemans, et al., 2012; Bao, C. Hommes, and Pei, 2021), interest changes (Bao and Zong, 2019) and selection (Anufriev and C. Hommes, 2012) work in these data.

A good design for our purposes is an asset market in which stocks pay dividends but discounting is induced by an opportunity cost because experimental currency earns interest, and assets are redeemed at a known value after period T (Bostian, Goeree, and C. A. Holt, 2005). This design creates a constant fundamental value; a bubble is easily recognized as prices persistently above this level.[2] This constand fundamental design produces reliable bubbles and has replicated successfully (A. Smith et al., 2014; Charles A. Holt, Porzio, and Song, 2017).

Our design also measures emotional arousal, motivated by findings from a few other emotiona studies. Emotions, such as exuberance or panic, have been shown to correlate with market volatility and price reversals in actual markets among

[1] Participants understand this qualitative relationship, but do not know the exact formula for price-setting.

[2] Several other designs have induced a constant fundamental value, and generally observe mis-pricing. C. F. Camerer and Weigelt (1993) and Ball and Charles A Holt (1998) used a design in which discounting was induced by a probability of assets expiring in each period. C. Noussair, Robin, and Ruffieux (2001)’s design created a constant fundamental by paying expected dividends of zero. Caginalp, D. Porter, and V. Smith (2001) deferred dividend payments rather than paying sequentially. These papers showed smaller bubbles than SSW.

professional traders (Lo and Repin, 2002; Fenton-O'Creevy et al., 2012; Weitzel et al., 2020).[3] Breaban and Charles N. Noussair (2018) found that positive emotional states (inferred from faces) correlated with purchases and overpricing, and that negative states correlated with selling, low prices, and price decreases. One study reported that inducing emotional excitement using short film clips magnified bubbles in the SSW design (Andrade, Odean, and Lin, 2015), but that finding failed to replicate (C. Huber et al., 2024). It has also been shown that the simple awareness of one's heart rate (interoception) among participants is correlated with better market performance and a longer professional career (Kandasamy et al., 2016; Bossaerts, 2021). Bossaerts et al. (2024) measured heart rate and found that showed that anticipatory emotional responses were correlated with higher earnings than reactive responses. Other papers linked asset bubble formation to EEG activity and eye-tracking (Toma and Miyakoshi, 2021; Toma, Cepoi, et al., 2023). Finally, a study using fMRI reported that insula activity in successful traders rose before bubble peaks and was associated with profitable selling (A. Smith et al., 2014).[4]

These findings motivated us to measure emotional arousal to see whether emotions are associated with individual behavior and market-level prices in the constant-fundamental domain.

In this paper we also elicited price forecasts (following (C. F. Camerer and Weigelt, 1993), (Haruvy, Lahav, and Charles N Noussair, 2007)). We compared properties of these forecasts to evidence from the growing amount of subjective belief data in finance and macro-finance[5] and explore their association with trader behavior (Corgnet et al., 2022).

There is evidence to suggest that participant experience has an attenuation effect on market bubbles, for example, Dufwenberg, Lindqvist, and Moore (2005) and D. P. Porter and V. L. Smith (2010) demonstrate that even a few experienced participants can eliminate bubble formation in lab-based markets. An "experience effect" was

[3] Other physiological states also matter; (Berkman et al., 2017) shows that circadian mismatch and sleep deprivation can significantly amplify volatility and bubble formation in global asset markets.

[4] The insula is region involved in translating bodily emotions into conscious feelings and is reliably associated with financial uncertainty Mohr, Biele, and Heekeren, 2010; Morriss, Gell, and Reekum, 2019

[5] Many analysts have found overreaction of beliefs about financial and macroeconomic variables(De Bondt and Thaler, 1990; Amromin and Sharpe, 2014; Greenwood and Shleifer, 2014a; Gennaioli, Ma, and Shleifer, 2016; Bordalo, Gennaioli, La Porta, et al., 2019; Bordalo, Gennaioli, Ma, et al., 2020; Barrero, 2022). A few studies show underreaction (Abarbanell and Bernard, 1992; Bouchaud et al., 2019; Ma et al., 2020). Afrouzi et al., 2023 find overreaction in experiments with artificial time series, echoing an older literature including (Lovell, 1986; Schmalensee, 1976)

first shown in SSW, who found that bubbles were reduced when participants had experience trading within the same participant group. Several other studies found similar results, and that experience effects held even when across different participant groups.(King, V. Smith, et al., 1993; Van Boening, Williams, and LaMaster, 1993; Haruvy, Lahav, and Charles N Noussair, 2007). Other studies, however, suggest that experience alone is not sufficient to prevent bubbles from forming, though experience often causes bubbles to form in smaller magnitude (Sunder, 1995). For instance, Hussam, D. Porter, and V. L. Smith (2008) found that the "experience effect" can be mitigated with different market structures, such as changing the amount of market liquidity and implementing dividend shocks. Kopányi-Peuker and Weber (2021) also found that experienced participants did not entirely eliminate bubble formation, and that learning effects alone are insufficient to prevent market bubbles. The impact of subject experience on bubble formation, therefore, remains an open question.

We provide additional background on time-varying risk premia, High-Stakes experiments, and more detailed information regarding subject experience in section 3.9 of the Appendix.

3.3 Experimental Methods

Our experiment created a series of markets in which participants traded a single risky asset with each another using Otree software.[6] In each session, 3-5 participants participated in a lab at Caltech, wearing Empatica wristband sensors. More experimental details are in Appendix 3.14.

Participants traded a risky asset (stock) for an experimental currency (CASH) in a call market used in A. Smith et al. (2014) and Charles A. Holt, Porzio, and Song (2017). At the end of every trading period t, the stock paid a dividend D_t, which was equally likely to be 0.4 or 1 units of CASH (so $\mathbb{E}(D_t) = .7$). The dividend D_t was the same for all traders in period t, and independent and identically distributed (iid) across periods. CASH earned a fixed 5% interest.

Each participant submitted buy or sell orders for stocks, stated in units of CASH. After all orders were submitted, a uniform market-clearing price was calculated where supply met demand and every trade was executed at that single price (see Appendix 3.12).

The equilibrium price P^* is the value at which an investor is indifferent between

[6]The software was developed at Virginia Tech using open source Otree (Chen, Schonger, and Wickens, 2016) and was hosted on the cloud computing platform *heroku.com*.

investing P^* units in cash and earning an r percent return, or buying a share at price P and earning a dividend D_t plus an expected capital gain. Equating these two returns (and assuming risk-neutrality) gives an "Euler equation"

$$rP_t = \mathbb{E}(P_{t+1}) - P_t + \mathbb{E}(D_t) \tag{3.1}$$

In the steady-state solution, $P_{t-1} = P_t$, so $rP_t = 0.7$ and $P_t = 0.7/r = 14 \equiv P^*$. Intuitively, if prices are below 14, participants should bid up the price because the expected 0.7 dividend yield $(.07/P)$ is greater than their return from interest. If prices are above 14, participants should sell and invest the proceeds into cash, earning more than the dividend yield. After period 30, stock shares were redeemed for 14 units of CASH, the equilibrium price. Because the terminal redemption prices is commonly known to all participants, the solution $P^* = 14$ can also be computed by backward induction.

There is a small technical wrinkle to the discussion above; ignoring the finite time horizon behaviorally, there can be growing bubbles, so that $P_t = P^* + B_t$, if the bubble premium, above intrinsic value P^*, grows at an expected 5%.[7]

In our design (and in theory), these kinds of 'quasi-rational' bubbles should not exist because the asset's finite life should kill expected bubble growth at some point (Blanchard and Watson, 1982; Tirole, 1985). If traders are forward-looking, once bubble growth is expected to be too low, the bubble should crash.[8]

A market session was 30 trading periods, each lasting approximately 80 seconds. Each of those periods consisted of four pages: an Order Submission Page, a Forecast Page, a Round Results Page, and a Risk Elicitation Page. Images of the pages shown to participants are available in the section 3.12 in the Appendix.

On the first page, the Order Submission Page, participants were instructed to place bid and ask limit orders in a grid. Participants had twenty seconds to submit their orders and could simultaneously input up to three bids and three asks during any given period. They could not place bids above their own ask order prices. Short

[7]Suppose $E(B_{t+1}) = B_t(1 + r) + \epsilon_t$ where $\mathbb{E}(\epsilon_t) = 0$. Then prices $P_t = P^* + B_t$ could include a growing-bubble component of that type, $B_t \geq 0$, as long as there is an expected return of r on the bubble component. Indeed, Figure 3.5 shows that the average price return declines slowly across periods in most markets, but it is not far from the predicted 5% for the first 15-20.

[8]However, there are many examples in related literature, such as repeated prisoners' dilemma and centipede games, which illustrate the empirical limits of backward induction in long-lived games (McKelvey and Palfrey, 1992).

sales and margin purchases were not allowed, although some studies have included those features. [9]

After participants submitted orders they saw a Forecast Page. Participants then had 30 seconds to make forecasts of future market prices at four different time horizons– 0, 2, 5, and 10 periods ahead.[10] We analyze whether participants traded consistently with price forecasts, whether forecasts predict prices, whether forecasts satisfy rationality tests, and what kinds of models fit forecasts.

After the Forecast Page, the price is calculated and participants see the price and a summary of their trades, shares, and cash.

Participants were then shown a Risk Elicitation Page (Figure 3.18), which consisted of a sequence of four pairs of monetary lotteries. Each pair of lotteries consisted of a riskier lottery and a less risky lottery, and participants had five seconds to make each selection. The task followed the Dynamically Optimized Sequential Experiment (DOSE) protocol introduced by Chapman et al. (2024), which delivered two parameters for each participant and period: the constant-relative-risk-aversion coefficient ρ and a risk-consistency parameter μ. After each choice, the probabilities of the lotteries were updated according to DOSE so that the next choice provided the most information about the participant's risk parameters. This short exogenous design allows us to gauge whether risk preference variation is a potential driver of price bubble formation, while separating the measure of risk from the market-generating process.

Data Collection

Throughout each market, we record electrodermal activity (EDA), heart rate (HR), blood volume pulse (BVP), and skin temperature (TEMP), enabling us to link moment-by-moment changes in arousal and affect to the trading decisions participants made as bubbles inflated and deflated.[11]

We measured physiological responses for in-lab participants using a combination of

[9] Some experimental asset-market studies have varied these trading constraints directly. For example, King, V. L. Smith, et al. (1993) examine the robustness of bubbles under treatments including short selling and margin purchases; Haruvy and Charles N. Noussair (2006a) study the effect of relaxing short-selling constraints in experimental spot asset markets; and Ackert et al. (2006) study margin, short selling, and lotteries in experimental asset markets. See also Hauser and J. Huber (2012) for related evidence on short-selling and borrowing constraints in laboratory markets.

[10] Forecasts are truncated when there are few periods left in the market. If, for example, there were only nine periods left to play, participants were not asked to predict the price ten periods ahead.

[11] More details on our biometric measures can be found in section 3.13 in the Appendix. Note this paper only focuses on EDA results, though these other measures were captured simultaneously.

Empatica E4 devices and Empatica EmbracePlus devices.[12] Most studies that have measured biological responses and their relationship to finance have employed large, complicated devices to measure various biological responses. For instance, Lo and Repin (2002) required six separate biosensors and over five minutes of set-up time. These complicated setups usually meant that only one subject could be monitored at a time. Our use of Empatica E4 wristbands circumvents this limitation, allowing us to gather data on multiple participants per market simultaneously. We collect the following from each in-lab participant: (EDA), heart rate (HR), blood volume pulse (BVP), and skin temperature (TEMP).[13] Full details of these measures can be found in section 3.13 in the Appendix. We did not record the biophysiology of the online participants.

We used these psychophysiological data to predict stock price and bubble crash dynamics. In addition, we conducted exploratory analyses to examine how psychophysiology responds to specific events, such as the revelation of forecast errors.

The experiment was approved by Caltech's IRB. Participants were given a brief demographics survey before being shown the written and video instructions (full instructions are available in section 3.30 of the Appendix). Subjects were then entered into the market experiment. After the market ended, participants completed a post-experiment survey that included a free-response question asking participants to describe their trading strategy. Full details of these processes can be found in the section 3.32 in the Appendix.

We collected data from 19 standard markets, with an average of 18 participants per market (342 participants in all). In each of these markets, we collected biometric data from approximately three to five in-lab participants, totaling 72, representing around 21% of all participants, detailed break down can be found in section 3.28 in the Appendix.[14]

We refer to these markets as Standard markets. They serve as a benchmark for the two treatments in our study.

[12] These devices have been shown to provide high-quality data in several experiments (McCarthy et al., 2016; Schuurmans et al., 2020; Milstein and Gordon, 2020; Tashjian et al., 2022).

[13] Of primary interest for data analysis is the Round Results page, where participants learn the updated market price and presumably update their forecasts about the market. It is at this time that a subject is most likely to be surprised and emotionally aroused.

[14] Note we did identify that there were differences in the behavior between our in-lab and online participants, with in-lab participants being significantly more adept. Breakdown of these differences can be found in section 3.11 of the Appendix.

Our first variation was running four “High-Stakes markets", for which we recruited 59 participants, roughly 16 participants each. These participants were recruited from *prolific.com*. The key difference in these markets was the calculation of the bonus payment. Instead of paying one USD for every 200 units of CASH, the bonus was raised to five USD for every 200 units of CASH earned by participants during the market. Results of the High-Stakes markets are further discussed in section 3.4.

In the second treatment, we invited prolific participants who had previously traded in one of the Standard markets and conducted five new "Experienced markets." Each market comprised $\tilde{1}9$ participants on average, yielding a total sample of 93 participants. We recruited these subjects exclusively from *prolific.com*, so we did not collect any biometric data from them. The configuration of these Experienced markets was identical to the Standard markets in every way (including the bonus payout ratio of 200 CASH to one USD). More details are further discussed in section 3.4.

In all markets, participants received a base payment of $20 and a bonus payment based on their performance in the various experimental tasks. A full description of the payment scheme can be found in section 3.12 in the Appendix.

3.4 Price and forecast data

In this section, we focus on insights about the aggregate market from the various collected data. We first attempt to predict stock returns using market measures and the biometric data. Figure 3.1a displays the average price path of the 19 experimental markets. Figure 3.1b shows all the market paths.

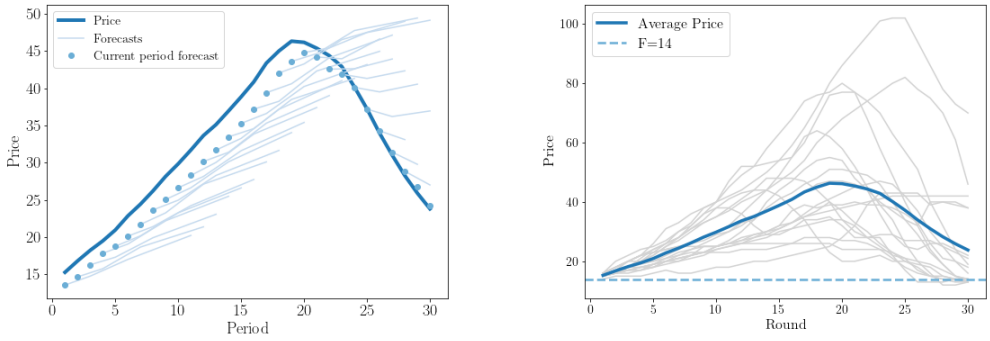
Bubble magnitudes and duration have conventionally been measured by a series of market-level statistics.[15] Figure 3.2 illustrates these measures for a single exemplar market.[16]

High-Stakes Markets

A long-standing interest in experimental economics is whether behavior is fundamentally changed by increasing the financial stakes. Higher stakes may increase attention and cognitive effort, potentially moving choices closer to rational-model predictions and reducing behavioral noise (V. L. Smith and Walker, 1993; C. F.

[15] See section 3.27 in the appendix for an explanation of how we calculated these metrics. A listing of observed bubble metrics from all markets can be found in Table 3.34

[16] This market was chosen as an exemplar because it generated the lowest sum of normalized squared deviations from the all-market means, summed across all measures.



(a) Average Standard markets (b) All Standard markets

Figure 3.1: (a) Average prices and forecasts across 19 Standard markets. The thick line shows the average price path. Each circular point on the graph represents the average current period forecast (just before the current price is announced) averaged across participants. The thin lines extending from the circular point are the "tongues" of average future forecasts in period t for each of the 2-, 5-, and 10-period ahead prices. Market-specific graphs can be found in section 3.15 in the Appendix. (b) All Standard markets are shown here as thin light gray lines. The average standard price over time is shown as the bold line. Fundamental value is shown as a dashed line at F=14.

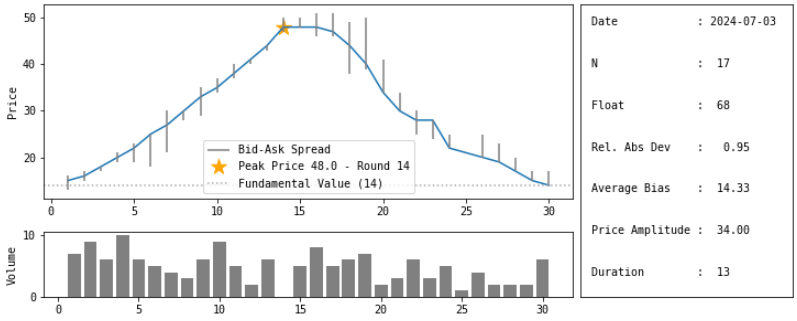
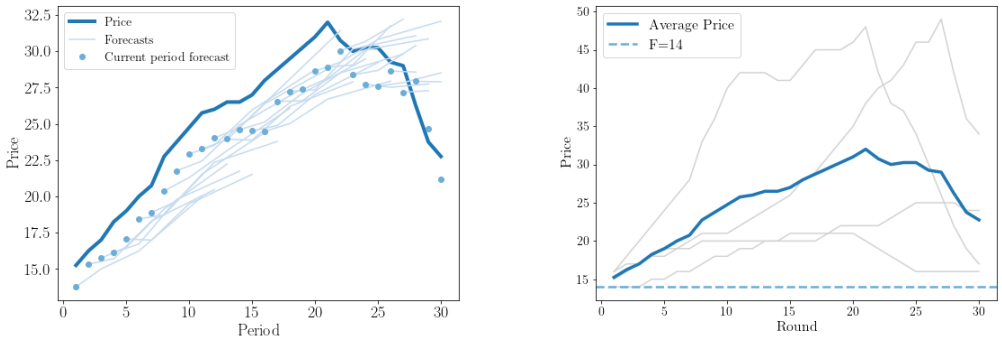


Figure 3.2: Price path and volume of a typical market along with commonly used bubble metrics. **Rel Abs Dev** is a volume-weighted average of how far the price deviates from fundamental value, **Average Bias** is a simple average of price deviations. **Price Amplitude**, shows the difference between the market's highest and lowest prices. This gives an idea of the range of market prices observed. **Duration** is the longest number of periods in the market during which the price increases.

Camerer and Hogarth, 1999). In asset-market experiments, this logic suggests that larger incentives could reduce mispricing and result in smaller bubbles.

To investigate, we ran four High-Stakes markets in which the market bonus was 5 times as large as in the Standard treatment. These sessions had N=59 (12-17 traders per market). Figure (3.3a) shows the results. We observed bubbles and crashes to the fundamental in all but one market. High-Stakes markets peaked about three periods later (22.00 compared to 18.84, $p = 0.178$) and had smaller price peaks (52.0 compared to 35.8, $p = 0.256$). More details are in Appendix Figure 3.19). With these small samples, there is not much power to detect differences among these different market conditions.



(a) Average of All High-Stakes markets

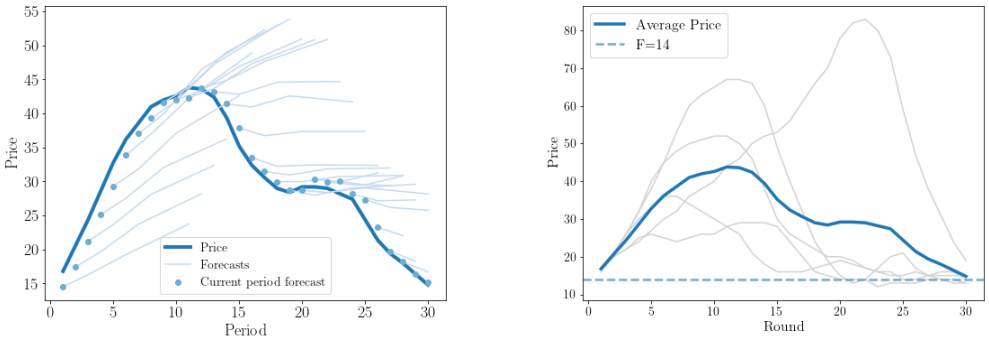
(b) All High-Stakes markets

Figure 3.3: (a) Average prices and forecasts across four High-Stakes markets. The thick line shows the average price path. Each circular point on the graph represents the current period forecast (just before the current price is announced) averaged across participants. The thin lines extending from the circular point are the "tongue" of future forecasts in period t for each of the 2-, 5-, and 10-period ahead prices. Market-specific graphs can be found in section 3.15 in the Appendix. (b) All Experienced markets are shown here as thin light gray lines. The average standard price over time is shown as the bold line. Fundamental value is shown as a light dashed line at F=14.

Experienced Markets

As noted in the brief literature review, there is substantial, but mixed evidence, that prior experience participating in market sessions reduces bubbles. The closest design to our is (Kopányi-Peuker and Weber, 2021) with some small differences (which prevent the market from crashing to the fundamental).[17] Three consecutive

[17] They have only six traders per session, interest cash cannot be used to buy assets, and subjects are told the terminal period is sometime between 25 and 40 periods.



(a) Average Experienced markets

(b) All Experienced markets

Figure 3.4: (a) Average prices and forecasts across five Experienced markets. The thick line shows the average price path. Each circular point on the graph represents the current period forecast (just before the current price is announced) averaged across participants. The thin lines extending from the circular point are the "tongue" of future forecasts in period t for each of the 2-, 5-, and 10-period ahead prices. Market-specific graphs can be found in section 3.15 in the Appendix. (b) All Experienced markets are shown here as thin light gray lines. The average standard price over time is shown as the bold line. Fundamental value is shown as a light dashed line at F=14.

market experiences, in a single session, did not reduce prices but did reduce price variation across the three experiences.

To study effects of market experience, we conducted an exploratory analysis in which *prolific.com* participants from Standard sessions were invited to participate again. This created five sessions with N=74 *prolific.com* participants (the number of participants ranged from 13 to 22). We ran each of these returning participants through the same instructions and market procedure as in their first market.

Figure 3.4a shows the average price path and forecasts for these Experienced markets.[18] In our Experienced markets, peak prices were similar to those in Standard markets (53.4 compared to 52.0, $p = 0.831$). However, Experienced markets peaked significantly earlier, with an average peak period of 12.27 compared to 18.84 in Standard markets ($p = 0.038$). We view this finding as exploratory because it is based on a small sample of only five Experienced markets, though broadly, it is consistent with replicated evidence from the experimental asset-market literature that trader experience affects bubble dynamics (C. Huber et al., 2024).[19]

[18]Figures for each session are in Appendix 3.15.

[19]C. Huber et al. (2024) conduct high-powered, preregistered replications of 17 claims from prior experimental asset-market studies. They replicate prior evidence that experience reduces but does not

Note we did not find any significant selection effects present in the experienced sample.[20]

[21]

Predicting Returns in High-stakes and Experienced Markets

We again repeat the OLS regression from the previous sections, and once again we find that lag returns are highly predictive of future returns (Table 3.11). The full model (column 3) explains approximately 59% of the variance in next-period returns ($R^2 = 0.593$). The only significant predictor here is the own-lagged return R_{t-1} ($p = 0.005$).

3.5 Predicting Market Phenomena

Our analysis proceeds in two stages: first, a pre-registered regression to understand which features are most correlated with asset returns, and a pre-registered machine learning model to understand the amount of predictive power our dataset has for returns prediction.

Predicting Returns: Regression

We find that our measure of arousal, average EDA tonic activity (EDA^{ton}), is not predictive, while numerous traditional measures remain informative, suggesting that our current measure of individual arousal does not improve predictive accuracy above and beyond established price and trading metrics.

Predicting Returns OLS Model Standard Markets

We use OLS regression to examine the relationships between market returns, biometric indicators, and trading behavior variables across the 19 experimental markets (Table 3.2). Our analysis follows our preregistered plan, the only deviation is that we eliminated f^t_{t-10} because of suspected multicollinearity and the shift required to incorporate those drops to many rows for our tests to maintain power.[22]

eliminate bubbles, supporting the broader claim that experience can change asset-market dynamics. In our setting, experience primarily shifts the timing of the bubble peak rather than reducing peak magnitude. Distribution of Experienced participants is available in Figure 3.24 in the Appendix.

[20] A detailed breakdown of selection effects regarding these participants is available in section 3.16 in the Appendix. Note that in all, we had 93 Experienced market participants, some of whom repeated multiple times.

[21] We ran a model that mirrors the pre-registered regression except for the EDA since it is not collected from online-only subjects. Table 3.20

[22] A faithful execution of the pre-registered model can be found in Table 3.18

Table 3.1: Summary of Regression Notation

Symbol	Definition
f^t_{t-h}	Cross-Sectional Average participant forecast for period t, from horizon $h \in \{0, 2, 5, 10\}$ periods earlier.
f^{t+h}_t	Cross-Sectional Average participant forecast, formed at time t, for the asset price in period $t + h$, using horizons $h \in \{0, 2, 5, 10\}$.
ρ	Cross-Sectional Average participant risk aversion parameter from DOSE risk elicitation task.
$30 - t$	Number of periods remaining until the end of the experiment.
Ω_t	Sentiment measure of bid vs. ask order imbalance.
Υ_t	Asset Allocation (Cash Value/Stock Value) across market participants in period t.
V_{t-1}	Volume of trades in the previous period.
EDA^{ton}_t	Cross-Sectional Average of tonic EDA activity, averaged across participants for a given period t.
$EDA^{ton}_t\text{MA}_5$	Backward-looking five-period moving average of EDA^{ton}_t. $EDA^{ton}_t\text{MA}_5 = \frac{1}{5}\sum_{i=0}^{4} EDA^{ton}_{t-i}$
R_{t-1}	Log return from the previous period ($= log(\frac{P_{t-1}}{P_{t-2}})$)

Notes: Unless otherwise noted, forecasts are averaged across participants at the market level; individual-level notation is introduced explicitly when used. Note that f^t_t denotes the nowcast, i.e., the forecast for the current period's price formed immediately prior to price revelation. Rounds without applicable forecasts (rounds 1–5) were excluded from the analysis.

The final specification (column (5)) explains approximately 70% of the variance in next-period returns ($R^2 = 0.704$). One predictor emerges as highly statistically significant: the own-lagged return R_{t-1} ($\beta = 0.623$, $p < 0.001$). While the three forecast variables (f^t_t, f^t_{t-2}, f^t_{t-5}) maintain marginal significance in the final model ($p < 0.1$ for all three), they become even more significant when lagged returns are excluded (column (4), where all three carry $p < 0.01$). These results indicate that recent return dynamics and near-term participant forecasts contain meaningful information about subsequent price movements.

By contrast, $\text{EDA}^{\text{ton}}_{t,\text{MA}_5}$ enters positively but is not statistically significant in the full specification (Table 3.2, column (5), $\beta = 0.129$), suggesting that tonic arousal does not materially predict period-by-period returns once traditional market controls are included. A five-period moving average of tonic EDA shows marginal significance in a bivariate specification (Table 3.2, column (2), $\beta = 9.783$, $p < 0.1$), but this association lacks robustness to the inclusion of additional controls (Table 3.2, columns (4) and (5)). Lagged returns, by contrast, are highly significant across all specifications in which they appear.

We suspected multicollinearity among the forecast variables, and our analysis in 3.20 bears that out. We found a high degree of correlation among those variables and the first principal component explained over 91% of their variance. We then reran these models replacing the three forecast variables with their first principal component (see Table 3.21 in the Appendix) with no major changes in results.

Table 3.2: Predicting Next-Period Log Returns R_t (returns $\times 100$)

	(1)	(2)	(3)	(4)	(5)
Intercept	-0.621***	-5.032*	-0.266	-5.097	-0.331
	(0.115)	(2.660)	(2.805)	(4.889)	(2.972)
R_{t-1}	0.827***		0.623***		0.623***
	(0.034)		(0.051)		(0.051)
$EDA_t^{ton}\text{MA}_5$		9.783*		-0.430	0.129
		(5.420)		(2.791)	(1.570)
ρ			0.520	-0.831	0.531
			(0.739)	(0.694)	(0.769)
V_{t-1}			-0.089	-0.029	-0.088
			(0.064)	(0.075)	(0.067)
$30 - t$			0.179*	0.560***	0.179*
			(0.092)	(0.154)	(0.092)
Ω_t			0.010	0.034	0.010
			(0.025)	(0.032)	(0.025)
Υ_t			0.001	0.007	0.001
			(0.007)	(0.011)	(0.007)
f_t^t			0.138*	0.603***	0.138*
			(0.071)	(0.067)	(0.072)
f_{t-2}^t			-0.130*	-0.466***	-0.131*
			(0.071)	(0.107)	(0.070)
f_{t-5}^t			-0.089*	-0.242***	-0.088*
			(0.044)	(0.070)	(0.042)
Observations	456	456	456	456	456
R^2	0.668	0.012	0.704	0.582	0.704
Adjusted R^2	0.668	0.010	0.698	0.574	0.697

Notes: Dependent variable: future log return R_t (scaled by 100). Standard errors (in parentheses) are heteroskedasticity- and cluster-robust (clustered by market). $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Why do our markets not display the commonly observed "cash-inflow" effect, as shown by the lack of significance of our asset allocation measure Υ_t? One possible mechanism linking liquidity to bubbles is a house-money channel: prior gains

or greater cash holdings may make traders more willing to take risk, increasing speculative demand for the risky asset (Thaler and Johnson, 1990; Hsu and Chow, 2013; Frino, Grant, and Johnstone, 2008). In many experimental asset markets, this channel may be reinforced by the fact that idle cash is unattractive. When uninvested balances earn zero, additional cash can "burn a hole in traders' pockets", creating incentives to move excess liquidity into the risky asset.

Our environment differs in a crucial respect. Cash earns a 5% risk-free return each period, so holding cash is itself a profitable strategy. As a result, additional liquidity does not automatically translate mechanically into greater speculative demand, because buying the risky asset requires giving up a substantial safe return. Consistent with this, markets with more cash in our experiment do not exhibit systematically larger bubbles, suggesting that the impact of liquidity on mispricing depends on the return of safe assets: when cash is remunerated, cash-inflow effects and house-money effects from accumulated gains may be attenuated by the opportunity cost of holding the risky asset.

Although asset allocation and risk-preference measures do not significantly predict returns in our full specification, non-lagged return variables still contain substantial predictive information. In particular, when lagged returns are excluded, the model retains strong predictive power for future returns (adjusted $R^2 = 0.537$), driven largely by the forecast variables. We interpret this as evidence that participant forecasts are highly correlated with recent return dynamics and capture much of the same information. This relationship can be seen in Table 3.22 in the Appendix.

Predicting Returns in Standard Markets: Machine Learning

Using a machine learning approach focusing on Support Vector Regression (SVR),[23] we find that while traditional market variables predict returns with substantial accuracy, the inclusion of EDA^{ton} does not meaningfully improve out-of-sample performance. This provides evidence that the null result for EDA^{ton} in our simple OLS does not reflect model misspecification but rather a true lack of predictive content, further supporting the idea that physiological arousal, as measured by EDA^{ton}, does not add incremental predictive power beyond standard metrics.

[23] Support Vector Regression (SVR) is a kernel-based machine learning algorithm that is effective for capturing both linear and non-linear relationships in data. SVR aims to find a function that deviates from the actual observed targets by a value no greater than a specified margin while being as flat as possible.

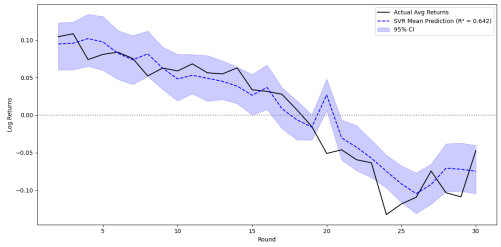
We choose this framework as SVRs are relatively simple ML models and thus often excel in generalization when the data are complex and limited. We also used several other well-established ML models to predict returns; results are presented in section 3.22 in the Appendix.

Our model is trained on period-by-period data in a given market. We first split our data by market, so that 80 percent of the markets were included in our training set and 20 percent were included in our test set. Given that we had 19 total markets, this resulted in 15 markets for training and four held-out markets for testing the final model. This split provides us with an out-of-sample test-set predictability measure, ensuring that our model correctly generalizes to out-of-sample data. We chose to compare these two models to understand which approach was most effective for our dataset. Within the training set, we trained our model using a 10-fold cross-validation schema; each fold had 13 markets in the train set and two markets in the validation set. This split was repeated 10 times in order to fit the parameters that were then tested on the four held-out test markets. In each fold, we ensured that our training dataset was split by market, so that all the data from a given market's periods were present in the same fold. This was done to help combat "data leakage", which can be a significant problem when using machine learning techniques on time-series data.

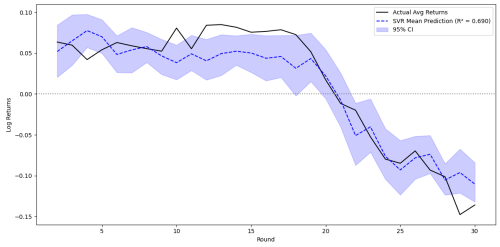
To tune the model's performance, we employed a grid search over two key hyperparameters: the regularization parameter and the kernel function. The regularization parameter (c) was tested at values 0.1, 1, and 10. This parameter controls the trade-off between achieving a low training error and maintaining a low model complexity. A smaller (c) allows for a more generalized model, while a larger (c) fits the training data more tightly, potentially at the cost of overfitting. We tested our model with two different kernel functions, as kernel choice has a significant impact on model flexibility and interpretability: the linear kernel and the radial basis function (RBF) kernel. The linear kernel assumes a linear relationship between features and the target, while the RBF kernel maps data into a higher-dimensional space to capture more complex, non-linear patterns.

The features used for training the SVR were carefully selected to capture the most plausible predictors of asset returns and mirror the variables selected in our regression model, with a notable absence of lagged returns. These features include averages of DOSE risk variables (ρ, μ), forecasted price levels at different horizons, and statistical measures of (EDA^{ton}). We also included lagged volume and the

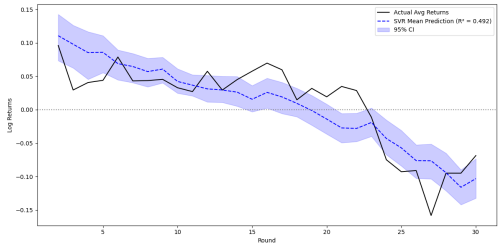
number of rounds remaining. The exclusion of lagged returns allowed us to see how predictive our variables are, separate from the return autocorrelation. Detailed descriptions of these features and the selection process are provided in section 3.18 in the Appendix.



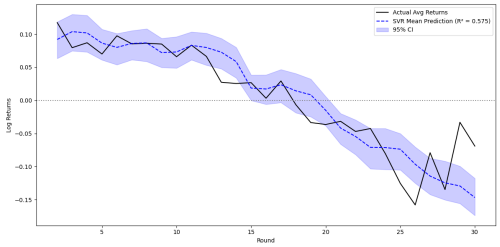
(a) Holdout: Markets 1–4



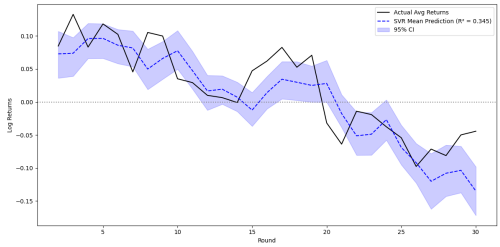
(b) Holdout: Markets 5–8



(c) Holdout: Markets 9–12



(d) Holdout: Markets 13–16



(e) Holdout: Markets 17–19

Figure 3.5: Mean out-of-sample SVR return predictions. Shaded regions show 95% confidence intervals from 1,000 bootstrap resamples of the training data, evaluated on a fixed held-out test set. The majority of realized returns lie within the bootstrap confidence bands across all held-out market groups, suggesting the model captures common market dynamics and provides well-calibrated out-of-sample uncertainty. For figure clarity, we choose not to plot the confidence interval of our held-out samples; for interested readers, this can be seen in Figure 3.29 in the Appendix.

Confidence intervals (CIs) for model performance metrics (MAE and R^2) were estimated using bootstrapping[24] with 1000 iterations.

[24]Bootstrapping is a statistical resampling technique that estimates variability and uncertainty by repeatedly taking random samples from the original dataset with replacement to create multiple

Model	Stage	MAE Mean	Range (MAE) Low	Range (MAE) High	R^2 Mean	Range (R^2) Low	Range (R^2) High
SVR (w/ EDA^{ton})	Test	0.048	0.044	0.052	0.517	0.336	0.663
	Train	0.044	0.043	0.044	0.627	0.609	0.646
SVR (w/o EDA^{ton})	Test	0.046	0.042	0.056	0.544	0.261	0.680
	Train	0.043	0.042	0.045	0.626	0.610	0.651

Table 3.3: Average SVR model performance across held-out market validation groups. Mean values are computed by averaging the scores across all held-out samples. Ranges report the minimum and maximum score observed across the held-out market groups.

Our analysis reveals that traditional measures exhibit moderate predictive power for asset returns (MAE mean 0.046, R^2 mean = 0.544 for the non-EDA model)[25]. However, the inclusion of (EDA^{ton}) data does not improve model performance. Details of feature weights can be found in Table 3.4.

Feature	Mean Importance	Range Low	Range High
$\bar{f}_t^t$	0.348	0.308	0.365
$\bar{f}_{t-2}^t$	0.221	0.187	0.281
$30 - t$	0.130	0.116	0.148
$\bar{f}_{t-5}^t$	0.126	0.076	0.156
$\sigma(EDA_t^{ton})$	0.053	0.028	0.076
skew(EDA_t^{ton})	0.053	0.025	0.070
$\overline{EDA}_t^{ton}$	0.028	0.010	0.044
$DOSE_{\rho,t}$	0.015	0.004	0.031
V_{t-1}	0.013	0.005	0.030
$DOSE_{\mu,t}$	0.012	0.001	0.025

Table 3.4: SVR feature importance weights. Mean importance weights are reported across held-out market validation groups. Ranges indicate the minimum and maximum importance weight observed across held-out groups.

simulated datasets. After 1000 iterations, the 95 percent confidence intervals were derived from the 2.5th and 97.5th percentiles of the bootstrapped distributions, providing robust estimates of model uncertainty and generalizability.

[25]To assess robustness to market selection, we repeated the w/o EDA^{ton} evaluation across 100 different random holdouts of 4 markets. Across these splits, test performance for the SVR model averaged MAE = 0.045 (SD = 0.006) and R^2 = 0.559 (SD = 0.111). The variability in R^2 across splits reflects heterogeneity in predictability across market environments rather than estimator instability; performance remained positive across the vast majority of holdouts.

3.6 Associations of Arousal with Trading Behavior

Earlier sections show that measures of tonic arousal (EDA^{ton}) contribute little to predicting one-period-ahead returns using standard OLS and machine-learning approaches. This null result likely reflects sampling constraints inherent to our hybrid experimental design, where biological data were captured from only a fraction of the participants in each market. Because short-run price changes are driven by the aggregate actions of the entire market, a partial biological sample naturally weakens direct price predictions. Consequently, we shift our focus toward longer-horizon market patterns and non-price dimensions of individual trading behavior.

Crucially, the analyses presented in this section are entirely exploratory. Rather than testing preregistered causal hypotheses, we examine statistical associations to highlight promising avenues for future research. Across these investigations, we find that physiological arousal is meaningfully associated with participant earnings, de-risking decisions, and revealed risk aversion.

EDA^{ton} Around Price Peak

We examined whether individual differences in EDA^{ton} are associated with trading performance across the market cycle. To do this, we define $\tau = t - t_p$ as the number of periods a given round t sits from its market's peak round (t_p), establishing a consistent temporal anchor to determine whether physiological responses are anticipatory or reactionary. Grouping our physiological data by this distance-from-peak variable, we define our primary metric EDA^{ton}_t as the average EDA^{ton} activity observed at each relative period τ.

A pervasive intuition among practitioners is that successful trading requires emotional control rather than heightened arousal or reactivity. Consistent with this view, we find that high earning participants exhibit lower volatility in EDA^{ton} than low earning participants. This aligns with field evidence that day-traders with higher emotional reactivity experience poorer returns (Lo and Repin, 2002) and that higher autonomic emotional regulation (measured by heart rate variability) distinguishes expert traders from less experienced peers (Fenton-O'Creevy et al., 2012).

To observe this, we sorted participants into terciles based on their relative market bonus performance. Figure 3.6a shows that EDA^{ton}_t activity is less volatile among the highest earners compared to the lowest earners; which is confirmed by a Levene test comparing variance of EDA^{ton}_t between the to groups ($\sigma_{\text{hi}} = 0.025$, $\sigma_{\text{lo}} = 0.068$, $p = 0.003$). A plot of participants sorted into above- and below-median earnings shows

a similar pattern (Figure 3.32).

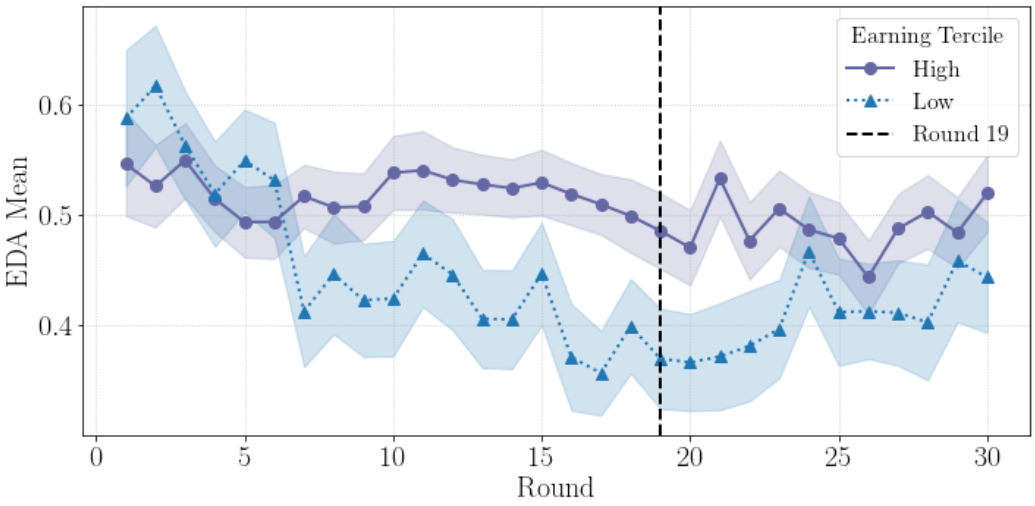
We further explored how the paths of EDA responses for high and low earners compare by estimating separate OLS regressions of $\text{EDA}_t^{\text{ton}}$ on the distance-from-peak measure and its square according to Equation (3.2), where the sign of γ determines the directional curvature of the model.

$$\text{EDA}_t^{\text{ton}} \;=\; \alpha + \beta\tau \;+\; \gamma\tau^2 \;+\; \varepsilon \tag{3.2}$$

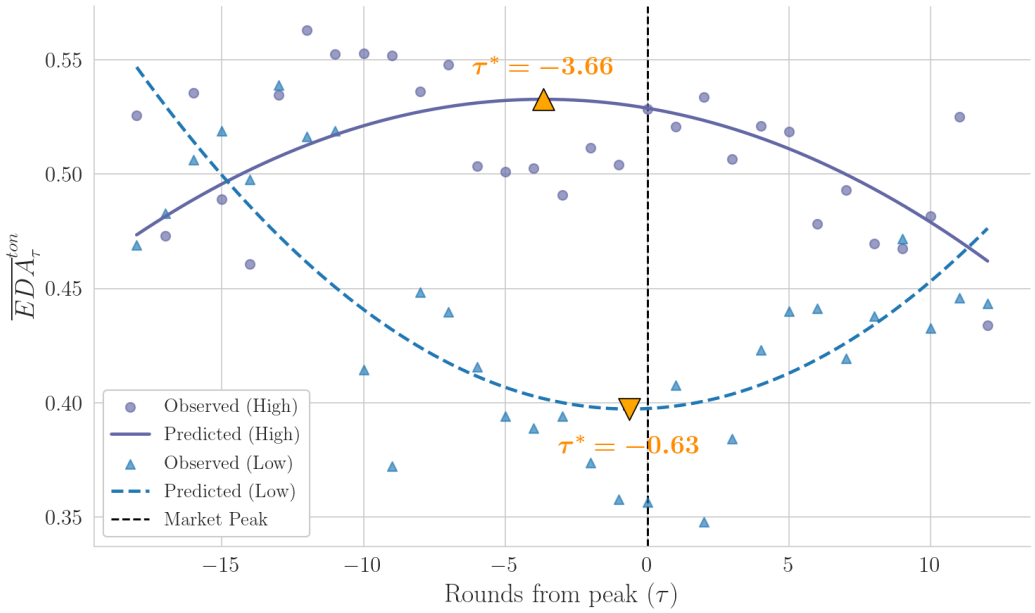
Both regressions explain a substantial portion of the variance in group-mean tonic EDA ($R^2 = 0.545$ for high earners and 0.388 for low earners), suggesting that τ is a meaningful predictor of physiological arousal.[26] The results reveal a striking dissociation between the two groups. For high-tercile participants, the coefficient γ is negative ($\gamma_{\text{hi}} = -0.000288$, $p < 0.001$), yielding an inverted parabola. Solving for the vertex of this fitted model reveals an optimal arousal peak at $\tau^*_{\text{hi}} = -3.663$ (Figure 3.6b). This significantly negative baseline optimum indicates that the arousal peak for high earners substantially precedes the actual market crash, a timeline heavily consistent with heightened vigilance and anticipatory processing among better-performing traders leading up to the market turning point. Conversely, for low-tercile participants, the pattern is reversed: γ is positive ($\gamma_{\text{lo}} = 0.000495$, $p < 0.05$), yielding a U-shaped curve. The calculated minimum for this group occurs much closer to the market crash at $\tau^*_{\text{lo}} = -0.628$. This indicates that arousal among poorer performers is driven primarily by reaction to the market crash.

This dynamic closely mirrors the findings in A. Smith et al. (2014), who observed a nearly identical pattern in neural activity within the anterior insula. Across both their fMRI data and our skin conductance metrics, high earners exhibit an anticipatory increase in arousal leading up to the market crash, whereas low earners display a reactionary increase after the crash occurs. The cross-method consistency of these internal physiological markers stands in notable contrast to the fragility of experimental attempts to induce external emotional states. Specifically, while Andrade, Odean, and Lin (2015) originally concluded that video-induced excitement causally inflates asset bubbles, a high-powered preregistered replication by C. Huber et al. (2024) failed to find any such effect, underscoring the limitations of non-physiological emotional manipulations in bubble environments.

[26] Full regression results are available in the appendix (Table 3.5). Inference presented here is base on heteroskedasticity-robust standard errors



(a)



(b)

Figure 3.6: **(a)** Tonic EDA by earnings tercile around the market peak. Panel A shows mean $\text{EDA}_t^{\text{ton}}$ with 90% confidence intervals for High and Low tercile participants across rounds relative to the market peak ($\tau = 0$). The x-axis is cropped to only include τ bins for which there are 10 or more observations to avoid showing extremely large CI bands. The plot in panel (b) is cropped to match. **(b)** OLS quadratic regressions of mean tonic EDA ($\text{EDA}_t^{\text{ton}}$) on squared distance from the market peak (τ^2) with robust standard errors (Equation (3.2)), estimated separately for High and Low earning terciles. Points are group means; curves are fitted predictions from $\text{EDA}_t^{\text{ton}} = \alpha + \beta\tau + \gamma\tau^2$. Fitted curves were optimized and those optimal points are indicated by large orange triangular markers. Detailed values can be seen in Table 3.5

Table 3.5: Quadratic Regressions of Tonic EDA Around the Market Peak

	High Tercile	**Low Tercile**
Intercept α	52.900***	39.700***
	(0.900)	(1.700)
Distance from peak τ (β)	-0.211***	0.062
	(0.064)	(0.290)
Squared distance τ^2 (γ)	-0.029***	0.050*
	(0.007)	(0.023)
Vertex $\tau^* = -\beta/2\gamma$	-3.66	-0.63
Observations	45	42
R^2	0.545	0.388
Adjusted R^2	0.523	0.356

Notes: The dependent variable is mean tonic EDA $\left(\text{EDA}_t^{\text{ton}}\right)$ multiplied by 100 for readability. Regressions are estimated separately for High and Low earning terciles on binned means by distance from the market peak, $\tau = t - t_p$. Thus, observations denote the number of observed τ bins for each tercile, not participant-round observations. Heteroskedasticity-robust standard errors are reported in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

De-risking and Physiological Arousal

"De-risking" describes the practitioner behavior of reducing portfolio risk. In our experimental setting, the most direct form of de-risking is selling shares of the risky asset. We find that higher physiological arousal is associated with a greater likelihood of de-risking through executed sell orders within a trading period. A logistic regression on all subject–period observations (N = 2,280) shows that higher peak arousal (EDA^{peak}) significantly increases the odds of executing a sell order (see Appendix 3.27). A one–standard deviation increase in peak arousal raises the probability of selling by approximately 9%, evaluated at the sample mean. We also observe differences in de-risking behavior between the high and low earning terciles. By looking at average share positions of the risky asset, we observe that high earners tend to accumulate shares early in the session and begin to offload them prior to the market peak, while low earners tend to gain shares during the post peak crash. (Figure 3.11). This is a pattern of behavior that supports the anticipatory physiological response of the high earners, and the reactionary response of the low earners.

Risk Aversion and Physiological Arousal

We next evaluate whether trial-by-trial physiological arousal is associated with revealed risk preferences. Physiological arousal is measured using the in-period peak ($EDA_{i,t}^{peak}$), while risk aversion is proxied by $\hat{\rho}_{i,t}$, the parameter estimated from participants' choices in the DOSE task. In the DOSE specification used to generate these estimates, utility over gains is given by $u(x) = x^{1-\rho}/(1-\rho)$, so higher values of ρ imply greater risk aversion and lower values imply greater risk tolerance.

We estimate an OLS regression mapping $EDA_{i,t}^{peak}$ with interactions for medium and high earner tercile to $\hat{\rho}_{i,t}$ across 2,219 individual observations, using cluster-robust standard errors adjusted at the subject level (Panel A Table 3.30). The baseline effect (which is the full effect for the low earners) is negative and statistically significant ($\beta_{\text{lo}} = -0.499, p = 0.013$) applies to the participants in the low earning tercile. Because lower values of ρ indicate greater risk tolerance, this negative coefficient indicates that heightened arousal is associated with increased willingness to accept risk among low earners. For medium and high earners the estimated effects are similarly signed but statistically indistinguishable from zero ($\beta_{\text{med}} = -0.213, p = 0.343; \beta_{\text{hi}} = -0.204, p = 0.258$; Panel B, 3.30). So, while low-earners show a propensity to become more risk-seeking in response to physiological arousal, the same response is not shared (or is possibly attenuated) by higher earning participants.

3.7 Models of Forecasting

This section examines the quality and structure of individual price forecasts. In Appendix section 3.17, we also observe that most participants trade consistently with their stated beliefs. We find that forecasts systematically deviate from rational-expectations benchmarks. Despite these departures from rationality, individual forecasting behavior is well described by a parsimonious three-parameter "double adaptive" model, with closely related models performing nearly as well. Moreover, heterogeneity in both forecasting parameters and forecast–action consistency is strongly correlated with participants' realized trading profits.

Evaluating Forecast Rationality

The theory of rational expectations posits that agents form forecasts using all available information, so their predictions should not display systematic errors or predictable biases over time. In the context of experimental asset markets, this means participants' price forecasts at time t should efficiently incorporate all relevant mar-

ket information, including past prices and previous forecast errors, and that forecast errors should not be systematically related to any variables in the agent's information set at the time the forecast was made.

We investigate whether participants in our experiment exhibit forecast biases; specifically, signs of irrationality, similar to those found in previous studies based on economic outlook surveys (Zarnowitz, 1985; Kohlhas and Walther, 2021). Following prior research (Zarnowitz, 1985; Lovell, 1986), we apply three standard tests to assess rationality: the unbiasedness of forecast errors, the absence of autocorrelation in forecast errors, and the lack of correlation between forecast errors and forecasts. These tests allow us to assess whether forecast failures arise from biased beliefs, persistent errors, or internal inconsistency in expectation formation, regardless of how forecasts are translated into trades.

Forecasting Framework and Aggregate Measures

We examine aggregate forecasting behavior by horizon and find systematic patterns: participants tend to underestimate prices at short horizons (current and two periods ahead) while overestimating prices at longer horizons (five and ten periods ahead). Median forecast errors by period are shown in Appendix Figure 3.41.

For each period t and forecasting horizon h, we compute mean and median forecasts across participants. Let I_t denote the set of participants active at period t, with size $N_t = |I_t|$. Participant i's forecast at period t for the price in period $t + h$ is denoted $f_{i,t}^{t+h}$.[27] The aggregate mean and median forecasts are:

$$\bar{f}_t^{t+h} = \frac{1}{N_t} \sum_{i \in I_t} f_{i,t}^{t+h}, \quad \tilde{f}_t^{t+h} = \text{median}_{i \in I_t} \{ f_{i,t}^{t+h} \}.$$

Individual forecast errors are defined as the difference between realized prices and an individual's forecasts:

$$E_{i,t}^{h} = P_{t+h} - f_{i,t}^{t+h}, \tag{3.3}$$

where positive values indicate underestimation and negative values indicate overestimation. Summary statistics for forecasts and forecast errors are reported in Appendix Table 3.35.

[27] The experiment was run in multiple markets. For expositional simplicity, we suppress the market index s; all quantities are defined at the market level.

Rationality Tests

Unbiasedness Test First, we examine whether participants are unbiased across different horizons. Specifically, if participants are unbiased, the distribution of forecast errors at each horizon should center around zero, indicating no systematic bias.

Figure 3.40 in the Appendix illustrates the distribution of forecast errors at each horizon. We observe that as the periods progress ($t \in \{1, \ldots, 28\}$ for $h = 2$; $t \in \{1, \ldots, 25\}$ for $h = 5$; $t \in \{1, \ldots, 20\}$ for $h = 10$), the distribution gradually shifts downward suggesting that participants increasingly over-forecast as periods progress, consistent with systematic bias that becomes more pronounced at longer horizons.

Next, we compute the mean forecast error for each period across different horizons. Figure 4.9 illustrates how the mean forecast error evolves over periods for each horizon. Initially, participants tend to under-forecast prices, but as the periods progress, they gradually shift toward over-forecasting. This pattern is consistent across all horizons, but it emerges earlier for longer horizons.

Before conducting the t-test to examine whether the mean forecast error $E^h_{i,t}$ is zero across periods t for each forecasting horizon h, we first assess whether the normality assumption is satisfied using the Shapiro–Wilk test. The first row of Table 3.6 presents the results. Among the 510 participants, more than half fail the normality test, indicating that the distribution of forecasts for each horizon and individual significantly deviates from normality. Therefore, we employ the Wilcoxon rank-sum test, a nonparametric method that does not assume normality and is well-suited for non-normally distributed data. Compared with a simple sign test, which only evaluates whether the proportion of positive errors equals 50 percent, the Wilcoxon test also incorporates information on the magnitude of errors, thus providing greater power to detect systematic bias. The results are reported in the second row of Table 3.6. When forecasting the price for the current period, more than half of the agents appear to be unbiased. However, as the forecasting horizon increases, the proportion of unbiased agents declines, indicating that forecast bias tends to increase with longer prediction horizons.

Taken together, the results in Table 3.6 indicate that violations of rational-expectations conditions become increasingly prevalent as forecast horizons lengthen, with especially sharp declines in the absence of error–forecast correlation at longer horizons.

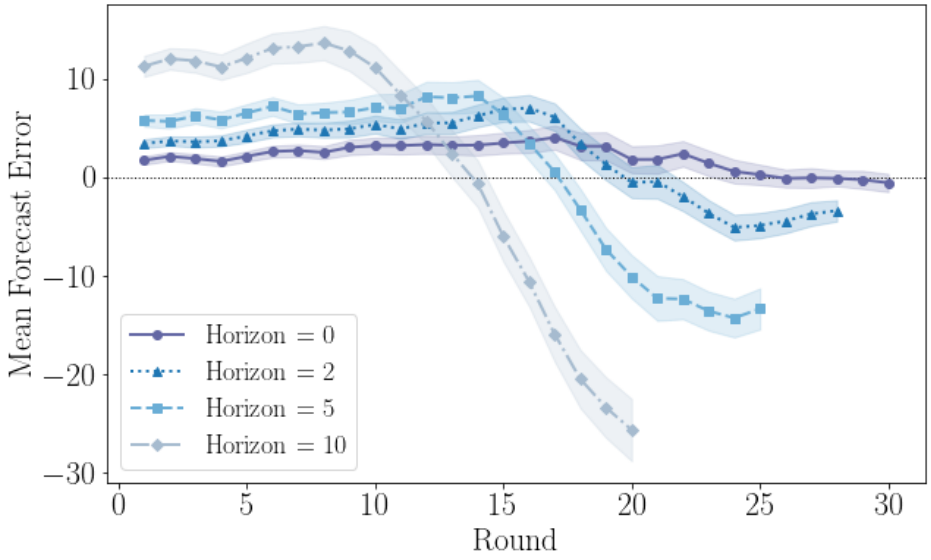


Figure 3.7: Mean forecast error by period for each of the four forecast horizons. As the forecast horizon increases, errors grow because projecting several periods ahead introduces greater uncertainty, causing small expectation errors to compound over longer intervals. 95 percent confidence intervals are shown by shaded regions. As time passes and the bubble begins to peak (and crash), the forecast error becomes more and more negative, indicating that participants are slow to update their price expectation.

Table 3.6: Normality and Rationality Acceptance Rates (%) Across Horizons

Statistic	**Test**	**0**	**2**	**5**	**10**
	Normality	42.5%	45.9%	44.7%	46.5%
$E^h_{i,t} = 0$	Unbiasedness	62.0%	61.8%	59.4%	47.6%
$\beta_1 = 0$ (Eq 3.4)	Zero Autocorrelation	42.2%	–	–	–
$\beta_1 = 0$ (Eq 3.5)	Zero Error–Forecast Correlation	62.5%	34.9%	14.5%	11.7%

Notes: "Normality" refers to results from the Shapiro–Wilk test, with normality rejected at $p < 0.05$; the percentages for each horizon (0, 2, 5, 10) indicate the proportion of agents whose forecast errors are normal. "Unbiasedness" denotes that the mean forecast error is not significantly different from zero ($p > 0.05$), and the reported percentages reflect the proportion of unbiased agents at each horizon. "Zero Autocorrelation" and "Zero Error–Forecast Correlation" indicate agents whose forecast errors exhibit no significant autocorrelation or no significant correlation with forecasts at the 0.05 level, respectively, with the percentages representing the proportion of such agents for each horizon. Scatter plot of β's shown in Figure 3.8

Zero Autocorrelation Test To test the autocorrelation of forecast errors, we examine whether forecast errors from the current period are independent of those in subsequent periods. The underlying principle is that, for rational participants, future

forecast errors should not be systematically related to any information available at the time the forecast was made, including the previous forecast error, which would be part of that information set. We find that, as shown in the third row in Table 3.6, 42.2 percent of agents exhibit no significant autocorrelation in their forecast errors.

Specifically, we estimate the following regression model:

$$E^0_{i,t} = \beta_0 + \beta_1 E^0_{i,t-1} + \epsilon_{i,t}, \tag{3.4}$$

where the null hypothesis is $\beta_1 = 0$. A statistically insignificant β_1 suggests the absence of autocorrelation and supports the assumption of the rational expectations benchmark.

We estimate the coefficient β_1 for each individual using ordinary least squares (OLS), and count the number of agents for whom β_1 is not significantly different from zero ($p > 0.05$). We restrict this test to the current period forecast error ($h = 0$) because forecast errors at longer horizons mechanically overlap across periods, which would induce spurious autocorrelation even under rational expectations.

Correlation Between Forecast Errors and Forecasts The third rationality test examines the correlation between forecast errors and the forecasts themselves. Under rational expectations, forecast errors should be uncorrelated with the forecasts on which they are based.

This condition is based on two related intuitions. First, the forecast $f^{t+h}_{i,t}$ is part of the agent's information set at time t; if any component of this information set can systematically predict future errors, a rational agent would incorporate that information to reduce the errors. Second, a negative correlation implies that when an agent is optimistic (i.e., makes a high forecast), they tend to be overly optimistic, resulting in negative forecast errors. A positive correlation implies the opposite; optimistic forecasts are not optimistic enough.

We compute the correlation between forecast errors and forecasts separately for each forecast horizon by estimating the following regression:

$$E^h_{i,t} = \beta_0 + \beta_1 f^{t+h}_{i,t} + \epsilon^h_{i,t}, \tag{3.5}$$

This regression is estimated separately for each agent and horizon, and we count the number of agents for whom β_1 is not significantly different from zero ($p > 0.05$).

The fourth row in Table 3.6 reports these results. When forecasting the current period ($h = 0$), more than half of our participants (62.5%) exhibit no significant correlation between forecast errors and forecasts, consistent with rational expectations. However, as the forecasting horizon increases, the proportion of agents satisfying this condition declines sharply, falling to just 11.7% at $h = 10$. This pattern indicates that forecast errors become increasingly predictable from agents' own expectations at longer horizons, violating a key implication of rational expectations.

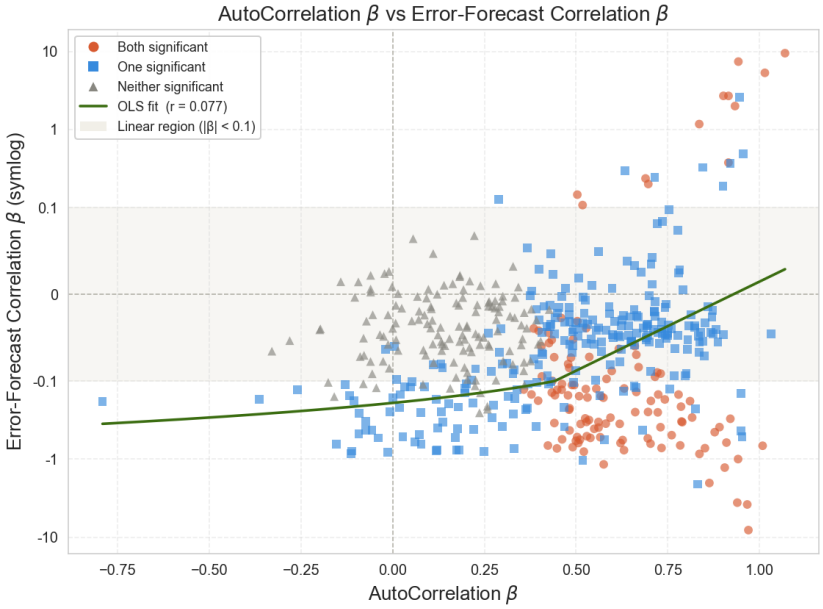


Figure 3.8: Scatter plot of forecast rationality βs; the means of which are presented in Table 3.6. Figure shows that non-significance is distributed around the sample population instead of clustered around a few participants. The vertical axis is presented as a `symlog` scale, which is a logarithmic scale adapted to handle negative numbers and is linear in the light gray band between -0.1 and 0.1. This is why the trend line appears kinked. Joint significance counts can be found in Table 3.7

Table 3.7: Counts of participants by joint significance of the zero auto-correlation β (AC, rows) and the zero error-forecast correlation β (EF, columns).

	EF β		
AC β	Not Sig.	Significant	Total
Not Sig.	140	75	215
Significant	177	118	295
Total	317	193	510

start for condensing

Tests of Competing Models of Forecasts

Next, we assess how well common expectation models account for these forecasts. We test four models: adaptive expectations (Cagan, 1956; Nerlove, 1958), extrapolation (Barberis, Greenwood, et al., 2015; Hirshleifer, Li, and J. Yu, 2015), overreaction (Afrouzi et al., 2023), and double adaptive expectations (Charles A. Holt, Porzio, and Song, 2017).

To avoid redundant notation and methodological descriptions, we define common elements upfront. Let $f_{i,t}^{t+h}$ denote the forecast made by agent i at period t for the price in period $t + h$, after observing the realized price P_{t-1}. Unless otherwise specified, we estimate all model parameters at the individual level using ordinary least squares (OLS) regression for horizon $h = 0$, computing standard errors robust to heteroskedasticity.

Adaptive Expectations The adaptive expectations model is specified as:

$$\text{Adaptive Expectations: } f_{i,t}^{t} = \delta P_{t-1} + (1 - \delta) f_{i,t-1}^{t-1} + \epsilon_{i,t} \tag{3.6}$$

The parameter δ captures the weight agent i assigns to the most recent price. A higher δ indicates a stronger reliance on the latest information relative to past expectations.[28]

The first panel of Figure 3.9 displays the distribution of the estimated δ parameters. Most agents have values close to 1, indicating they place substantial weight on the most recent price when forming expectations. The mean estimate of 0.917 (Table 3.8) further supports this finding. Similar patterns emerge in the Experienced and High-Stakes markets (Figures 3.42 and 3.43), where participants consistently weight the most recent price heavily.

Extrapolation The extrapolative expectations model incorporates the recent price trend into the forecast:

$$\text{Extrapolation: } f_{i,t}^{t} = P_{t-1} + \phi_i (P_{t-1} - P_{t-2}) + \epsilon_{i,t} \tag{3.7}$$

Here, $\phi_i > 0$ measures the degree of extrapolation. Because the coefficient on the lagged price P_{t-1} is fixed at 1, ϕ_i exclusively captures the sensitivity to the most recent price change. A smaller ϕ_i implies weaker trend extrapolation and yields forecasts closer to the last observed price.

[28] This notation is equivalent to $f_{i,t}^{t}$ in previous sections, suppressing the $h = 0$ term.

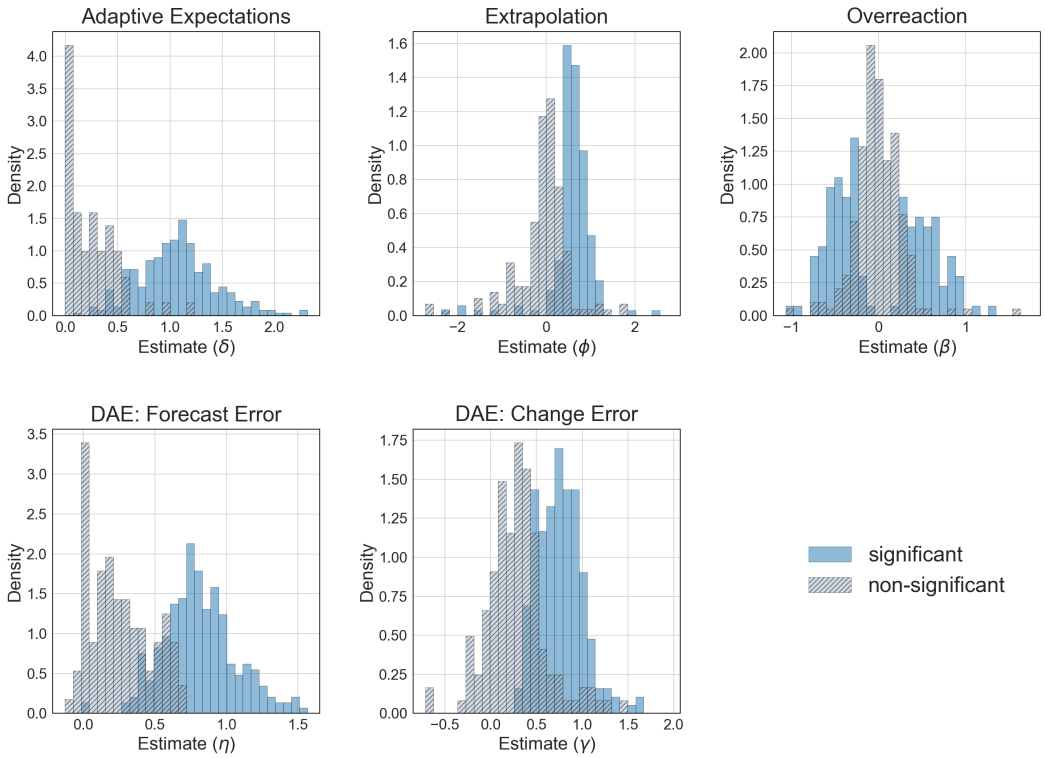


Figure 3.9: Histograms of Parameter Estimates for Different Expectation Models. **Notes:** Different colors represent different levels of statistical significance. Light gray indicates that the estimate is not significantly different from zero ($p > 0.05$), while dark gray indicates that the estimate is significantly different from zero ($p < 0.05$). All models are estimated using OLS regression, with standard errors that are robust to heteroskedasticity.

The second panel of Figure 3.9 shows that most agents have ϕ_i estimates between 0 and 1. The mean estimate of 0.199 suggests a moderate degree of extrapolation in the Standard markets. This extrapolative tendency strengthens in the Experienced markets (mean $\phi = 0.389$) and the High-Stakes markets (mean $\phi = 0.273$).

Overreaction Following Afrouzi et al., 2023, we estimate the overreaction model as:

$$\text{Overreaction: } P_t - f^t_{i,t} = \alpha_i + \beta_i \left(f^t_{i,t} - f^t_{i,t-1} \right) + \epsilon_{i,t} \tag{3.8}$$

The left side of the equation represents the forecast error in period $t+1$, while the parenthetical term on the right denotes the forecast revision between periods t and $t+1$. Because the strict one-period-ahead forecast is unavailable in our experimental design, we approximate it using linear interpolation between the current-period and two-period-ahead forecasts. This approach preserves monotonicity without imposing additional structural assumptions on belief dynamics.

Under rational expectations, forecast revisions should not systematically predict forecast errors ($\beta_i = 0$). A negative β_i indicates overreaction, meaning larger upward revisions lead to negative forecast errors as agents overshoot the realized price. Conversely, a positive β_i reflects underreaction.The third panel of Figure 3.9 reveals an even split among agents between overreaction ($\beta_i < 0$) and underreaction ($\beta_i > 0$). In the Standard markets, the mean estimate is 0.009 with a standard deviation of 0.612. However, clear treatment differences emerge. Experienced markets shift toward underreaction with a mean β of 0.121 (SD = 0.395). High-Stakes markets tilt slightly toward overreaction with a mean estimate of -0.214 (SD = 0.337) and show less dispersion, suggesting that higher financial incentives may help curb extreme expectation errors.

Double Adaptive Expectations (DAE) The double adaptive expectations model accounts for prior errors in both price levels and price trends:

$$\begin{aligned}\text{Double Adaptive: } f_{i,t}^{t} = f_{i,t-1}^{t} &+ \eta \left(P_{t-1} - f_{i,t-1}^{t-1}\right) \\ &+ \gamma \left[(P_{t-1} - P_{t-2}) - \left(f_{i,t-1}^{t} - f_{i,t-1}^{t-1}\right)\right] + \epsilon_{i,t} \end{aligned} \tag{3.9}$$

The coefficient η captures adjustments based on recent forecast errors in price levels, while γ captures adjustments based on errors in price trends. As the specific forecast $f_{i,t-1}^{t}$ is not directly measured, we proxy it using the average of the current-period and two-period-ahead forecasts from period $t-1$.

The fourth and fifth panels of Figure 3.9 display the parameter distributions for the adaptive level and adaptive trend, respectively. Most estimates are positive and bounded between 0 and 1. The mean adaptive level parameter is 0.648, and the mean adaptive trend parameter is 0.572 (Table 3.8).

We observe distinct heterogeneity across treatments. Agents in Experienced markets rely more heavily on level adjustments (mean $\eta = 0.772$) and less on trend changes (mean $\gamma = 0.483$). High-Stakes markets show the inverse pattern, with lower sensitivity to level errors (mean $\eta = 0.608$) and higher responsiveness to trend deviations (mean $\gamma = 0.644$). This indicates that experience and financial incentives alter how agents weight levels versus trends when forming expectations. As shown in Figure 3.10, the DAE model's fitted predictions are highly correlated with actual participant forecasts, indicating a strong model fit.

Table 3.9 shows parameter estimates across the three experimental conditions.

Table 3.8: Summary Statistics of Parameter Estimates and Fit by Forecasting Model

Model (Parameter)	Mean	SD	Mean R^2	N
Adaptive Expectations (σ)	0.917	0.479	0.501	354
Extrapolation (ϕ)	0.199	1.019	0.301	353
Overreaction (β)	0.009	0.612	0.168	359
DAE: Level Error (η)	0.648	0.361	0.679	358
DAE: Trend Error (γ)	0.572	0.471		

Notes: Mean and SD denote the mean and standard deviation of the OLS model parameter estimate across participants for each model, respectively. Mean R^2 reports the average OLS goodness-of-fit across participants, and N indicates the number of participants with non-missing estimates. The slight variation in sample size N across models is due to differences in model specifications (e.g., lag structures), which prevented estimation for a small subset of participants with insufficient observations. For the Double Adaptive Expectations model, results are reported separately for the level forecast error and trend forecast error parameters.

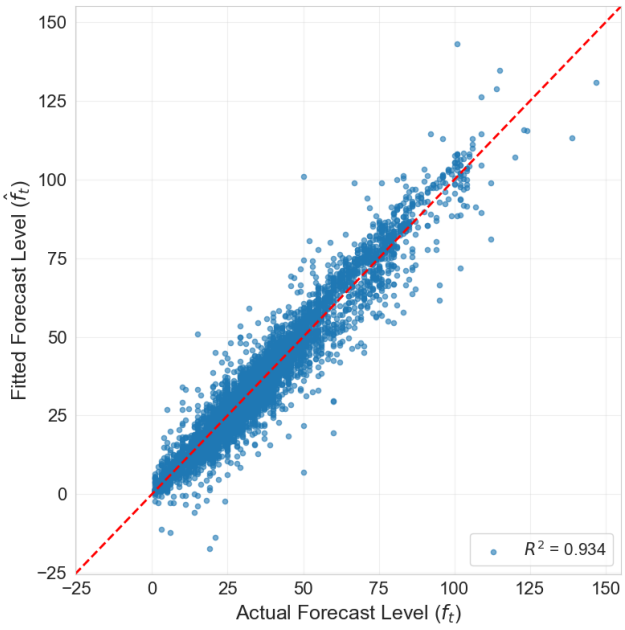


Figure 3.10: Double Adaptive Model: Actual vs. Fitted Forecasts.
Notes: In the appendix, we check robustness and compare realized forecast levels to the model's fitted predictions under different representations: Figure 3.44 forecasts expressed as a percentage of lagged price; Figure 3.45 high- vs. low-earners; and Figure 3.46 pre- vs. post-crash periods. The dashed line indicates the 45-degree line where fitted values equal observed values.

Table 3.9: Expectation Model Parameters Across Markets

Model	Param.	Std.	Exp.	HS	Std.–Exp.	Std.–HS	Exp.–HS
AE	δ	0.917	1.036	0.746	0.027	0.005	< 0.001
EXT	ϕ_i	0.196	0.381	0.268	0.009	0.826	0.734
OR	β_i	0.009	0.121	-0.214	0.033	< 0.001	< 0.001
DA-L	η	0.648	0.772	0.608	0.004	0.441	0.009
DA-T	γ	0.572	0.483	0.644	0.037	0.370	0.055

Notes: Reported values are cross-participant means from individual-level OLS estimates with heteroskedasticity-robust standard errors. The final three columns report pairwise Welch two-sample p-values comparing participant-level parameter estimates across market conditions. Market abbreviations are Std. = Standard, Exp. = Experienced, and HS = High-Stakes. Model abbreviations are AE = adaptive expectations, EXT = extrapolation, OR = overreaction, DA-L = double adaptive level adjustment, and DA-T = double adaptive trend adjustment. Parameter interpretations are as follows: δ is the weight on the most recent price; ϕ_i is sensitivity to recent price trends; $\beta_i < 0$ indicates overreaction and $\beta_i > 0$ indicates underreaction; η adjusts to level forecast errors; and γ adjusts to trend forecast errors. A value of $\delta > 1$ implies overweighting the most recent price relative to prior expectations.

- **AE** is the adaptive expectations model. The parameter δ measures the weight placed on the most recent price when updating forecasts. Values above one indicate overweighting of the most recent price relative to prior expectations.
- **EXT** is the extrapolation model. The parameter ϕ_i measures sensitivity to recent price trends; larger values indicate stronger trend-following expectations.
- **OR** is the overreaction model. The parameter β_i captures adjustment to prior forecast errors; $\beta_i < 0$ indicates overreaction, while $\beta_i > 0$ indicates underreaction.
- **DA-L** is the level component of the double adaptive model. The parameter η measures adjustment to level forecast errors.
- **DA-T** is the trend component of the double adaptive model. The parameter γ measures adjustment to trend forecast errors.

Detailed parameter estimates for the Experienced and High-Stakes markets are reported in Appendix Tables 3.36 and 3.37, reinforcing the treatment-specific patterns discussed above.

As a more demanding joint test, we regress both experimental bonus measures on all five expectation-formation parameters simultaneously (Table 3.10). For each bonus type (Market & Forecast), we ran a LASSO regression against the five forecast model parameters and a battery of controls. Most controls came from survey questions

and include self-reported risk preferences and demographic data. Other controls like average DOSE ρ and total number of orders placed came from the participants' choices during the experiment. Each LASSO selection was followed up by a post-selection OLS regression with robust standard errors on the selected regressors. The five forecast model parameters are considered key variables in these models and so are present in the post-selection regressions regardless of whether they were selected by the LASSO. Any regressor present in Table 3.10 that LASSO shrank to zero are indicated by a dagger (†).

Only the adaptive and extrapolation parameters where significant in predicting forecast bonuses. For the market bonus, none of the forecast parameters were significant, though the extrapolation (ϕ), overreaction (β), and double adaptive change error (γ) parameters did survive the LASSO process. These parameters show more significance when predicting bonuses in the absence of controls (Table 3.38), but do not seem to be robust. We repeated this analysis but restricting the data to only the in-lab participants with biometric data. To these regressions, we included the participants' average tonic EDA activity as well as the standard deviations. Neither of these measures were selected by the LASSO regression. If fact, all of the forecast parameters where shrunk to zero by the LASSO.

We also find that these high earners understood the experiment better (based on the score of the post-instruction quiz), were more rational in following their forecasts, and consisted of a higher proportion of Caltech students.

These joint regressions that heterogeneity in forecasting behavior are correlated with the variation in participant payoffs. The adaptive model serves as the most robust predictor of the forecast bonus, while the overreaction model consistently explains the market bonus. The weaker evidence for extrapolation after controlling for session differences suggests that unobserved session-level factors may drive much of that variation.

3.8 Conclusion

This paper studies the role of psychophysiological and behavioral factors in the formation and collapse of asset price bubbles in controlled experimental markets with constant fundamental value. We combine high-frequency trading data with measures of (EDA), elicited risk preferences, price forecasts, and post-experiment textual responses to examine how beliefs, arousal, and behavior jointly shape market outcomes. To our knowledge, no prior study has examined how physiological

Table 3.10: Market & Forecast Bonus Explained by Forecast Parameters and Controls

Dependent Variable	(1) MB	(2) FB	(3) MB w/ EDA	(4) FB w/ EDA
Adaptive (α)	-0.045†	0.089**	-0.031†	-0.052†
	(0.057)	(0.038)	(0.166)	(0.119)
Extrapolation (ϕ)	0.050	0.038***	0.287†	0.098†
	(0.037)	(0.014)	(0.177)	(0.159)
Overreaction (β)	0.094	0.015†	-0.010†	0.127†
	(0.065)	(0.024)	(0.211)	(0.122)
DAE Forecast Error (η)	0.017†	-0.027†	-0.175†	0.026†
	(0.057)	(0.023)	(0.202)	(0.145)
DAE Change Error (γ)	0.098	0.006†	-0.196†	0.144†
	(0.073)	(0.040)	(0.267)	(0.191)
DOSE ρ	0.040	0.042***	-	-
	(0.026)	(0.014)		
Net Orders	-	-	-	-0.044
				(0.027)
Net Volume	0.019**	-	-	-
	(0.008)			
Attempted Trades	-0.005***	-	-	-
	(0.001)			
Executed Trades	-	-0.003*	-	-
		(0.002)		
Quiz Grade	-	-0.062***	-	-0.042
		(0.012)		(0.048)
Age	-0.003	-	-	-
	(0.002)			
General Risk (Pre)	-0.043***	-	-	-
	(0.014)			
Invest 100 (Pre)	0.001	0.001*	-	-
	(0.001)	(0.000)		
Own Stock	-0.051	-	-	-0.090*
	(0.037)			(0.048)
Self Place	-	-	-0.004**	-
			(0.002)	
Invest 100 (Post)	-	-	-	-0.001
				(0.001)
N	274	274	55	55
R^2	0.253	0.248	0.185	0.351

Notes: LASSO with 5-fold cross-validation was used to select control variables. The five forecast-parameter variables are retained in all post-selection OLS specifications regardless of LASSO selection because they are the main variables of interest. Coefficients are from post-selection OLS with heteroskedasticity-robust standard errors in parentheses. † indicates forecast-parameter variables that LASSO would have set to zero but that were retained by design. Dependent variables are Market Bonus (MB) and Forecast Bonus (FB), transformed by log(1 + bonus). Adaptive expectations, extrapolation, overreaction, and double adaptive expectation parameters are individually estimated from their respective forecasting models and then used as regressors. Columns (1)–(2) show all participants. Columns (3)–(4) show only in-lab participants with biometric data and include EDA tonic mean and EDA tonic standard deviation, which were not selected by LASSO. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

arousal, beliefs, and trading behavior jointly interact to generate and unwind bubbles in experimental markets with a constant fundamental value.

At the aggregate level, we show that market dynamics are strongly driven by standard financial predictors, particularly lagged returns and near-term expectations. Measures of tonic physiological arousal do not improve period-by-period return prediction once these variables are included. However, when examined over longer horizons, smoothed measures of market-level arousal are systematically associated with future price movements and crash likelihood. Periods of elevated tonic arousal are followed by higher forward returns (See Figure 3.33), while periods of lower arousal precede weaker returns and heightened instability. These results suggest that physiological measures capture slower-moving market states rather than high-frequency price innovations.

At the individual level, physiological arousal is closely linked to trading behavior and performance heterogeneity. Elevated arousal increases the probability of de-risking behavior within a trading period, while lower volatility in tonic arousal is associated with higher market earnings. In addition, arousal is negatively correlated with elicited risk aversion at the trial level, consistent with arousal influencing risk-related decision-making rather than prices directly. Taken together, these findings indicate that physiological signals operate through behavior (portfolio adjustments, timing, and execution) rather than through direct price forecasting.

We also observed differing patterns of tonic arousal around the market turning point. High-earning participants exhibit an anticipatory peak in arousal that occurs well before the market crash. Conversely, low earners demonstrated a large increase in arousal starting just before the market peak and continuing until the end of the session. This divergence closely mirrors the neural findings of A. Smith et al., 2014 in the anterior insula and aligns with evidence from Lo and Repin, 2002 linking emotional reactivity to poorer trading returns. Given recent high-powered replication failures of prominent behavioral market phenomena (e.g., C. Huber et al., 2024), the cross-method robustness of these physiological markers is notable. Crucially, while prior studies relied on restrictive neuroimaging (fMRI) or complex field setups, capturing these dynamics with relatively inexpensive EDA monitoring wristbands provides a highly scalable, low-cost alternative.

Our analysis of expectations reveals systematic deviations from rational benchmarks that grow with the forecast horizon. Participants tend to underestimate prices early in the market and overestimate prices near the peak, with forecast errors becoming

increasingly biased, autocorrelated, and predictable as horizons lengthen. Formal rationality tests show that the fraction of forecasts consistent with rational expectations declines sharply with horizon. Importantly, these deviations are economically meaningful: participants whose forecasts are more accurate earn significantly higher market payoffs and reduce exposure earlier in the price cycle, whereas lower-earning participants tend to increase exposure into downturns.

We further examine several canonical models of expectation formation. Adaptive, extrapolative, and double-adaptive models explain a substantial portion of individual forecast variation, and the estimated parameters from these models are significantly related to market earnings. In joint regressions, forecast parameters collectively explain a nontrivial share of cross-sectional variation in trading performance, despite modest explanatory power at the individual level. These results indicate that forecast heterogeneity is structured rather than random and that how participants form expectations has economically relevant consequences.

Finally, we test whether bubbles in our setting can be explained by time-varying risk premia. Despite eliciting individual risk preferences in each period, we find no meaningful relationship between contemporaneous risk premia and asset prices. This evidence suggests that bubble dynamics in our experiment are not driven by rational updating of risk preferences but instead arise from belief dynamics and behavioral responses.

Overall, our findings indicate that price dynamics in experimental asset markets are shaped by the interaction of beliefs, behavior, and emotional states, even though traditional financial variables remain central for short-horizon prediction. Physiological and behavioral measures provide complementary information about market conditions, participant heterogeneity, and the timing of instability. These results contribute to the literature on belief formation and limits to rationality in asset markets and motivate further work on how behavioral and biological signals interact with financial decision-making in more complex market environments.

Predicting Returns in High-Stakes and Experienced Markets Extended

Table 3.11 repeats the same OLS regressions predicting return as for the Standard markets, except that there are no EDA recordings in these markets.

The OLS model explains roughly 60% of the variance in next-period returns ($R^2 = 0.603$). As seen in the standard markets, the lagged return, R_{t-1} predict the majority

Table 3.11: High-Stakes and Experienced markets: Predicting Future Returns R_t (returns $\times 100$)

	High Stakes			Experienced		
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	-0.158	4.269	3.429	-1.483***	2.622	4.092
	(0.395)	(4.672)	(2.510)	(0.269)	(4.806)	(4.974)
R_{t-1}	0.716***		0.479**	0.744***		0.532***
	(0.089)		(0.083)	(0.085)		(0.081)
ρ		-3.359	-1.925		1.563	-2.350
		(1.616)	(1.056)		(6.801)	(4.730)
V_{t-1}		-0.022	-0.128		-0.153	-0.077
		(0.215)	(0.140)		(0.317)	(0.288)
$30 - t$		0.163	0.088		0.107	0.087
		(0.193)	(0.095)		(0.195)	(0.101)
Ω_t		0.012	-0.061		-0.023	0.005
		(0.090)	(0.051)		(0.079)	(0.064)
Υ_t		-0.006	-0.003		-0.006	-0.004
		(0.021)	(0.014)		(0.018)	(0.011)
f_t^t		0.532*	0.332**		0.904**	0.303
		(0.180)	(0.096)		(0.226)	(0.258)
f_{t-2}^t		-0.225	-0.168		-0.889***	-0.401
		(0.153)	(0.085)		(0.181)	(0.210)
f_{t-5}^t		-0.442***	-0.194		-0.158	-0.033
		(0.074)	(0.115)		(0.076)	(0.042)
Observations	96	96	96	120	120	120
R^2	0.503	0.482	0.587	0.572	0.521	0.629
Adjusted R^2	0.497	0.435	0.544	0.569	0.487	0.599

Notes: Dependent variable: future log return R_t (scaled by 100). Standard errors (in parentheses) are heteroskedasticity- and cluster-robust (clustered by market).

of the variance, but even when R_{t-1} is omitted, the forecasts have similar predictive power. [29] It is notable that predictability of returns is similar for all three types of markets.

3.9 Extended Literature Review

Time-varying risk premia

Time-varying risk premia have long been posited as a key mechanism underlying asset price bubbles. The concept followed soon after the onset of the efficient market hypothesis, where Shiller found that asset prices fluctuate more volatily than their fundamental values. Further work by Fama (1984) and Fama and Bliss (1987) uncovered systematic, time-varying patterns in excess returns across the term structure, stylized facts that any dynamic model must explain. Follow-up yield-curve regressions by Campbell and Shiller (1991) confirmed that these movements could not be reconciled with the traditional rational expectations hypothesis, pointing instead to sizable and predictable fluctuations in term premia. Taken together, this evidence firmly established that risk compensation was state-dependent rather than constant over time and that risk premia of financial assets appears to fluctuate over time.

A central unresolved debate concerns what drives time variation in expected returns: changes in risk preferences (or risk-bearing capacity) versus changes in beliefs about future payoffs (Cochrane, 2011). Under the first view, predictable excess returns arise because the representative (or marginal) investor becomes more risk averse in downturns (or faces tighter intermediary constraints), raising required compensation for holding risky assets; in these models, prices fluctuate largely through discount-rate variation (Campbell and Cochrane, 1999; Cochrane, 2011). Under the second view, predictable returns may instead be driven by belief dynamics—learning, disagreement, extrapolative expectations, or shifts in perceived probabilities of rare disasters—so that fluctuations in expected returns reflect changing expectations about fundamentals and risks rather than changes in preferences (Hong and Stein, 2007; Gabaix, 2012; Greenwood and Shleifer, 2014b).

Subsequent work in macro-finance has sharpened the empirical distinction between these competing mechanisms by linking each to different observable proxies. A prominent approach measures time-varying risk aversion (or the price of risk) using option-implied quantities such as the variance risk premium, which isolates

[29] We ran a model that mirrors the pre-registered regression except for the EDA since it is not collected from online-only subjects. Table 3.19

investors' willingness to pay for protection against future volatility and has strong predictive power for aggregate equity returns (Bollerslev, Tauchen, and Zhou, 2009). A different literature measures belief dynamics using survey-based expectations and disagreement, showing that dispersion in forecasts captures slow-moving shifts in market beliefs and helps explain predictable variation in expected returns, particularly in fixed income markets (Kim and Orphanides, 2007). This measurement-based distinction reflects a broader state-of-the-art view: time-varying risk premia are generated by both preference-based and belief-based channels, but the literature increasingly operationalizes them separately to evaluate their relative quantitative importance.

Modern empirical work increasingly attempts to separately identify these channels, recovering evidence of time-varying both risk aversion and beliefs in household and institutional portfolio choice settings.

Extended to experimental markets, as participants update their perceptions of risk, whether due to changing economic conditions, policy shifts, or psychological factors, their required compensation for bearing risk may also fluctuate. These changes in required compensation (risk premia), which can arise either from belief dynamics about the amount of risk or from shifts in risk tolerance or risk-bearing capacity, can generate asset price fluctuations independent of fundamentals, potentially leading to mispricing and bubbles.

Empirical investigations, such as those by (Cohn et al., 2015) and (Guiso, Sapienza, and Zingales, 2018), provide supporting evidence that individual risk assessments can vary substantially over time in response to evolving market uncertainties. Collectively, this body of work reinforces the view that time-varying risk premia are a rational driver of bubble phenomena. Brandt and K. Q. Wang, 2003 - also found TVRA in stock and bond markets in response to inflation.

Real-world high-Stakes environments, such as financial system rescues, executive compensation, and asset price bubbles, often require insights that extend beyond field data. Empirical validation of whether High-Stakes influence behavior can take months or years for suitable real-world data to emerge. Which is why lab experiments are ideal incubators for simulating various behaviors, including those that occur when higher stakes are at play. By varying payoff schemes and framing within the lab, researchers can efficiently test theories about High-Stakes decision-making without waiting months or years for suitable field data to become available.

Empirical evidence on the influence of large incentives outside the laboratory is mixed. Using brokerage account records, Sui and B. Wang, 2023 shows that investors exhibit more substantial biases and perform worse in higher-stakes real accounts than in lower-stakes simulated accounts, contrary to theoretical predictions that High-Stakes should improve performance. The authors find that across tasks conducted in both the United States and India, very high reward levels often lead to trading performance declines. This is in line with R. Yu, 2015, who argue that, contrary to the predictions of efficiency-wage models, extremely high monetary incentives can impair rather than enhance productivity (i.e., "choking under pressure"), mainly due to high psychological stress.

Higher incentives can also influence how beliefs are formed, not merely the decisions themselves. Ambuehl (2017) shows that when information acquisition is costly, larger incentives lead participants to select and interpret information in a manner that favors taking action. In asset market experiments, findings are mixed. Cui, Feltovich, and Zhang, 2022 demonstrates that both relative-performance bonuses and loss-framed payment schemes independently drive the price of a risky asset above its fundamental value. Other studies find different results. Enke et al., 2023 argues that, even when stakes are raised from none to standard to very high, there is only a modest impact on reducing cognitive biases across four tasks—testing base-rate neglect, anchoring, and failure of contingent thinking.

Given the mixed and limited findings on the impact of higher incentives on asset bubble formation, we conduct a secondary analysis in section 3.4 to investigate the effect of elevated stakes on bubble formation within our experimental framework.

3.10 High vs Low Earner Behavior in experimental markets

There is substantial evidence that asset-market outcomes often reflect systematic differences between traders with more information or stronger task-relevant skills and those with less expertise (Brunnermeier and Nagel, 2004).

Our hybrid lab–online design provides new leverage on this issue. In our study, we analyze the differences in behavior and biometrics that exist between our highest-earning participants and our lowest-performing participants. The results are from exploratory analyses. We found from analysis of shareholding patterns that we saw three clusters of behavior: high earners bought shares early and sold later, medium earners did not make many transactions, and low earners sold early and bought late. Because of this, we split our participant analysis into these three groups. These

earning terciles are determined within each market, so it is possible that a medium earning participant in one market might actually earn a larger market bonus than a high earner in another. Indeed, we do observe some overlap in earnings by tercile (Table 3.29 in the Appendix). We also analyze the strategic differences between our high and low earner participants using NLP techniques in 3.25 in the Appendix.

Trading Behavior High vs Low Earners

Our goal in these analyses is to understand the behavioral differences that exist between high and low earners. Figure 3.11 shows that the top market-earnings median tended to accumulate shares early in the experiment, and then steadily reduce their holdings in later market periods, which suggests that most of the trading was done between the high and low earners. This unloading behavior intensified just before the market peak. In contrast, the low earning participants showed the opposite pattern: they gradually increased their share holdings throughout the market's rise and continued to do so even through the later market periods. This behavior may suggest that high earning participants may anticipate the market's run-up and crash more effectively than low earning participants. As shown in the Appendix, we analyzed the strategy differences outlined by the participants in the different terciles; high earning participants are much more likely to adopt a strategy of "buy early, sell later" (Figure 3.35).

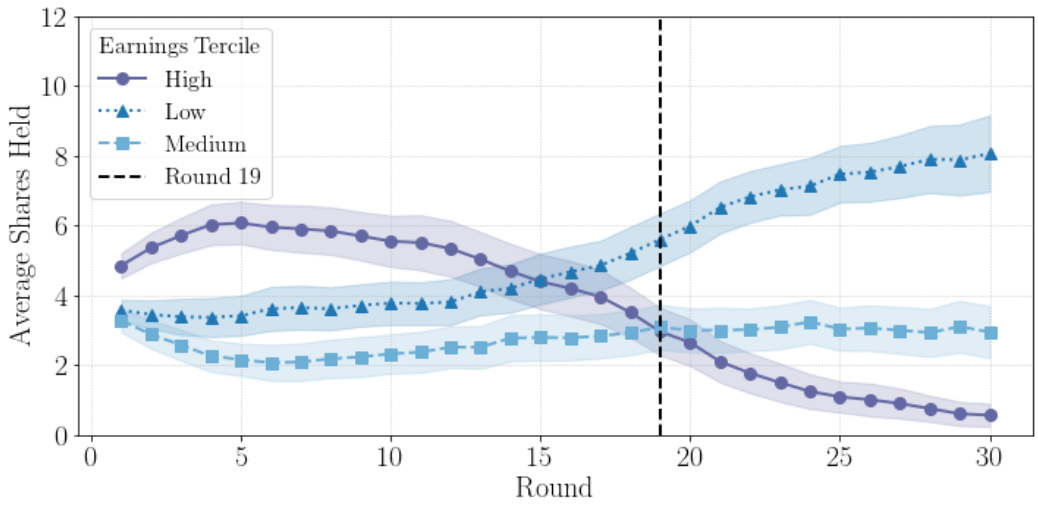


Figure 3.11: Average shares held over 30 periods for participants grouped by market-earnings tercile (low, medium, high). The high-earning group accumulated shares early in the market and sold them off later. The lowest-earning group increases their holdings late in the market, particularly after the average peak period.

Forecast Behavior of High and Low Earners

The idea that the highest earning participants were better able to anticipate market patterns is substantiated in our results, which show that high earning participants are significantly more accurate forecasters in all time horizons, as shown in Figure 3.12 and Table 3.12.

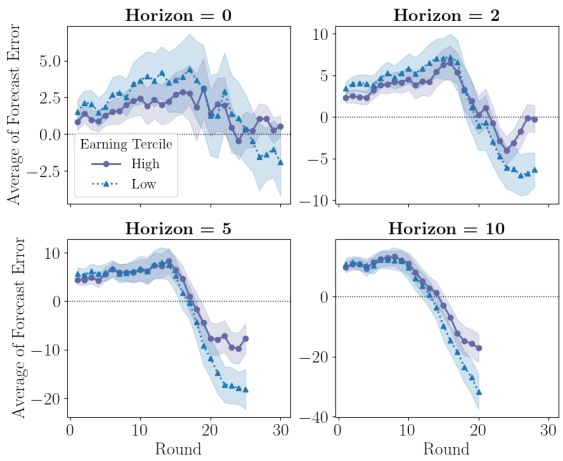


Figure 3.12: Mean forecast error for the highest and lowest earning terciles. Shaded regions represent 95 percent confidence intervals.

Table 3.12: Forecast Error Comparison

H	High	Low	p-value
0	7.553	12.064	0.023
2	10.167	13.990	0.021
5	15.071	19.978	0.003
10	22.754	27.512	0.016

Note. Mean squared error was calculated for each participant and each horizon (H). Participants were then grouped by earnings tercile, and t-tests were conducted to compare the MSE measure between the high and low earners.

Together, these findings suggest that higher earnings are not merely the result of luck or random variation but reflect a systematic forecasting advantage. Specifically, high earning participants appear to generate more accurate forecasts of future prices, especially at longer horizons, which in turn may enable them to strategically accumulate and offload shares in anticipation of market peaks and crashes.

Peak Eda EDA^{peak} High Earners vs Low Earner

We also wanted to understand whether there were different arousal patterns between the highest earners and the lowest earners. Are our higher earners calmer than our lower earners? To investigate, we ran a mixed linear model comparing the period-by-period EDA^{peak} levels of individuals in the highest-earning tercile with those in the lowest-earning tercile. We found a trend $p < 0.06$ showing that higher earners, contrary to our expectation, actually operated closer to their EDA^{peak} throughout the experiment than lower earners [30]. The data suggest that higher earners operate

[30]Note that this measure is standardized by participant, so it is possible that this effect is because higher earners never reach the EDA spike seen in low earners

closer to their maximum arousal state than lower earners, although this relationship appears to be reversed in the early periods of the experiment.

3.11 Subject Differences In-Lab vs. Prolific

Laboratory participants were Caltech students, who tend to score highly on standardized academic measures, while online Prolific participants come from a broader population with more heterogeneous backgrounds. Because of this difference, we performed multiple analyses to quantify if/and by how much the trading behavior differed between these two groups. This allows not only to understand our data better but also to gain some insight (albeit somewhat crudely) whether differences in measurable analytical skills, rather than education labels, are associated with systematic differences in trading behavior. We find that Caltech students statistically outperformed the prolific students in the market bonus, forecast bonus, and on the quiz. Table 3.13 summarizes the differences between the two groups of in-lab and Prolific participants. Below is the demographic information for each of our markets.

Table 3.13: Summary statistics comparing survey responses from participants based on site.

	Age		Risk Gen Pre		Risk Gen Post		Risk Fin. Pre		Risk Fin. Post		Inv Pre		Inv Post	
Site Session	L	P	L	P	L	P	L	P	L	P	L	P	L	P
100hm05f	30.33	40.76	5.33	4.53	5.00	4.35	3.00	4.12	2.33	4.12	66.67	53.53	58.33	54.12
19jfynhh	27.00	34.50	3.80	4.80	4.50	5.00	2.80	3.80	3.75	4.67	57.00	43.00	60.00	54.44
3olyo3ti	31.60	37.94	**2.40**	**4.06**	2.00	3.29	3.00	3.94	**1.67**	**3.65**	44.00	47.18	30.00	39.53
898aphuw	30.00	31.25	2.67	3.42	3.00	3.58	2.67	3.75	3.00	3.50	40.00	49.58	47.50	45.83
8h0lckuw	**24.86**	**41.85**	4.00	3.85	4.00	3.64	3.29	4.23	3.17	4.27	48.57	87.08	42.50	58.00
b4opeq8h	28.50	36.94	4.00	3.89	2.50	4.27	2.00	3.22	2.00	3.40	63.50	46.39	65.00	48.93
ea9wnj36	34.40	35.82	**3.20**	**5.00**	3.80	3.90	3.80	4.73	3.60	3.90	40.00	38.27	42.00	31.50
j9yi8qx8	**23.67**	**39.93**	3.50	3.64	3.17	3.29	2.50	2.71	2.17	2.64	34.17	49.29	30.83	47.29
jtf9f20p	**22.40**	**37.00**	3.00	4.00	3.00	3.44	3.40	3.78	3.60	3.67	68.00	45.44	**83.00**	**42.78**
lrx4o1bb	34.67	38.53	4.00	4.53	3.83	4.43	3.67	4.80	3.50	4.71	51.67	54.33	50.00	50.71
m48xf7cp	38.25	31.39	4.00	4.22	3.50	3.75	3.25	4.06	3.00	3.44	37.50	38.89	40.00	44.38
mhuck8jp	**24.00**	**37.33**	3.50	4.33	3.25	3.89	3.25	4.17	3.75	3.89	75.00	59.17	65.00	57.22
nhrerg4w	**21.83**	**33.27**	4.00	4.36	4.40	3.80	4.50	3.64	4.20	3.40	**100.00**	**47.73**	**100.00**	**42.50**
pmyan048	**20.33**	**35.89**	3.67	4.53	4.00	4.25	3.33	4.21	3.33	4.12	40.00	51.58	50.00	52.19
rcnn52km	30.00	38.00	4.33	3.50	4.33	4.00	3.00	3.38	3.33	4.14	50.00	48.12	66.67	59.29
txyy1leq	**20.00**	**37.94**	**4.75**	**3.06**	4.00	2.80	**4.25**	**2.81**	3.25	2.53	46.25	25.94	43.75	21.67
wvjdln40	**24.00**	**39.46**	3.00	3.77	3.50	3.77	2.80	3.23	3.00	3.69	45.00	40.00	46.25	44.23
yh2bjr4n	—	—	—	—	—	—	—	—	—	—	—	—	—	—
z1vrkziq	23.50	35.50	4.00	4.06	3.50	3.81	2.50	4.00	**2.00**	**3.81**	66.25	57.25	53.75	55.69

Note. Bolded values indicate a significant difference at the 5% level. Site abbreviations: L = Lab, P = Prolific.

The market bonuses in a given market were all correlated since there is a predeter-

Table 3.14: Summary statistics comparing market performance measures from participants based on site (Lab L or Prolific P.

Site Session	Quiz Grade L	Quiz Grade P	Market Bonus L	Market Bonus P	Forecast Bonus L	Forecast Bonus P	Risk Bonus L	Risk Bonus P	Attempted Trades L	Attempted Trades P	Executed Trades L	Executed Trades P	DOSE r L	DOSE r P	DOSE μ L	DOSE μ P
100hm05f	4.00	3.06	3.45	3.43	0.02	0.01	1.55	1.23	50.33	35.59	12.00	8.82	-0.03	0.38	3.09	3.09
19jfynhh	**4.00**	**2.17**	4.44	2.65	**0.05**	**0.01**	0.95	1.14	33.80	30.45	**12.60**	**7.45**	0.34	0.81	3.06	3.17
3olyo3ti	3.80	2.53	**4.85**	**2.93**	0.55	0.49	0.78	0.82	29.00	33.47	9.60	7.71	0.55	0.79	2.90	3.08
898aphuw	3.75	3.17	3.53	3.35	0.04	0.04	0.75	1.00	27.25	33.25	10.50	12.25	0.80	0.76	3.37	3.28
8h0lckuw	0.00	0.00	8.37	5.44	**1.38**	**0.87**	2.00	2.00	33.71	40.43	7.71	10.14	0.46	0.61	3.10	2.89
b4opeq8h	3.00	2.17	3.95	3.34	0.06	0.04	0.95	0.83	35.50	39.89	11.00	10.61	0.88	0.56	3.14	3.15
ea9wnj36	3.40	2.38	4.22	2.92	0.01	0.01	1.02	0.91	32.20	42.00	9.00	12.50	0.59	0.39	3.16	3.09
j9yi8qx8	**3.50**	**2.29**	**4.23**	**3.02**	0.60	0.47	0.95	1.07	41.00	36.00	11.33	8.79	0.83	0.76	3.13	3.02
jtf9f20p	**4.40**	**1.80**	**3.74**	**3.21**	0.02	0.02	1.01	0.59	28.20	50.90	**6.00**	**14.60**	-0.01	0.40	2.99	2.92
lrx4o1bb	3.33	3.19	3.39	3.44	0.06	0.05	0.85	0.77	38.17	38.06	10.67	9.81	0.41	0.52	3.10	3.05
m48xf7cp	3.00	2.79	2.14	2.89	0.02	0.01	0.69	0.75	42.50	33.21	12.50	11.00	0.29	0.70	2.82	3.12
mhuck8jp	3.50	2.75	1.36	2.25	0.03	0.02	0.31	0.64	30.00	35.58	8.75	10.25	0.52	0.25	2.78	3.07
nhrerg4w	**4.33**	**2.92**	3.61	3.22	**0.05**	**0.02**	0.61	1.02	27.17	32.82	7.83	9.09	**-0.36**	**0.73**	**2.90**	**3.27**
pmyan048	4.20	3.53	2.89	3.21	0.00	0.01	0.92	0.95	35.60	36.79	12.80	11.05	0.99	0.66	3.12	3.04
rcnn52km	3.50	3.22	3.70	3.14	0.01	0.02	1.16	1.10	31.50	26.89	8.50	7.44	0.11	0.54	3.04	3.02
txyy1leq	3.40	3.12	2.88	3.54	0.02	0.02	0.75	1.03	41.80	34.75	9.00	9.00	0.83	0.90	3.15	3.08
wvjdln40	4.20	2.93	3.35	3.07	0.05	0.03	1.10	0.82	37.60	41.36	12.00	12.29	0.51	0.34	3.14	3.18
yh2bjr4n	3.00	2.00	4.35	2.93	0.31	0.19	1.36	0.99	30.20	34.09	11.80	7.56	1.02	1.00	3.18	3.15
z1vrkziq	**4.00**	**2.12**	**4.28**	**2.86**	0.02	0.02	0.59	1.06	**19.50**	**47.12**	8.25	14.88	0.97	0.31	3.04	3.10

Note. Bolded values indicate a significant difference at the 5% level. Site abbreviations: L = Lab, P = Prolific. **Quiz** is the score on the initial attempt at the graded comprehension quiz. **Mkt Bonus** is the market bonus earned during trading. **FC Bonus** is the forecast bonus. **Risk Bonus** is the result of the randomly selected DOSE trial. **Att Trades** is the number of attempted trades. **Exe Trades** is the number of executed trades. The **DOSE** parameters are the final two columns.

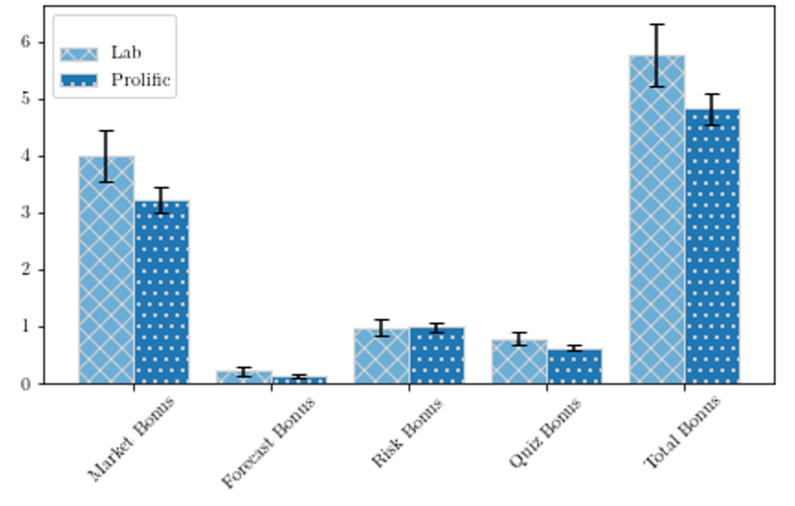


Figure 3.13: Plot of mean bonuses by site. Standard markets only. 95 percent confidence intervals are shown.

Table 3.15: Mean bonuses by group with p-values and 95 percent confidence intervals.

Bonus Type	μ Lab	μ Prolific	p-value	95% CI
Market Bonus	3.976	5.144	0.041	[-1.846, -0.489]
Forecast Bonus	0.218	0.125	0.004	[0.007, 0.179]
Risk Assessment Bonus	0.983	0.991	0.903	[-0.152, 0.135]
Quiz Bonus	0.787	0.682	0.035	[-0.006, 0.216]
Total Bonus	5.763	6.852	0.068	[-1.855, -0.322]

Notes: Mean bonuses by group with p-values and 95% confidence intervals. μ_{Lab} and μ_{Prolific} denote mean bonuses for in-lab and Prolific online participants, respectively. The p-value is from a two-sample Welch's t-test of $H_0 : \mu_{\text{Lab}} = \mu_{\text{Prolific}}$, which does not assume equal variances. The 95% confidence interval is for the difference $\mu_{\text{Lab}} - \mu_{\text{Prolific}}$.

mined amount of wealth in the market. Therefore, a high bonus for one participant or a group of participants must result in a low bonus for another participant or group. Because the two groups (in-lab vs. prolific) are not i.i.d., a simple t-test will not suffice to compare the relative performances.

We instead employed a bootstrapping technique in which the market bonuses for the participants within each market were randomly shuffled, and the new group means were compared. This gave us a distribution against which we could compare our observed market bonus means. Figure 3.14 shows the results of 20,000 iterations of this exercise. Assuming that the bootstrapped distribution of group mean differences is Student's t with 19,999 degrees of freedom, the observed difference ($0.763) was significantly different from the bootstrapped mean of $0.099 ($p = 0.002$).

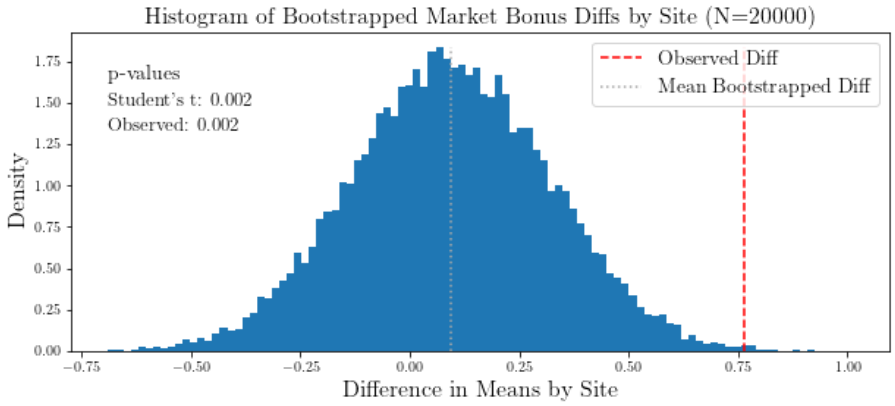


Figure 3.14: Distribution of the difference in means of the market bonus by site (μ = \$0.099; σ = 0.226; df=19,999). Market bonuses were shuffled within each market. We then calculated the difference in the means of market bonuses by site. The observed difference of \$0.763 is shown by a red dashed line. The p-value of this comparison was calculated by calculating $1 - \text{cdf}(\mu, \sigma, \text{df}) = 0.002$ (labeled Student's t). We also computed the p-value by calculating the proportion of the bootstrapped observations that are greater than the observed difference (labeled "observed").

3.12 Design Details

Experimental Pages

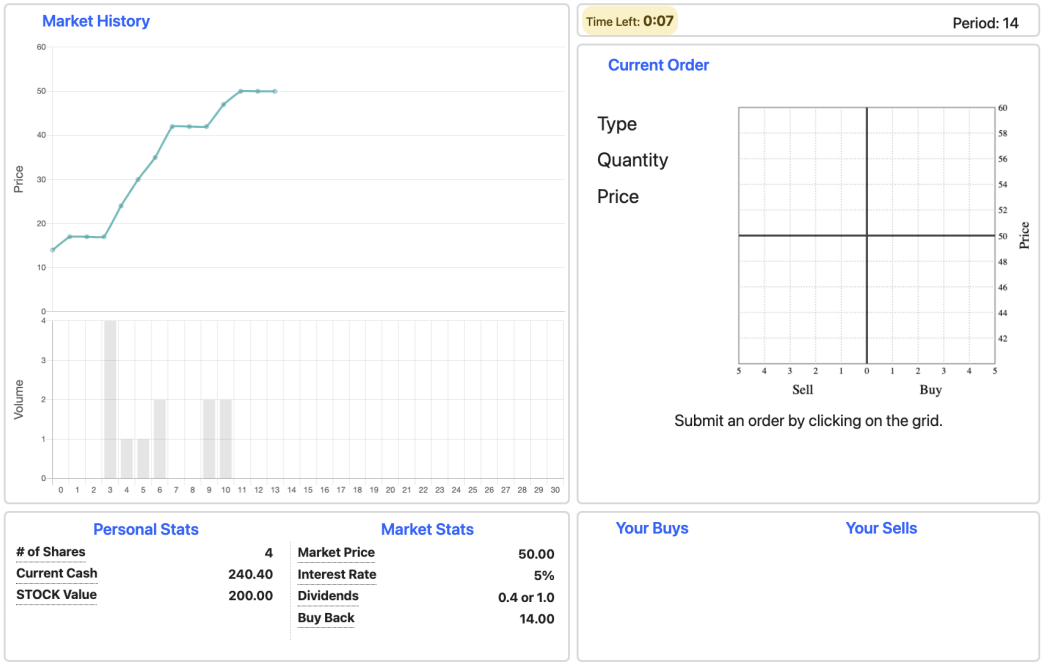


Figure 3.15: Order submission page (20 seconds). Participants are asked to submit the price and volume for their limit orders using the grid.

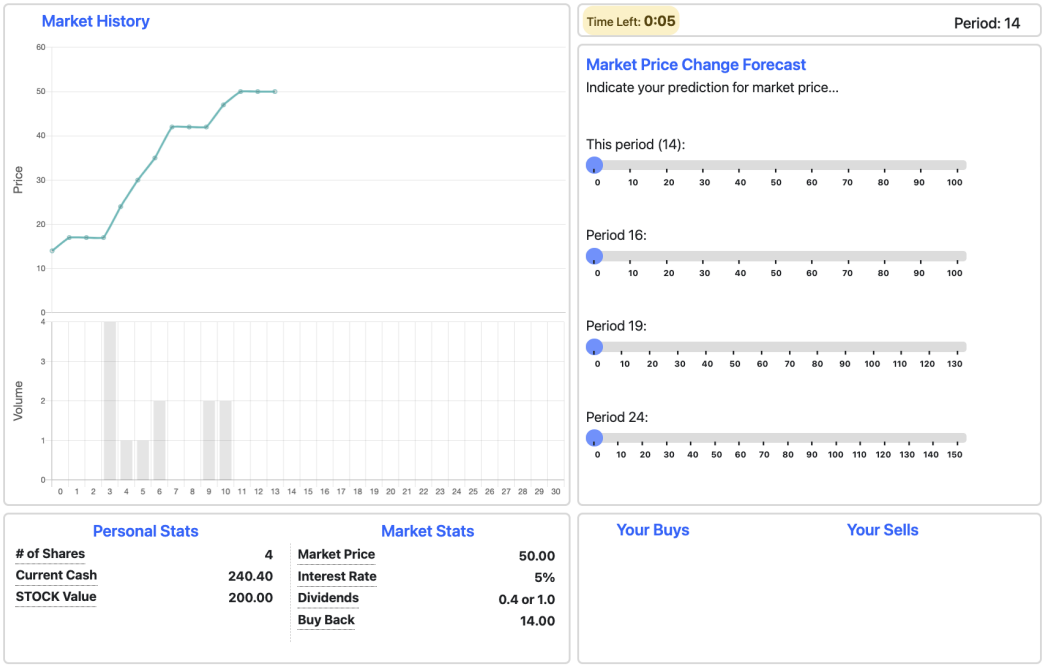


Figure 3.16: Forecast page. (30 seconds) Participants are asked to submit their estimation for the current market period and periods 2, 5, and 10 ahead.

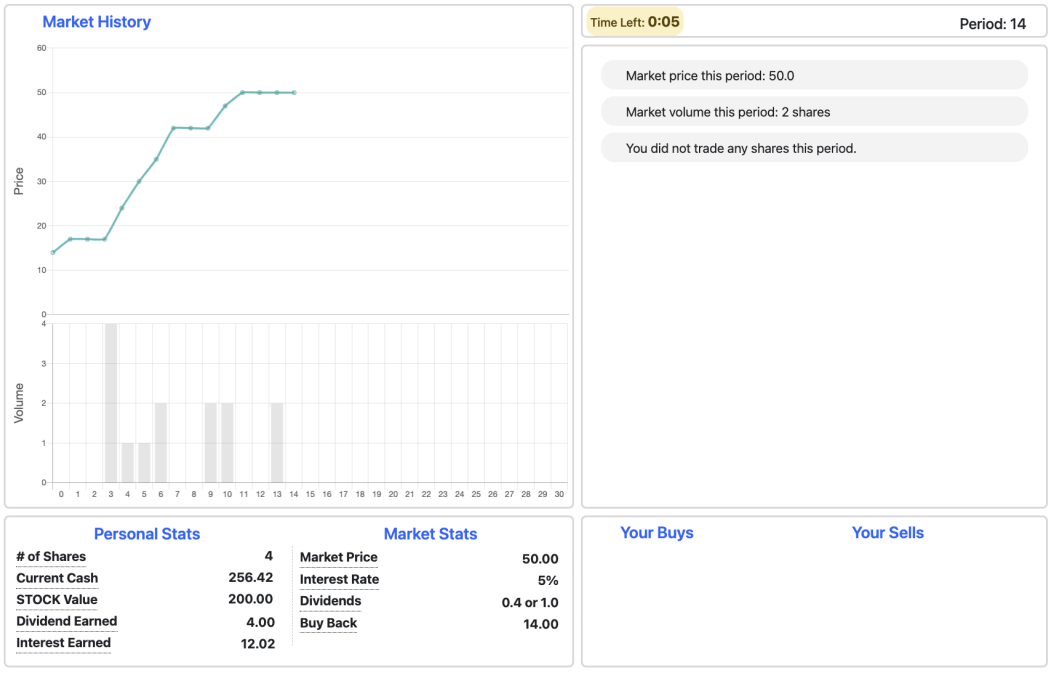


Figure 3.17: Round Results page. (10 seconds) Participants see the actual realized price for that period, the total number of shares traded, and the number of shares they traded.

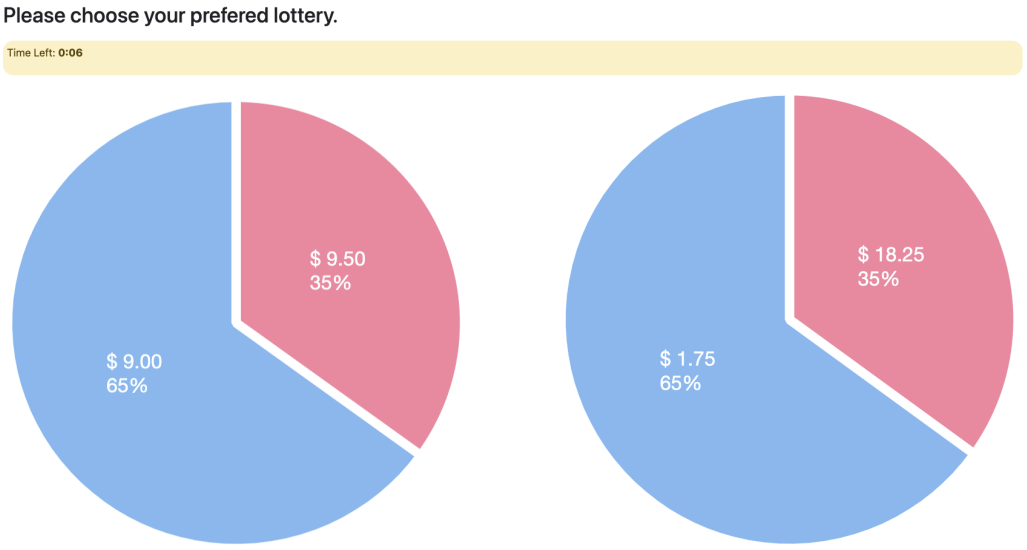


Figure 3.18: Risk Elicitation Page. (20 seconds). Participants are shown four pairs of a risky lottery, visually displayed as a pie chart, and a less risky lottery, and asked to choose between them. After each selection, the probabilities involved are adjusted according to the DOSE protocol (Chapman et al., 2024) so that the next choice offers the most information about the participants' risk preferences.

Price Formation

For every market s and period t, the market price is determined as the price that maximizes volume. A bid submitted by participant i in period t of market s is a tuple consisting of price and quantity, (p, q). Participants may submit up to three bids per period. The set of all bids submitted in a given market and period is represented by $B_{s,t}$. Likewise, the set of all asks is denoted by $A_{s,t}$, with similar notation for individual asks, (p, q).

Let $P_{s,t}$ be the set of all prices submitted as bids or asks,

$$P_{s,t} = \{p \mid (p, q) \in B_{s,t} \cup A_{s,t}\}, \tag{3.10}$$

We define the cumulative buy quantity Q^B available at a given price p as the total quantity of all bids at or above price p; formally:

$$Q^B(p) = \sum_{(p',q)\in B_{s,t}} q \cdot 1_{p' \geq p} \tag{3.11}$$

This represents the total quantity of shares that all participants are willing to purchase at a given price.

We define the cumulative sell quantity Q^A available at a given price p as the total quantity of all asks at or below price p; formally:

$$Q^A(p) = \sum_{(p',q)\in A_{s,t}} q \cdot 1_{p'\leq p} \tag{3.12}$$

For a given price p, volume is determined by

$$V(p) = \min(Q^B(p), Q^A(p)). \tag{3.13}$$

The market price is then

$$p^*_{s,t} = \arg\max_{p\in P_{s,t}} V(p). \tag{3.14}$$

All prices are rounded down to the nearest integer value, a consideration made to simplify the user interface.

If there is no intersection between the bid and ask, the mid-point between the highest bid and the lowest ask is reported as the market price for that period.

Payment Details

- A base payment of $20 and an average bonus of approximately $10.
- Bonus payment is calculated based on: (i) participants' trading results at a rate of 200 experimental CASH units to $1, (ii) forecasting ability (i.e. each price forecast that is within 2.5 units of actual price is rewarded with five CASH), (iii) a randomly chosen lottery that participants select from the experiment and (iv) the quiz result (participants receive $0.25 for each of the five comprehension questions answered correctly.

3.13 Bio-metric measure details

- Electrodermal activity (EDA), also known as skin conductance data, refers to the measurement of the electrical conductance of the skin, which varies with its moisture level. This physiological parameter is directly influenced by the autonomic nervous system, particularly the sympathetic nervous system, and is often used as an indicator of psychological or physiological arousal. When a person experiences emotional stress or excitement, the sweat glands in the skin become more active, resulting in increased skin conductance. EDA data are collected using sensors attached to the skin, which detect these changes in conductance. The data are released in the form of continuous signals that

reflect the variations in skin conductance over time. Analyzing these signals can provide insights into a person's emotional state, stress levels, and overall autonomic activity. The Empatica E4 devices that we used sampled EDA at a rate of 4Hz.

The EDA signal is preprocessed at a participant level by first winsorizing it to 99 percent and normalizing it to fall within 0 and 1 (inclusive) after (Cordes, Nolte, and Schneider, 2023). This step manages outliers and allows us to compare all participants. Further preprocessing and feature extraction were done using the `NeuroKit 2` package available from Python (Makowski et al., 2021). The EDA signal was then decomposed into phasic (rapid, transient changes) and tonic (slow, baseline level) components by using a built-in NeuroKit2 routine that performs these steps internally. We used the `NeuroKit 2` default cutoff frequency settings of 0.05 Hz using a 5th-order Butterworth filter, thereby allowing only frequencies higher than 0.05 Hz to pass through.[31]

- Heart Rate (HR) data refers to the measurement of the number of heartbeats per minute, which serves as an essential indicator of cardiovascular health and autonomic nervous system activity. It is influenced by factors such as physical activity, emotional states, and overall health. When a person experiences stress, excitement, or physical exertion, the heart rate typically increases as part of the body's autonomic response. EmbracePlus uses a total of 9 LEDs and four photodiodes to record the photoplethysmography (PPG) at the wrist. Hence, it records a peripheral and mechanical signal of cardiac activity, so it is more precise to refer to its measure of cardiac activity rate as pulse rate rather than heart rate. All wrist-worn biometric devices that use the PPG really are measuring pulse rate rather than heart rate, and FDA-cleared devices that use a similar technology usually refer to 'pulse rate' measurements. PPG is a non-invasive optical technique used to measure heart rate by detecting changes in blood volume in the microvascular bed of the skin. Since the PPG is a measure of blood circulation (which depends on the heart's pumping quality), it reflects the mechanical activity of the heart rather than the electrical activity of the heart.

- Blood Volume Pulse (BVP) data refers to the measurement of the change in blood volume within the microvascular bed of tissue. It is typically detected

[31] All results are robust to frequency cutoffs of 0.01 Hz as well.

through PPG sensors placed on the skin. BVP is a measure of the cardiovascular system's activity, specifically the pulsatile component associated with each heartbeat. By illuminating the skin and detecting the intensity of light that either passes through or is reflected by the blood vessels, BVP sensors can capture variations in blood flow that correspond to cardiac cycles. These fluctuations allow researchers to infer heart rate and heart rate variability, both of which are valuable indicators of autonomic nervous system function, particularly the balance between sympathetic and parasympathetic activity. The BVP signal is obtained from our Empatica Devices by a proprietary algorithm (of Empatica) that combines light signals of multiple specific wavelengths to illuminate tissue, and the reflected (or transmitted) light is then detected. The variations in that light—caused primarily by changes in arterial blood volume—yield the PPG signal. It has a fixed sampling rate of 64 Hz (64 times per second).

- Temperature (TEMP) refers to the measurement of a participant's body temperature, measured in °C, at the wrist. Fluctuations in skin temperature can provide insights into an individual's physiological state. For example, changes in skin temperature might correlate with stress levels or cognitive efforts that may occur during trading.

3.14 Experimental Procedure

1. Participants sign a consent form agreeing to participate in the experiment.

2. Participants fill in a short socio-demographic questionnaire where they provide information on their age, gender, study background, work experience, and level of English.

3. Subjects watch a video with the experimental instructions on how to trade in the software. Participants were provided with written instructions and had to respond to five comprehension questions on the experimental software. If participants could not answer at least one question correctly, they were unable to participate. This was done to ensure that we only allowed participants who understood the instructions of the experiment.

4. Participants then logged in to the market to perform the actual trading experiment. In-lab participants will have their biometric data measured using

either an Empatica E4 SCR device or a newer (acquired April 2024) Empatica EmbracePlus device.

5. Participants were then asked to respond to a post-experimental survey where they were asked to qualitatively outline their strategy.

3.15 Individual Market and Forecast Plots

Comparison of Descriptive Statistics Across Experimental Treatments

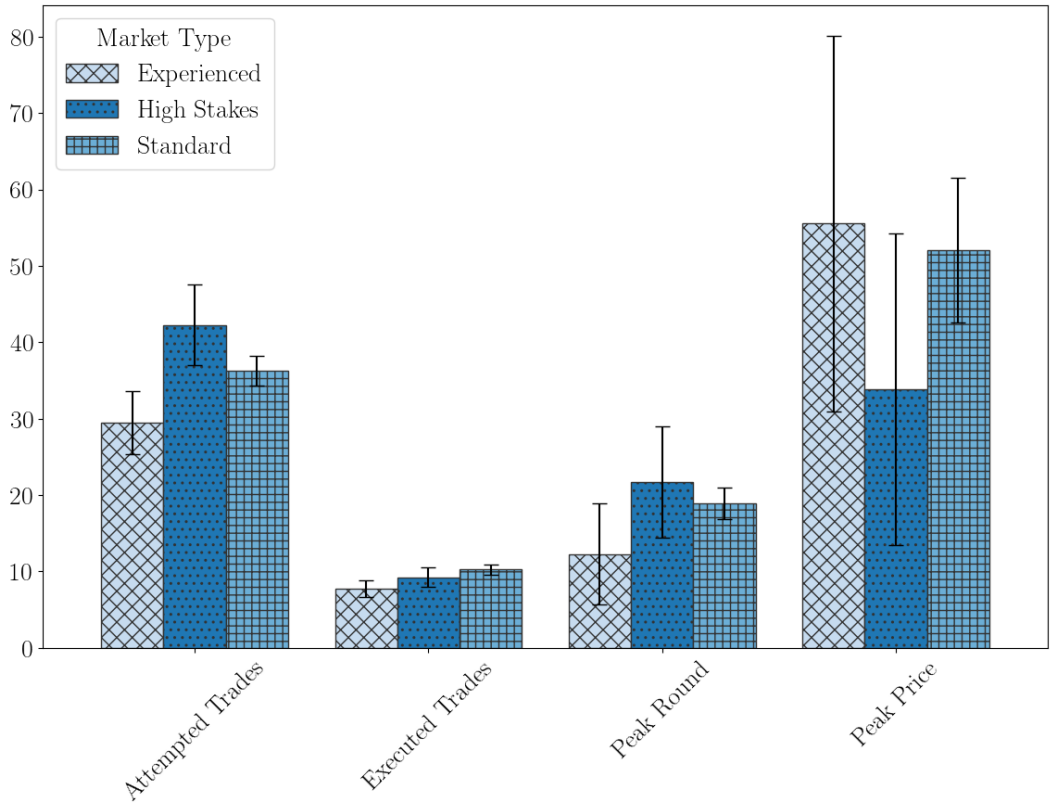
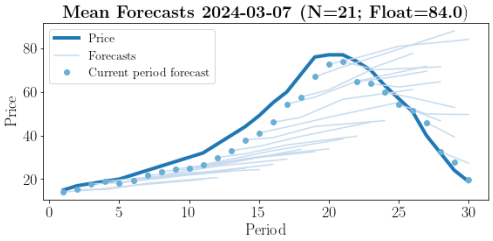
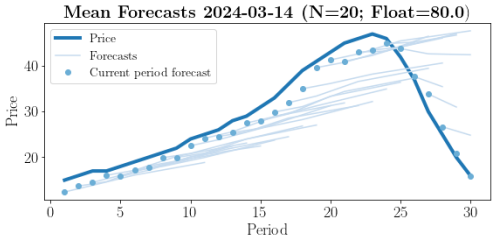


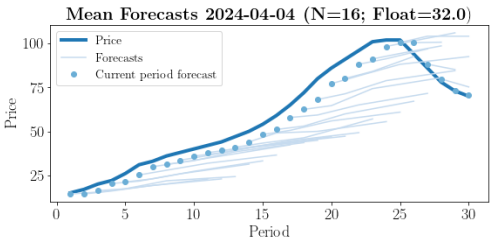
Figure 3.19: A comparison of market measures by market type. *Note.* 95 percent confidence intervals shown. Because there are only four High-Stakes markets and five Experienced, the tests are too weakly powered to detect significant differences. Notice the standard error bars are larger for High-Stakes and Experienced statistics compared to the standard. Attempted (Executed) Trades refer to the average number of trades attempted (or executed) per participant in each market.

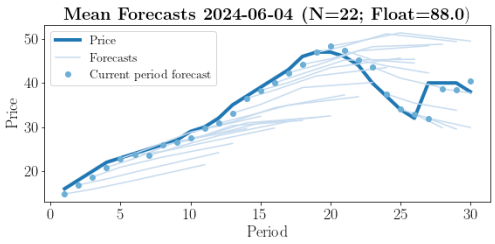
Standard Market Plots

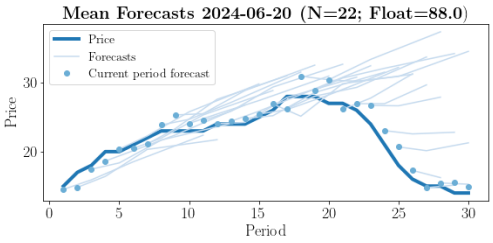
We ran nineteen Standard markets.

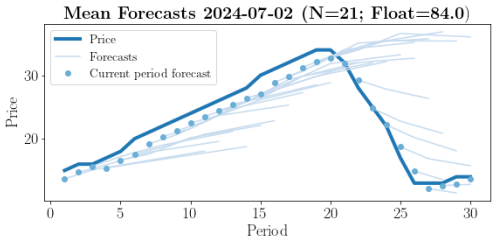

Mean Forecasts 2024-03-07 (N=21; Float=84.0)
Price
Forecasts
Current period forecast
Price
Period

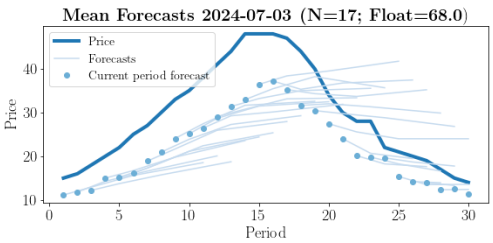

Mean Forecasts 2024-03-14 (N=20; Float=80.0)
Price
Forecasts
Current period forecast
Price
Period

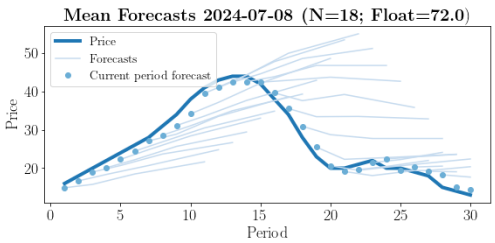

Mean Forecasts 2024-04-04 (N=16; Float=32.0)
Price
Forecasts
Current period forecast
Price
Period

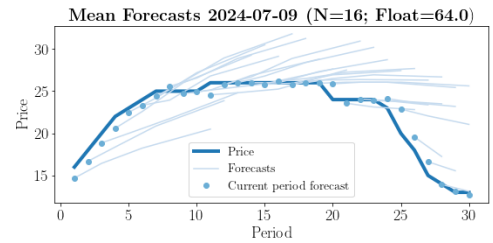

Mean Forecasts 2024-06-04 (N=22; Float=88.0)
Price
Forecasts
Current period forecast
Price
Period

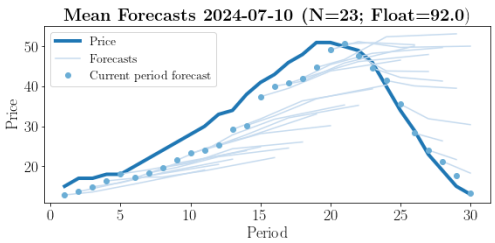

Mean Forecasts 2024-06-20 (N=22; Float=88.0)
Price
Forecasts
Current period forecast
Price
Period


Mean Forecasts 2024-07-02 (N=21; Float=84.0)
Price
Forecasts
Current period forecast
Price
Period


Mean Forecasts 2024-07-03 (N=17; Float=68.0)
Price
Forecasts
Current period forecast
Price
Period


Mean Forecasts 2024-07-08 (N=18; Float=72.0)
Price
Forecasts
Current period forecast
Price
Period


Mean Forecasts 2024-07-09 (N=16; Float=64.0)
Price
Forecasts
Current period forecast
Price
Period


Mean Forecasts 2024-07-10 (N=23; Float=92.0)
Price
Forecasts
Current period forecast
Price
Period

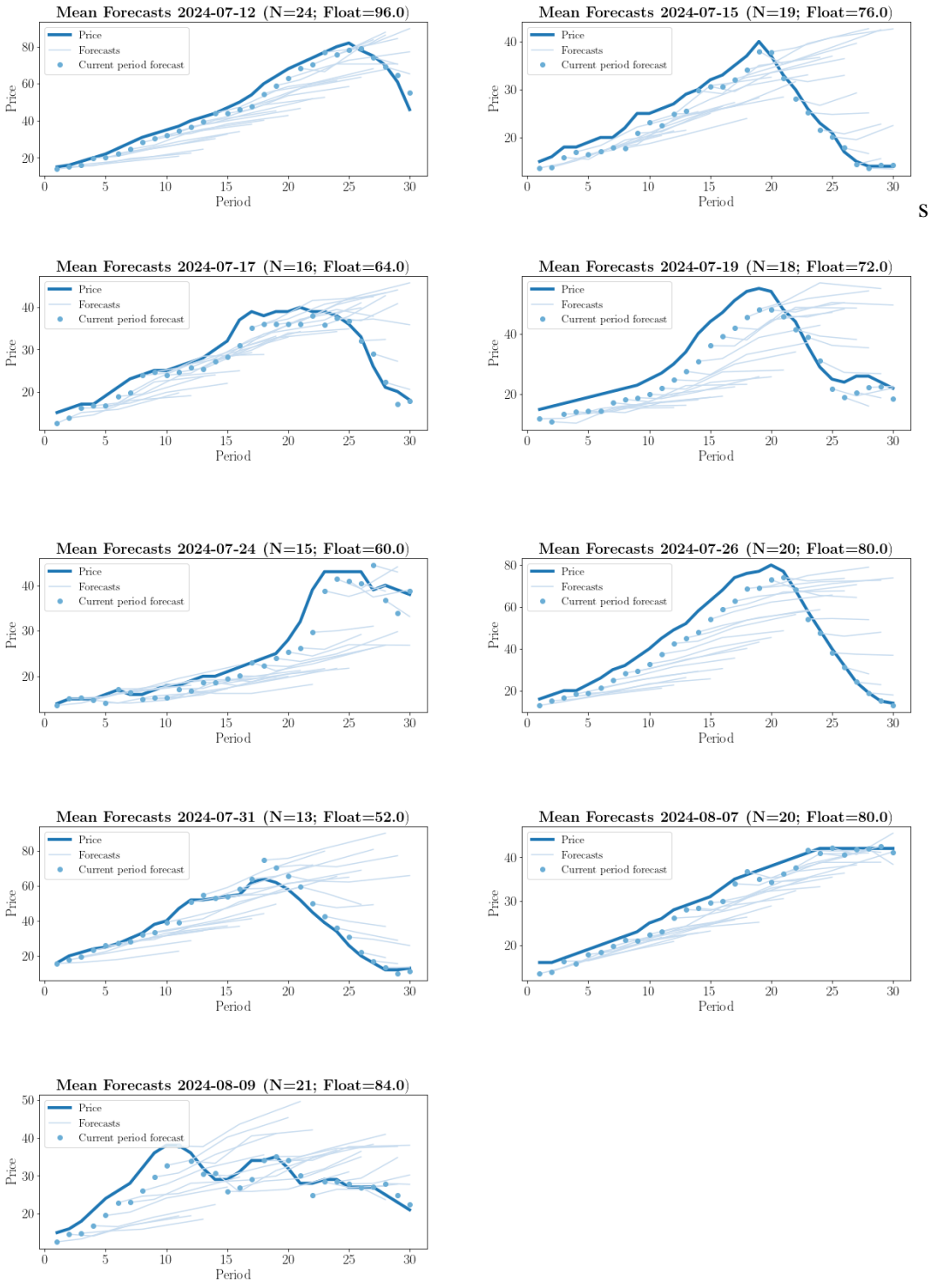
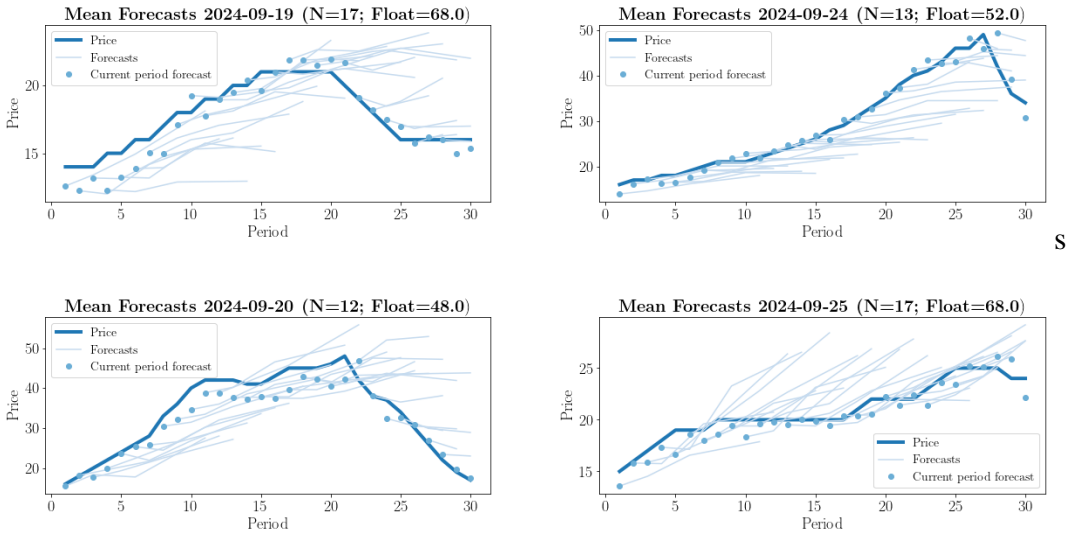


Figure 3.21: 19 Standard markets. The lines represent the dynamics of price levels for 30 periods. Each point represents the current period forecast, and subsequent lines represent average price forecasts for the 0-, 2-, 5-, and 10-period ahead periods. Note: There was an error in Market (2024-04-04); participants were only endowed with an average of two shares of stock rather than a set number of four. The qualitative performance was similar to that of our markets; therefore, we combined it into our analysis.

High-Stakes Markets Plots

We ran four High-Stakes markets. Each market had a payout that was five times greater than that of the Standard markets. We modified the conversion rate of the final payout from $1.00 USD per 200 CASH to $1.00 USD per 40 CASH. All participants in these markets were inexperienced. The plots below represent average forecasts.



s

Figure 3.22: Four High-Stakes markets. The lines represent the dynamics of price levels for 30 periods. Each point represents the current period forecast, and subsequent lines represent average price forecasts for the 0-, 2-, 5-, and 10-period ahead periods.

Experienced Markets Plots

We ran five Experienced markets.

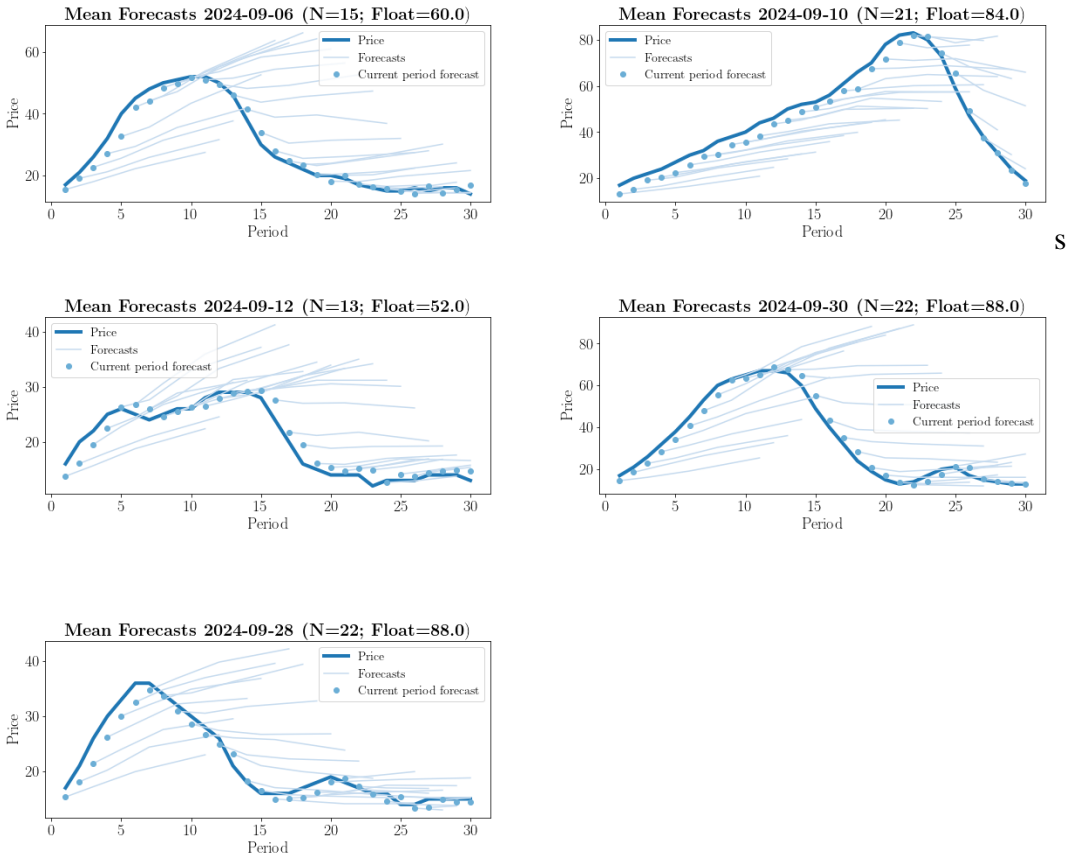


Figure 3.23: Five Experienced markets. The lines represent the dynamics of price levels for 30 periods. Each point represents the current period forecast, and subsequent lines represent average price forecasts for the 0-, 2-, 5-, and 10-period ahead periods.

3.16 Experienced Participants Selection Effects

To identify the impact of any selection effects, specifically, whether participants who returned for follow-up markets differed from those who did not, we compared the various survey responses and market performance measures of an Experienced participant's first market with those of participants who did not repeat the experience (see table (3.16)). We found that the only statistically significant differences between the two groups (as determined by a t-test) were the number of attempted and executed trades.

Tested Selection Effects:

- Age
- Similar

- Forecast Bonus
- Market Bonus
- Quiz Grade
- Attempted Trades
- Executed Trades
- General Risk
- Financial Risk

The survey measures were: Age; whether they had participated in a similar experiment before (coded as “Similar" 0-1); a self-reported preference for general and financial risk appetite (Exact measures can be found in section 3.32). Performance measures included market and forecast bonuses, their grade on the post-instruction quiz (section 3.31), the number of orders submitted, and the number of orders executed.

Table 3.16: Comparing means of measures of participants who repeated the experiment with the measures of their peers during their first market session.

	Only Once	Repeated	p-value
N	189.00	74.00	
Age	36.25	37.44	0.431
Similar	0.11	0.05	0.191
Forecast Bonus	0.12	0.15	0.531
Market Bonus	3.09	3.47	0.203 †
Quiz Grade	2.43	2.77	0.075
Attempted Trades	**39.03**	**32.93**	**0.023**
Executed Trades	**11.20**	**8.47**	**0.005**
General Risk	4.17	3.81	0.117
Finacial Risk	3.85	3.64	0.368

'Similar' is the percentage of participants who answered "yes" to the survey question: "Have you participated in similar trading experiments before?" Market and forecast bonuses are the bonuses earned during the experiment. Quiz Grade is the score out of 5 on the comprehension quiz. Attempted Trades are the number of orders submitted, while Executed Trades are the orders that were actually filled. General Risk is a Likert response to the question: "How would you rate your risk-taking behavior in general, on a scale from 1–7," where 1 is risk-averse, and 7 is risk-taking. Financial Risk is a Likert response to the question: "How would you rate your financial risk-taking behavior, on a scale from 1–7?" The dagger (†) on Market Bonus indicates that the p-value was computed by bootstrapping differences due to within-market correlations (see Figure 3.24).
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

On average, Experienced participants attempted 32.93 trades compared to 39.03 trades for the general participant population ($p = 0.023$). We also found that on average future repeat participants executed fewer trades (8.47) than participants that did not go on to repeat markets (11.20), ($p = 0.005$).

We also analyzed how repeat participant behavior changed over time. We find that compared to their first experiment, repeat participants seem to attempt and execute fewer trades, though these differences are not statistically significant. A full table detailing how these Experienced participants traded in each repeat period is available below in Table 3.17.

Table 3.17: Mean trades by Experienced Markets.

Session	N	Att Trades	Exe Trades	Market Bonus	Forecast Bonus
1	74	32.93	8.47	3.47	0.15
2	74	30.82	8.47	3.27	0.03
3	24	25.50	5.75	3.80	0.05
4	4	35.25	8.75	3.40	0.03

Note. Summary of mean trade activity and market performance based on experience level. A full breakdown is provided in Table 3.33 in Section 3.28.

The mean market bonus for one-time *prolific.com* participants (μ_o = \$3.09) was compared to the first market bonus earned by repeat participants (μ_r = \$3.47). The observed difference in means was $\mu_r - \mu_o$ = \$0.39. Because market bonuses are correlated within a market, gains for some traders necessarily imply losses for others; a simple t-test would not suffice. Instead, we bootstrapped the means by shuffling the market bonuses within each market and recalculating the group means ($\mu_{ri} - \mu_{oi}$ where i is the i^{th} iteration of the bootstrapping process). This process was repeated N=10,000 times to create a distribution of market bonus outcomes with mean μ_b = \$0.17 and standard deviation σ_b = \$0.26 (Figure 3.24). We then tested the null hypothesis ($\mathtt{H}_0 : \mu_r - \mu_o = \mu_b$) by counting the number of instances in the bootstrapping process that generated a difference greater than the observed difference of \$0.39 and dividing by N. This generated a p-value of 0.210, so we cannot reject $\mathtt{H}_0$. We also calculated a student's T distribution to compare the observed difference with the bootstrapped distribution with 9,999 degrees of freedom to get a p-value of 0.206.

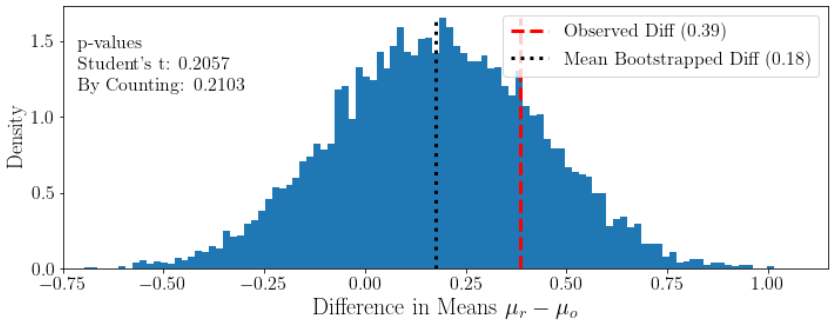


Figure 3.24: Market Bonus Comparison One-timers vs. Repeaters (Bootstrap N=10,000) The mean market bonus for one-time *prolific.com* participants (μ_o = \$3.09) was compared to the first market bonus earned by repeat participants (μ_r = \$3.47). Market bonuses were shuffled within each market. We then calculated the difference in the means of market bonuses, grouped by one-timers and repeaters. We repeated this 10,000 times to generate a distribution (μ = \$0.18; σ = 0.26; df = 9,999). The observed difference of \$0.39 is shown by a red dashed line. The p-value of this comparison was calculated by calculating $1 - \text{cdf}(\mu, \sigma, \text{df}) = 0.206$. We also computed the p-value by calculating the proportion of the bootstrapped observations that are greater than the observed difference (p=0.210).

3.17 Evaluating individual forecast consistency

While forecasting accuracy is a natural benchmark for assessing beliefs, accurate or biased forecasts alone do not determine market outcomes unless they are acted upon. A separate, conceptually distinct question is whether participants trade in a manner consistent with their stated expectations. Even if forecasts are biased or noisy, a trader who systematically conditions their trades on their beliefs may still behave coherently, whereas a trader who disregards their own forecasts exhibits a disconnect between beliefs and actions. Assessing this internal consistency, therefore, provides insight into whether forecasts meaningfully guide trading decisions and whether heterogeneity in performance reflects differences in how beliefs are operationalized, rather than differences in beliefs themselves.

In our experiment, participants provide forecasts prior to the revelation of actual prices. Specifically, at each period t, participants submit forecasts for the current period as well as for the next two, five, and ten periods. Let $f_{i,t}^{t+h}$ denote the forecast made by participant i at period t for the price in period $t+h$, where $h \in \{0, 2, 5, 10\}$. Thus, $f_{i,t}^{t}$ specifically refers to the forecast for the current period's price P_t made by participant i. For example, before the price for period 8 is revealed, each participant provides the forecasts $f_{i,8}^{8}$, $f_{i,8}^{10}$, $f_{i,8}^{13}$, and $f_{i,8}^{18}$. Throughout this paper, market identifiers are omitted from the forecast notation for simplicity.

To analyze the link between expectations and trading decisions, we regressed each participant's net trading on an indicator for whether their forecasted next-period price increase exceeded the 5% risk-free rate, estimating the regression separately for each individual.

If a participant forecasts that the price will rise by more than the risk-free rate of 5%, a consistent response (in the absence of constraints) would be to increase net buying. Therefore, our indicator regression provides a proxy for internal consistency: a significant, positive relationship would indicate that participants trade consistently with their own beliefs.

$$(\hat{\beta}_{0,i}, \hat{\beta}_{1,i}) = \arg\min_{\beta_0,\beta_1} \sum_t \left(\mathrm{Y}_{i,t} - \beta_0 - \beta_1 \mathbf{1}\left\{ \frac{f_{i,t}^t - P_{t-1}}{P_{t-1}} + \frac{0.7}{f_{i,t}^t} > 0.05 \right\} \right)^2 . \quad (3.15)$$

The estimated parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ are stored for each participant.

To ensure accurate estimation, the following preprocessing steps were performed.

1. Buy orders were assigned positive values ($+q$), while sell orders were assigned negative values ($-q$).

2. Net trading orders were computed as the sum of buy and sell orders for each participant in each period t:

$$Y_{i,t} = \sum_j q_{i,t}, \quad (3.16)$$

   where $q_{i,t}$ represents an order's quantity in period t.

Figure 3.25 shows the distribution of participant-level coefficients. A positive coefficient indicates increased net buying when a participant forecasts returns above the risk-free rate. More than half of our participants exhibit coefficients that are positive and we see a mean coefficient of 0.89, indicating that on average, participants trade 0.89 more units when they expect $\geq 5\%$ returns compared to when they do not.

Taken together, these results indicate that, for most participants, the link between stated beliefs and trading purchase behavior is strong, consistent with the findings of Grosshans and Zeisberger (2025). As illustrated in Figures 3.25 many individuals

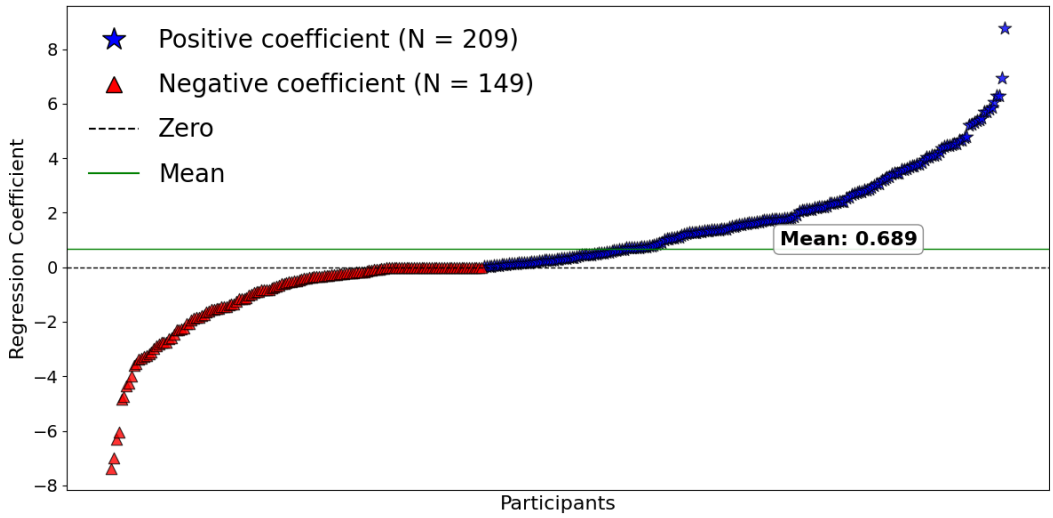


Figure 3.25: Plot of individual regression coefficients for net orders on an indicator for forecasting returns above the 5% risk-free rate. Each point is a participant-level coefficient from a regression of net orders on an indicator equal to one when the participant's forecasted return exceeds the 5% risk-free rate. Stars represent participants with positive coefficients, while triangles represent participants with negative coefficients. The coefficient measures how much net buying increases (or decreases) in high–expected-return states relative to low- or sub-risk-free-return states. More positive coefficients indicate a higher degree of internal consistency between beliefs and trading behavior.

exhibit strong alignment between their forecasts and their trading decisions. This pattern suggests that deviations from rational expectations does not arise from a failure to act on those beliefs systematically.

3.18 Model Features

Full Explanatory Variables Details

We suppress the market notation s for clarity; every variable below is calculated in a given market s. All explanatory variables are constructed using information available at or before the end of period $t-1$, unless otherwise noted.

Below are the specific explanatory variables used:

1. Previous Trading Period Returns

$$R_{t-1}$$

2. Averaged Risk-Aversion

$$\rho_t$$

3. Backward-Looking Forecasts

$$\bar{f}^{t}_{t-h}$$

where $\bar{f}^{t}_{t-h}$ denotes the cross-sectional mean forecast of price in period t, made h periods earlier.

4. Previous Trading Period's Volume

$$V_{t-1}$$

5. Remaining Periods

$$30 - t$$

6. OrderBook Pressure $\Omega_{s,t}$

$$\Omega_t$$

where Ω_t is given by:

$$\Omega_t = \frac{\sum_{p_b < P_{t-1}} (v_b p_a) - \sum_{p_a > P_{t-1}} (v_a p_b)}{TotalOrderVolume_{t-1}}$$

Here, the the best ask price (p_a) is paired with the total outstanding bid volume (v_b), and the best bid price (p_b) is paired with the total outstanding ask volume (v_a).

7. Average Asset Allocation

Υ_t = Average asset allocation (Cash Value/Stock Value) across participants in period t

8. Average Electrodermal Activity (EDA)

$$\frac{\sum_{i=1}^{N_t} \text{EDA}_{i,t}}{N_t}$$

where:

- N_t is the number of participants in period t,

EDA data captures both contemporaneous and lagged physiological responses.

9. Buying Ratio

$$\frac{\text{Total Buy Orders}_{t-1} - \text{Total Sell Orders}_{t-1}}{\text{Total Buy Orders}_{t-1} + \text{Total Sell Orders}_{t-1} + \epsilon}$$

Where:

- Total Buy Volume represents the sum of all purchase orders,
- Total Sell Volume represents the sum of all sell orders,
- n is the total number of transactions,
- $\epsilon = 10^{-5}$ is a small constant added to prevent division by zero,
- The ratio ranges from -1 to 1, with positive values indicating net buying pressure and negative values indicating net selling pressure.

Feature Selection The following transformations of the above variables are used as inputs to the predictive models (Machine Learning and Lasso):

- **Average DOSE risk parameters.**

$$\bar{\rho}_t = \frac{\sum_{i=1}^{N_t} \rho_{i,t}}{N_t}$$

$$\bar{\mu}_t = \frac{\sum_{i=1}^{N_t} \mu_{i,t}}{N_t}$$

  Where N_t is the number of participants in period t.

- **Lagged trading volume.** Total trading volume from the previous period, V_{t-1}.

- **Aggregate price forecasts.** Cross-sectional forecasts averaged across participants for period t, made in periods before period t at different horizons:

$$f_t^t, \quad f_{t-2}^t, \quad f_{t-5}^t, \quad f_{t-10}^t.$$

- **Physiological arousal (EDA).** Summary statistics of tonic electrodermal activity, including the mean, standard deviation, and skewness of EDA_t^{ton}, aggregated across participants within each period.

3.19 Pre-registered Analysis

Simple OLS

Table 3.18: Predicting Next-Period Log Returns R_t with Backward-Looking Forecasts (returns ×100)

	(1) Backward-Looking f_{t-h}^{t}	(2) Forward-Looking f_{t}^{t+h}
Intercept	-2.392	-0.933
	(2.979)	(2.540)
R_{t-1}	0.839***	0.822***
	(0.039)	(0.038)
ρ	0.020	-2.836**
	(1.504)	(1.117)
$30 - t$	0.015	0.105
	(0.084)	(0.090)
Ω_t	0.015	-0.021
	(0.029)	(0.029)
Υ_t	0.008	0.002
	(0.008)	(0.007)
V_{t-1}	-0.036	-0.055
	(0.062)	(0.053)
EDA_t^{ton}	2.786	-0.152
	(2.004)	(1.901)
f_t^t	0.322	0.282
	(0.244)	(0.171)
f_{t-2}^{t}	-0.541**	
	(0.204)	
f_{t-5}^{t}	0.202	
	(0.153)	
f_{t-10}^{t}	-0.012	
	(0.141)	
f_t^{t+2}		-0.215
		(0.199)
f_t^{t+5}		0.162
		(0.229)
f_t^{t+10}		-0.164
		(0.146)
N	355	336
R^2	0.697	0.689
Adjusted R^2	0.687	0.679

Notes: Dependent variable is future log return R_t, scaled by 100. Standard errors in parentheses are heteroskedasticity- and cluster-robust, clustered by market. The pre-registered specification used backward-looking forecasts; both specifications are shown. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 3.19: Predicting Next-Period Log Returns R_t with High-Stakes Sessions (returns ×100)

	Dependent variable: future_return × 100	
	(1) Backward-Looking f_{t-h}^{t}	(2) Forward-Looking f_{t}^{t+h}
Intercept	5.938	9.441
	(3.175)	(8.590)
R_{t-1}	0.488*	0.572**
	(0.169)	(0.170)
ρ	-1.502	-0.617
	(1.494)	(2.875)
$30 - t$	-0.070	-0.212
	(0.101)	(0.242)
Ω_t	-0.052	0.079*
	(0.055)	(0.025)
Υ_t	-0.013	-0.022
	(0.024)	(0.043)
V_{t-1}	-0.220	-0.454
	(0.226)	(0.224)
f_t^t	-1.457**	0.745
	(0.326)	(0.598)
f_{t-2}^{t}	0.894*	
	(0.289)	
f_{t-5}^{t}	0.494	
	(0.220)	
f_{t-10}^{t}	0.021	
	(0.188)	
f_t^{t+2}		-1.043
		(0.843)
f_t^{t+5}		-0.357
		(0.659)
f_t^{t+10}		0.503
		(0.269)
Observations	75	71
R^2	0.492	0.479
Adjusted R^2	0.413	0.392

Notes: Dependent variable is future log return R_t, scaled by 100. Standard errors in parentheses are heteroskedasticity- and cluster-robust, clustered by market. The pre-registered specification used backward-looking forecasts; both specifications are shown. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 3.20: Predicting Next-Period Log Returns R_t in Experienced Sessions (returns $\times 100$)

	Dependent variable: future return × 100	
	(1) Backward-Looking f_{t-h}^{t}	(2) Forward-Looking f_{t}^{t+h}
Intercept	-6.036	-15.059
	(11.702)	(7.711)
R_{t-1}	0.716***	0.722***
	(0.076)	(0.060)
ρ	10.908**	13.766
	(3.535)	(6.670)
$30 - t$	-0.066	0.129
	(0.465)	(0.377)
Ω_t	0.008	-0.043
	(0.061)	(0.061)
Υ_t	-0.032	-0.003
	(0.034)	(0.030)
V_{t-1}	0.210	0.057
	(0.351)	(0.371)
f_t^t	1.262	0.367
	(0.967)	(1.154)
f_{t-2}^{t}	-1.873	
	(1.089)	
f_{t-5}^{t}	0.549***	
	(0.092)	
f_{t-10}^{t}	0.070	
	(0.126)	
f_t^{t+2}		0.119
		(1.440)
f_t^{t+5}		-0.501
		(0.784)
f_t^{t+10}		0.151
		(0.462)
Observations	93	88
R^2	0.695	0.688
Adjusted R^2	0.658	0.647

Notes: Dependent variable is future log return R_t, scaled by 100. Standard errors in parentheses are heteroskedasticity- and cluster-robust, clustered by market. The pre-registered specification used backward-looking forecasts; both specifications are shown. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

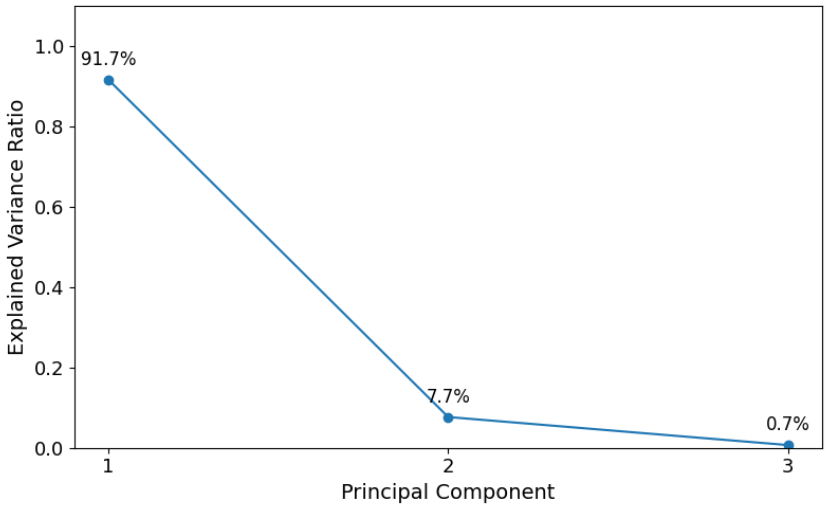


Figure 3.26: Scree plot showing the explained variation of each principle component of the three forecast measures used the main regression (Table 3.2). Forecast variables are f_t^t, f_{t-2}^t, f_{t-5}^t. Exact explained percentage of each component can be found beside the data point.

3.20 Standard Session Regressions

We performed a couple of robustness checks of the pre-registered model. First, we looked at the difference between backward and forward looking forecasts with no major changes in results (shown in the Pre-registration analysis section in Table 3.18). [32] We report both backward- and forward-looking forecast specifications to distinguish between the predictive content of beliefs formed prior to trading and beliefs stated contemporaneously about future prices. Next, We suspected multicollinearity among the forecast variables and Figure 3.27 shows a high degree of correlation among the them (f_t^t, f_{t-2}^t, f_{t-5}^t). Furthermore, a PCA analysis of those variables shows that the first principle component explains over 91% of the variation. We reran the main regression substituting the forecast variables with their first principal component (Table 3.2). The results are not much different from the main result with no major changes in significance.

3.21 Mathematical Explanation of the ANN Model

Model Architecture

The ANN model is a feed-forward neural network composed of:

- **Input layer**: Accepts input features $\mathbf{x} \in \mathbb{R}^n$, where n is the number of features.

[32] Regression notation and additional details can be found in section 3.22 in the Appendix.

Table 3.21: Predicting Next-Period Log Returns R_t with Backward- vs Forward-Looking Forecasts (returns ×100)

	(1)	(2)	(3)	(4)	(5)
Intercept	-0.621***	-5.032*	-4.240**	-19.007***	-4.340**
	(0.115)	(2.660)	(1.719)	(4.152)	(1.954)
R_{t-1}	0.827***		0.698***		0.698***
	(0.034)		(0.043)		(0.043)
$EDA_t^{ton}\text{MA}_5$		9.783*		-0.620	0.208
		(5.420)		(4.366)	(1.557)
ρ			0.654	-1.078	0.671
			(0.879)	(1.499)	(0.906)
V_{t-1}			-0.115*	-0.137	-0.114*
			(0.061)	(0.082)	(0.063)
$30 - t$			0.199**	0.916***	0.198**
			(0.090)	(0.186)	(0.091)
Ω_t			0.034	0.181**	0.034
			(0.029)	(0.064)	(0.029)
Υ_t			0.002	0.017	0.002
			(0.006)	(0.011)	(0.006)
Forecast PCA 1			-0.899*	-1.982*	-0.899*
			(0.494)	(0.999)	(0.494)
Observations	456	456	456	456	456
R^2	0.668	0.012	0.699	0.470	0.699
Adjusted R^2	0.668	0.010	0.694	0.461	0.693

Notes: Dependent variable: future log return R_t (scaled by 100). Standard errors (in parentheses) are heteroskedasticity- and cluster-robust (clustered by market). This model shows first principal component of the three forecast measure as a check for multicollinearity. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

- **Hidden layers**:
    1. **First hidden layer**: 64 neurons with ReLU activation.
    2. **Dropout layer**: Dropout rate of 0.1.
    3. **Second hidden layer**: 32 neurons with ReLU activation.
    4. **Dropout layer**: Dropout rate of 0.1.
- **Output layer**: Single neuron for regression output.

Dropout is used to help protect against overfitting; by randomly setting a fraction of

Table 3.22: Lagged Returns (x100) and Forecast Measures in Standard Sessions

Variable	(1)	(2)	(3)	(4)	(5)
$\bar{f}_t^t$	0.080***				-0.760*
	(0.030)				(0.440)
$\bar{f}_{t-2}^t$		0.11***			0.260
		(0.030)			(0.53)
$\bar{f}_{t-5}^t$			0.11***		0.490***
			(0.030)		(0.140)
$\bar{f}_{t-10}^t$				0.110***	0.020
				(0.030)	(0.030)
Constant	-1.320	-1.450	-0.150	2.020	3.270***
	(1.250)	(1.420)	(1.360)	(1.250)	(0.940)
Observations	532	494	437	341	341
R^2	0.020	0.041	0.064	0.083	0.318

Notes: The dependent variable is lagged log return. Sample restricted to Standard sessions. Standard errors clustered by market are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

the input units to zero during training, the network develops redundant representations, which typically improve generalization.

Mathematical Formulation

Let $\mathbf{x} \in \mathbb{R}^n$ be the input vector. The model computes the output $\hat{y}$ through the following transformations:

1. **Hidden layer architecture**:

$$\mathbf{z}^{(1)} = \text{ReLU}(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}), \tag{3.17}$$

where $\mathbf{W}^{(1)} \in \mathbb{R}^{64\times n}$ is the weight matrix, $\mathbf{b}^{(1)} \in \mathbb{R}^{64}$ is the bias vector, and ReLU is the Rectified Linear Unit activation function defined as:

$$\text{ReLU}(x) = \max(0, x).$$

Loss Function

The model is trained to minimize the Mean Squared Error (MSE) loss function:

$$\mathcal{L}(\theta) = \frac{1}{m}\sum_{i=1}^{m}(y^{(s,t)} - \hat{y}^{(s,t)})^2,$$

where:

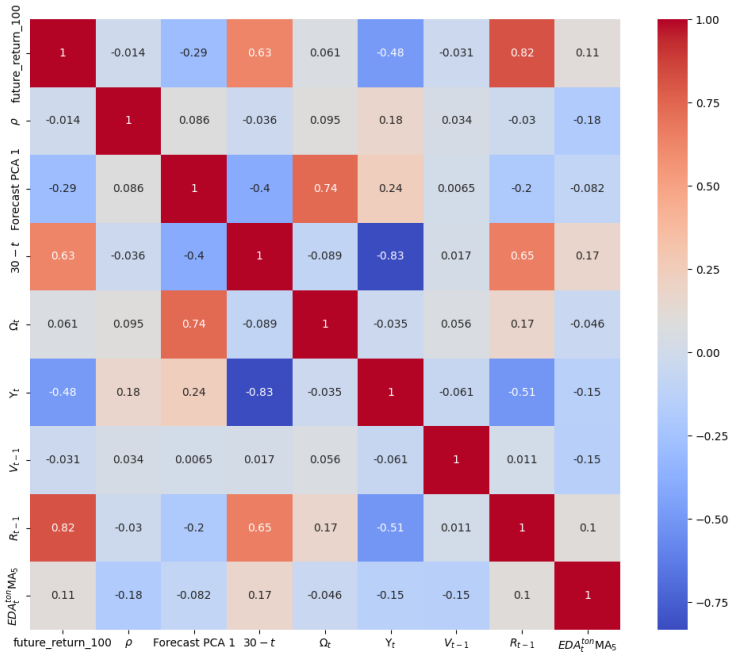


Figure 3.27: Correlational matrix of regressor in future returns predictions OLS (Table 3.2). The image shows a high level of correlation among the three forecast measures; indicating the possibility of multicollinearity. The first principle component of the three forecasts are included to avoid this in the regressions provided in the main text.

- m is the number of market period return observations.
- $y^{(s,t)}$ is the true returns for the market, period s, t.
- $\hat{y}^{(s,t)}$ is the predicted value for the market, period s, t.
- θ represents all the weights and biases in the network.

Optimization Algorithm

The model uses the Adam optimizer for training, which is an adaptive learning rate optimization algorithm that combines ideas from RMSProp and Momentum

methods. The update rules for the weights are:

$$\begin{aligned}
\mathbf{m}_t &= \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\nabla_\theta L(\theta), \\
\mathbf{v}_t &= \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)(\nabla_\theta L(\theta))^2, \\
\hat{\mathbf{m}}_t &= \frac{\mathbf{m}_t}{1-\beta_1^t}, \\
\hat{\mathbf{v}}_t &= \frac{\mathbf{v}_t}{1-\beta_2^t}, \\
\theta &= \theta - \alpha \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \epsilon}.
\end{aligned}$$

- $\mathbf{m}_t$ and $\mathbf{v}_t$ are the first and second moment estimates.
- β_1 and β_2 are exponential decay rates for the moment estimates.
- α is the learning rate (set to 0.001).
- ϵ is a small constant to prevent division by zero.

3.22 Mathematical Explanation of the SVR Model

Support Vector Regression(SVR)

Support Vector Machine Regressions (SVRs) estimates a function $f(\mathbf{x})$ that approximates the data within a tolerance ϵ, treating deviations smaller than ϵ as negligible. This allows the model to capture the underlying trend while ignoring minor fluctuations. Simultaneously, SVR enforces flatness of the function by minimizing the norm of the weight vector, thereby controlling model complexity. The trade-off between approximation accuracy and simplicity is governed by the regularization parameter C, while kernels (e.g., linear, RBF) enable the method to handle both linear and nonlinear relationships.

Below are the details of our SVR setup and corresponding hyperparameters.

- **Hyperparameters:**
  - C: Regularization parameter controlling the trade-off between model complexity and the amount up to which deviations larger than ϵ are tolerated.
  - `kernel`: The kernel function used to transform data into higher-dimensional space (e.g., `linear` or `rbf`).
  - ϵ: The width of the ϵ-insensitive tube within which no penalty is associated in the training loss.

Feature Pre-Processing

1. **Imputation**: Missing feature values are imputed using the median of each feature.

2. **Scaling**: The features are standardized using Z-score normalization:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}, \tag{3.18}$$

where μ and σ are the mean and standard deviation of the feature in the training set.

Mathematical Formulation

Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ be our training data, where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$. The goal of ϵ-SVR is to learn a function $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$ (in the linear case) or

$$f(\mathbf{x}) = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b \quad \text{(in the kernelized case)},$$

where $K(\cdot, \cdot)$ is the kernel function.

Primal Problem. SVR solves the following optimization problem:

$$\min_{\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} \left(\xi_i + \xi_i^*\right)$$

participant to

$$\begin{aligned} y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - b &\leq \epsilon + \xi_i, \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i &\leq \epsilon + \xi_i^*, \\ \xi_i,\ \xi_i^* &\geq 0, \quad i = 1, \ldots, m, \end{aligned}$$

where ξ_i and ξ_i^* are positive slack variables penalizing predictions outside the ϵ-insensitive tube.

Dual Problem. By introducing Lagrange multipliers α_i and α_i^*, one obtains the dual form, and the resulting decision function can be written as:

$$f(\mathbf{x}) = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b.$$

Only a subset of the training points (the *support vectors*) will have non-zero $(\alpha_i - \alpha_i^*)$.

Loss Function

In ϵ-SVR, the loss function is known as the ϵ-insensitive loss. It penalizes deviations only when the prediction is outside the ϵ-insensitive region:

$$L_\epsilon(y, \hat{y}) = \begin{cases} 0, & \text{if } |y - \hat{y}| \leq \epsilon, \\ |\, y - \hat{y}\, | - \epsilon, & \text{otherwise.} \end{cases}$$

However, our models, when performing parameter selection (e.g., via cross-validation), often use a standard regression metric such as the Mean Squared Error (MSE) or Mean Absolute Error (MAE) to compare models. In our case, we use MSE:

$$\mathcal{L} = \frac{1}{m} \sum_{i=1}^{m} \big(y_i - f(\mathbf{x}_i)\big)^2.$$

Optimization Algorithm

The SVR model is typically trained via an iterative method that solves the underlying convex Quadratic Programming problem. In scikit-learn, this is handled internally by the `SVR` class (e.g., using `libsvm` or `liblinear` backends). We use `GridSearchCV` with 5-fold cross-validation to select the optimal hyperparameters. Our chosen grid for C and `kernel` is given by:

Parameter	**Values**
C	$\{0.1, 1, 10\}$
`kernel`	{`linear`, `rbf`}

We train the model to minimize `neg_mean_squared_error` (i.e., maximize − MSE), and then select the parameter combination yielding the best cross-validation performance.

Final SVR Model. After the grid search completes, we use the best combination of $(C, \texttt{kernel})$ to re-train the model on the full training set. The final prediction function is:

$$\hat{y} = f(\mathbf{x}) = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*)\, K(\mathbf{x}_i, \mathbf{x}) + b,$$

Where the learned weights and support vectors reflect the best-fit regression surface.

Table 3.23: Support Vector Regression (SVR) w/o EDA Feature Importance (Linear Kernel)

Feature	Normalized Importance
f_t^t	0.422
f_{t-2}^t	0.353
$30 - t$	0.137
f_{t-5}^t	0.037
f_{t-10}^t	0.021
$DOSE_\mu_t$	0.021
$DOSE_\rho$	0.008
V_{t-1}	0.000

Importance values are normalized absolute coefficients from the linear SVR model (sum to 1). Higher values indicate a stronger influence on predicted returns.

Random Split SVR

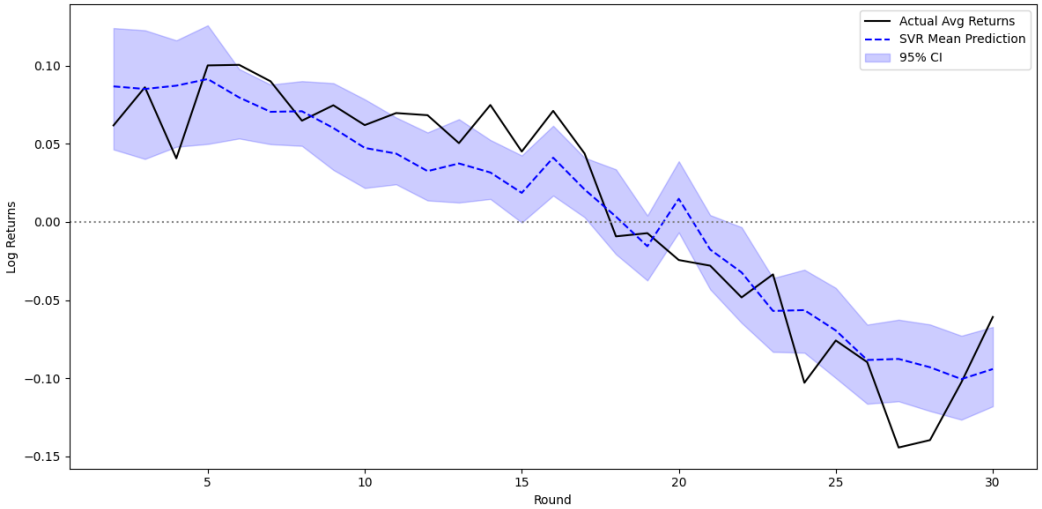


Figure 3.28: **SVR with EDA^{ton} Model Predictions vs Actual Returns.** The SVR mean prediction curve represents the model's out-of-sample forecasts averaged across all (four; [market: 1, 2, 6, 12]) held-out test markets for each trading period. Each SVR was trained on the remaining markets (15), cross-validated on training folds for tuning, and then evaluated once on the four held-out markets. Confidence intervals were generated using a nonparametric bootstrap applied to the training data: for each of 1000 iterations, the training data were resampled with replacement, a new SVR was fit to the bootstrapped sample, and its performance was evaluated on the fixed held-out test markets.

Table 3.24: Performance Metrics for Selected Models (with and without EDA^{ton})

		MAE	95% CI (MAE)		R^2	95% CI (R^2)	
Model	**Split**		**Low**	**High**		**Low**	**High**
SVR (w/ EDA^{ton})	Train	0.042	0.038	0.051	0.631	0.477	0.660
	Test	0.043	0.039	0.051	0.645	0.525	0.693
SVR (w/o EDA^{ton})	Train	0.041	0.029	0.048	0.628	0.483	0.776
	Test	0.041	0.040	0.050	0.671	0.548	0.695

Notes: Performance metrics are computed using a random train/test split (held out markets: [1,2,6,12]), as is standard in machine learning, and therefore differ from the held-out market validation reported in Figure 3.5. For CI related graph for these held-out markets see Figure 3.28.

Feature	**Normalized Importance**
f_t^t	0.372
f_{t-2}^t	0.292
$30 - t$	0.130
f_{t-5}^t	0.059
standard deviation EDA^{ton}	0.054
skewness EDA_t^{ton}	0.029
$DOSE_\mu_t$	0.026
$DOSE_\rho$	0.017
f_{t-10}^t	0.011
EDA_t^{ton}	0.008
V_{t-1}	0.002

Table 3.25: Importance values are normalized absolute coefficients from the linear SVR model (sum to 1). Higher values indicate a stronger influence on predicted returns. Using the same random split as above held out markets: [1,2,6,12].

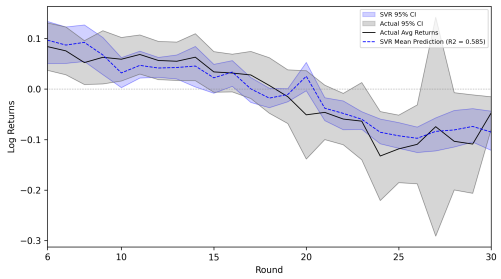
Our analysis reveals that traditional measures exhibit moderate predictive power for asset returns (MAE mean 0.041, R^2 mean = 0.686 for the non-EDA model)[33]. However, the inclusion of (EDA^{ton}) data does not improve model performance.

Our analysis reveals that traditional measures exhibit moderate predictive power for asset returns (MAE mean 0.041, R^2 mean = 0.686 for the non-EDA model). However, the inclusion of (EDA^{ton}) data does not improve model performance. Details of feature weights can be found in Table 3.4 and section 3.25 in the Appendix.

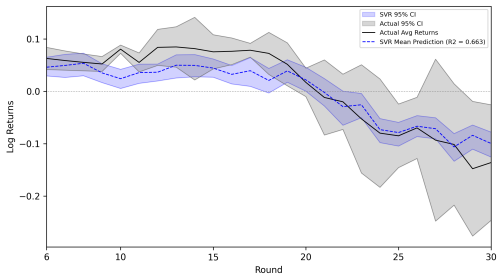
[33]To assess robustness to market selection, we repeated the w/o EDA^{ton} evaluation across 100 different random holdouts of 4 markets. Across these splits, test performance for the SVR model averaged MAE = 0.042 (SD = 0.005) and R^2 = 0.590 (SD = 0.117). The variability in R^2 across splits reflects heterogeneity in predictability across market environments rather than estimator instability; performance remained positive and economically meaningful across the vast majority of holdouts.

Notation is also explained further in Section 3.18.

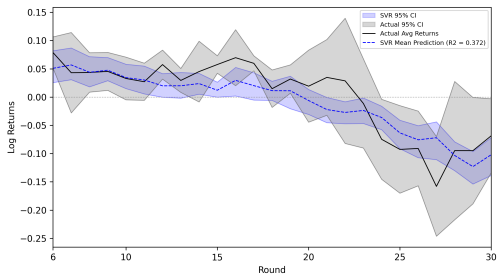
Regressions with Real value Confidence Intervals



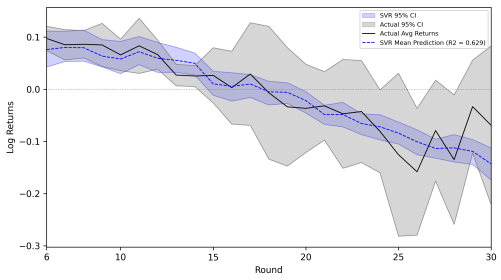
(a) Holdout: Markets 1–4



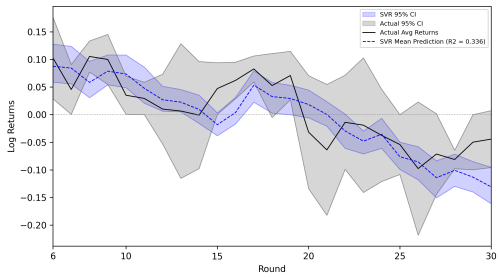
(b) Holdout: Markets 5–8



(c) Holdout: Markets 9–12



(d) Holdout: Markets 13–16



(e) Holdout: Markets 17–19

Figure 3.29: Mean out-of-sample SVR return predictions. Shaded purple regions show 95% confidence intervals from 1,000 bootstrap resamples of the training data, evaluated on a fixed held-out test set. Shaded gray regions show 95% confidence intervals from the standard deviation of the held-out test data. The majority of realized returns lie within the bootstrap confidence bands across all held-out market groups, suggesting the model captures common market dynamics and provides well-calibrated out-of-sample uncertainty.

Regression Details

Below is the outline of our regression model, which is used to predict returns and examine the relationship between our biometrics and asset prices.

Let $P_{s,t}$ be the price of the risky asset in market s, period t, and let:

$$R_{s,t} = \log\left(\frac{P_{s,t}}{P_{s,t-1}}\right)$$

be the log price ratio between periods t and $t-1$ in market s. Then, return dynamics can be captured by the following equation:

$$R_{s,t} = \alpha_s + \sum_{v=1}^{K} \beta_v X_{v,s,t} + \epsilon_{s,t} \tag{3.19}$$

Where:

- α_s is the market-specific intercept for market s,
- β_v represents the coefficient for the v-th variable in the vector of variables X_v,
- K is the number of variables,
- $X_{v,s,t}$ represents the v-th explanatory variable in market s, period t,
- $\epsilon_{s,t} \sim N(0, \sigma^2)$ is the error term, with $E(R_{s,t} \mid X_v)$ representing the conditional expectation.

LASSO regression

In empirical settings such as ours—where predictors can be highly correlated, and the number of potentially predictive variables is large—traditional Ordinary Least Squares (OLS) regression can yield highly variable estimates. This instability arises from multicollinearity, which makes it difficult to distinguish the contributions of correlated predictors. Moreover, including too many weakly informative variables can reduce out-of-sample predictive performance and lead to overfitting.

To address these issues, we estimated a Least Absolute Shrinkage and Selection Operator (LASSO) model using the same pre-registered set of explanatory variables as in our OLS specification. LASSO adds an ℓ_1-penalty to the regression loss function, shrinking small or redundant coefficients toward zero and effectively performing variable selection:

$$\hat{\boldsymbol{\beta}}^{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{3.20}$$

The regularization parameter λ was selected via a grid-search minimizing the Bayesian Information Criterion (BIC) over $\lambda \in [10^{-3}, 1]$. The optimal value was $\lambda = 0.091$. At this level of penalization, the LASSO retained only five non-zero predictors: lagged return, periods remaining, near-term forecast (f_t^t), 10-period forecast (f_{t-10}^t), and lagged volume. These variables were subsequently re-estimated using OLS with cluster-robust standard errors to obtain unbiased coefficient estimates and valid inference.

Table 3.26: Predicting next-period returns (R_t) using LASSO.

	Coefficient	Std. Error	t	p
Intercept	-0.755	1.209	-0.624	0.540
f_t^t	-0.068*	0.034	-2.014	0.059
f_{t-10}^t	0.046***	0.014	3.249	0.004
$30 - t$	0.141***	0.047	3.013	0.007
V_{t-1}	-0.088	0.051	-1.718	0.103
R_{t-1}	0.689***	0.036	19.017	< 0.001
N		532		
R^2		0.700		
Adjusted R^2		0.697		

Notes: Coefficients estimated by post-LASSO OLS with cluster-robust standard errors (clustered by market). Dependent variable: next-period log return R_t (multiplied by 100). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 3.27: Chow Test for Differences in Quadratic EDA Profiles by Earnings Tercile

Test	F**-Statistic**	p**-Value**
Classical Chow test	16.917	1.27×10^{-8}
Robust Wald interaction test	28.569	1.06×10^{-12}

Notes: Tests compare the High and Low earning-tercile quadratic EDA profiles estimated from $\overline{EDA}_\tau^{ton} = \alpha + \beta\tau + \gamma\tau^2 + \varepsilon$, where τ denotes rounds relative to the market peak. The test is implemented by estimating a pooled interaction model with Low-tercile indicators interacted with τ and τ^2. The null hypothesis is equality of the intercept, linear, and quadratic terms across High and Low terciles. The robust Wald test uses heteroskedasticity-robust standard errors. Regressions are estimated on tercile-by-τ binned means.

The resulting post-LASSO model achieves a similar R^2 as the standard OLS regressions, an $R^2 = 0.700$, indicating that the retained predictors explain 70% of the variation in next-period returns. Lagged returns are the dominant driver, but backward-looking forecasts and market-level variables (periods remaining, lagged

volume) also contribute significantly to short-horizon return dynamics. LASSO helps control the multicollinearity issue present in our standard OLS measures.

Other Machine Learning Models

We also conducted analysis with various other machine learning techniques, outlined below.

Table 3.28: Final model performance metrics. Other Machine Learning Models

Model	MSE	MAE	R^2	Max Error
XGBoost	0.005	0.053	0.448	0.291
SVR	0.003	0.043	0.645	0.211
RandomForest	0.005	0.054	0.493	0.237
HistGradientBoosting	0.004	0.047	0.562	0.259
ExtraTrees	0.004	0.053	0.512	0.222
ANN	0.004	0.047	0.573	0.245
MLP	0.005	0.058	0.395	0.224
LSTM	0.004	0.044	0.618	0.234

Mean squared error (MSE) and mean absolute error (MAE) summarize average prediction error, while R^2 measures the proportion of variance explained relative to a baseline model. Max Error reports the worst–case absolute residual. We report multiple metrics because different ML conventions emphasize different criteria.

3.23 DOSE Preference Measures

We use preference measures generated by DOSE, or Dynamically Optimized Sequential Experimentation (Chapman et al., 2024). DOSE is an adaptive Bayesian procedure for estimating individual-level economic preference parameters. The procedure begins with a model of preferences and a prior distribution over the parameters of that model. After each participant response, the prior is updated, and the posterior distribution is used to select the next question that is expected to be most informative.

We chose to calculate the DOSE parameters using the CRRA utility function:

$$u(x, \rho_i) = \frac{x^{1-\rho_i}}{1-\rho_i}$$

where risk aversion increases with ρ_i, and $x \in \mathbb{R}^+$ is the monetary payout presented in the DOSE task.

DOSE also explicitly models probabilistic choice. For any choice between options o_1 and o_2, with $V(o_1) > V(o_2)$, choice probabilities are given by the logit function:

$$\Pr(o_1) = \frac{1}{1 + \exp\left[-\mu_i \left(V(o_1) - V(o_2)\right)\right]},$$

where $\mu_i \in \mathbb{R}_+$ is a choice consistency parameter. Higher values of μ_i imply more consistent choices.

Question selection is based on expected Kullback–Leibler divergence. Consider a finite set of possible parameter vectors θ_k, for $k = 1, \ldots, K$, where $\theta_k = (\rho_k, \lambda_k, \delta_k, \mu_k)$, and let p_k be the posterior probability assigned to parameter vector θ_k. For a possible question Q_j, the expected Kullback–Leibler divergence is:

$$KL(Q_j) = \sum_{k=1}^{K} \sum_{a \in A} p_k l_k(a; Q_j) \log \left(\frac{l_k(a; Q_j)}{\sum_{m=1}^{K} p_m l_m(a; Q_j)} \right),$$

where $a \in A$ indexes the possible answers to the question and $l_k(a; Q_j)$ is the likelihood of answer a given θ_k. DOSE selects the question that maximizes $KL(Q_j)$, observes the participant's answer, updates the posterior over parameter vectors, and repeats this process for the next question.

The DOSE measures used in our analyses are posterior estimates of latent individual-level preference parameters rather than raw choice scores. The adaptive structure allows the elicitation procedure to tailor later questions to earlier answers, improving the precision of estimated risk, and loss parameters. Figure 3.30 represents the progress of one such DOSE task. We presented 33 tasks to participants one for each of the 3 practice rounds and 30 main market rounds. Tasked are distinguished from each other by varying the payouts presented to the participant, which in turn varied the percentages presented on screen in the pie charts.

A visual example of how the model generates ρ is shown in Figure 3.30.

3.24 EDA^{ton} activity by Earnings Tercile

For much of the individual-level analysis, we divided participants by earning terciles. This tercile split is done at the market level, so there is some overlap in earnings. That is, a medium-level earner in one market might actually earn more than a high earner in another. Table 3.29 shows the earnings ranges by tercile.

We also investigated whether there were differences in EDA^{ton} activity among high earners compared to other earners. To do so, we sorted participants into terciles

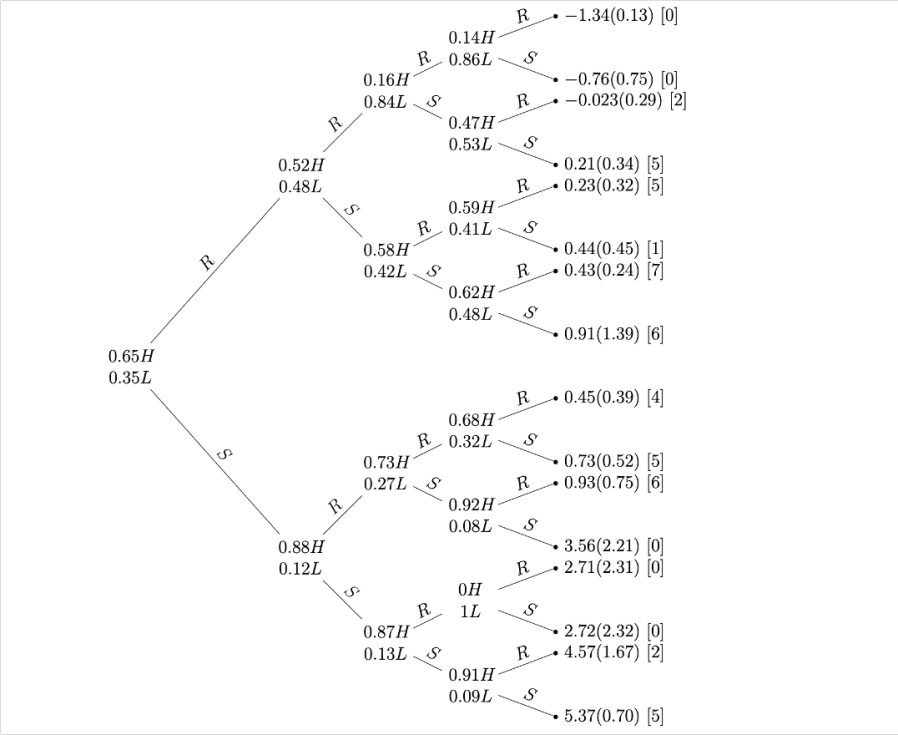


Figure 3.30: Example DOSE decision tree for generating estimates of ρ.

based on their individual market bonuses in each market session. For clarity, the main text only shows the top and bottom terciles. Here is the same plot showing all three terciles (figure 3.31)

Table 3.29: Ranges of market earnings by tercile.

Earnings Tercile	Min	Max
High	3.48	15.43
Medium	1.38	3.80
Low	0.00	3.21

Notes: The terciles are determined at the market level, which explains any inter-tercile overlaps.

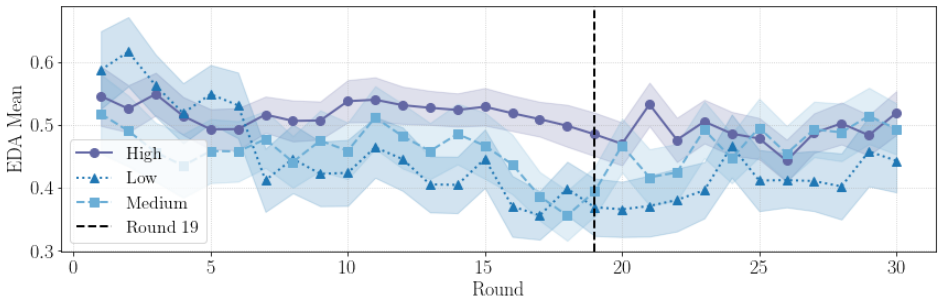


Figure 3.31: Average EDA^{ton} over 30 periods for participants grouped by market-earnings tercile (low, medium, high). 90% confidence intervals shown.

We also performed the same analysis with a median split (Figure 3.32). The Levene test for variance by median is similarly significant ($\sigma_{hi} = 0.025, \sigma_{lo} = 0.049, p = 0.009$).

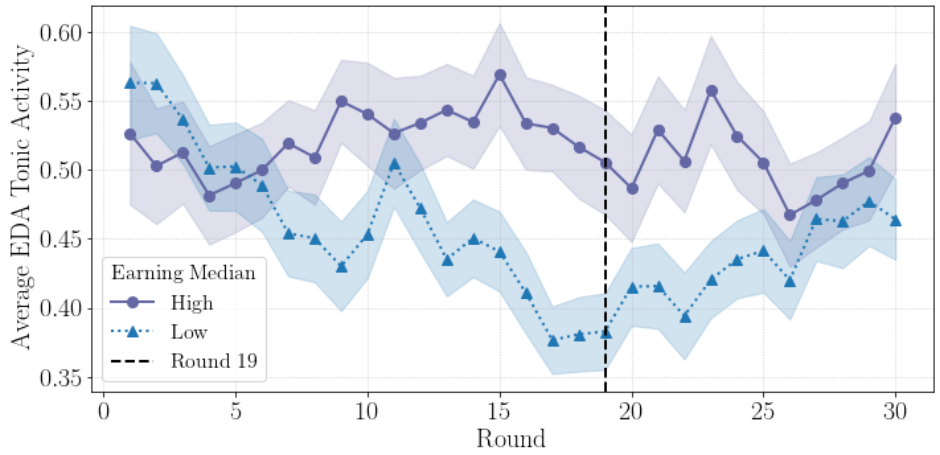


Figure 3.32: Average EDA^{ton} over 30 periods for participants grouped by market-earnings median (low, high). 95% confidence intervals shown.

Forward Returns Prediction from Smoothed Tonic EDA

While direct averages of EDA_t^{ton} did not predict subsequent returns R_t, we did an exploratory analysis of whether future returns might be predicted by smoothed EDA^{ton}. We find that when smoothed, there is a significant relationship between EDA^{ton} and forward returns.

The justification for this analysis is that emotion researchers distinguish between transitory emotions (often associated with events, and which are short-lived– minutes or hours) and moods, which are longer-lived (days or weeks) (Russell, 2003; Beedie, Terry, and Lane, 2005).

Neuroscientists distinguish between fast 'phasic' responses and slower time-averaged 'tonic' responses (Dawson, Schell, and Filion, 2007; Boucsein, 2012). Maybe our EDA^{ton} measures are associated with 'market moods' (several periods) rather than 'market emotions' (one period ahead)?

To investigate this relationship, we first computed the average EDA^{ton} signal across participants for each period t within a given market. We then constructed a five-period backward-looking moving average,

$$ma^{ton}_{EDA,t} = \frac{1}{5}\sum_{k=0}^{4} EDA^{ton}_{t-k},$$

computed separately within each market.[34] This procedure yields a market-specific, smoothed measure of market-level tonic arousal for each period.

Within each market, we then split observations at the median of $ma^{ton}_{EDA,t}$ into a "Top" and "Bottom" half, ensuring that each market contributes an equal number of observations to both groups. For each group, we computed the mean subsequent five-period forward return,

$$\frac{p_{t+5} - p_t}{p_t}.$$

Figure 3.33 shows a clear and statistically significant difference between the returns in periods below the EDA^{ton} median and the returns above the EDA^{ton} median. Higher emotional arousal measured by EDA^{ton} is associated with higher future returns ($p < 0.001$, permutation test for equality of distributions).

This is a rather strong and interesting exploratory finding. It could easily have been the case that a stronger market 'mood' is associated with rapid crashes. It is consistent with the hypothesis that the association between EDA^{ton} and future returns is slower (and hence mood-like)[35], while other kinds of emotion-returns association are either too noisy or too fast to be detected by our EDA^{ton} measures.

3.25 Debriefing and Language Modeling

We sought to gain an understanding of the types of strategies our participants used. To do so, we included a free-response question in the post-experiment survey in which we asked the participants to describe the strategy they used during the

[34] The first four periods of each market are excluded because a full five-period lookback is unavailable.

[35] Note in emotion literature "mood" is often associated with much longer timescales than we discuss in this work.

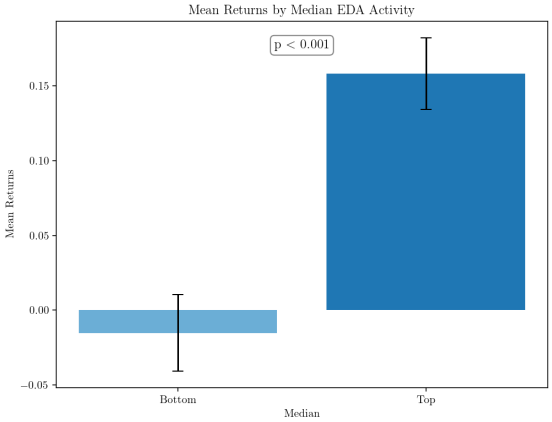


Figure 3.33: Mean Returns by EDA^{ton} median split. EDA^{ton} activity predicts future returns of the risky asset. Periods within each market were split into high and low activity based on a five-period moving average of EDA^{ton} activity. The bars show that the mean five-period forward return is higher for above-median EDA^{ton} activity.

market. The exact text of the question can be found in the *Post-experimental questionnaire* section of the supplementary material. For this analysis, we excluded the Experienced and High-Stakes markets to maintain consistency of participants' risk attitudes and experience. Of the 362 participants recruited, we received 302 responses.

Figure 4.3 presents a word cloud visualizing the raw frequency of words used in the responses, with a word's size corresponding to its frequency of use. The most frequently mentioned words are "sell" (140 occurrences), "buy" (133), "stock" (116), "price" (94), "market" (56), "low" (54), and "high" (48). A cursory glance shows that participants are keenly focused on buying and selling, and the appearance of "low" and "high" suggests that the classic "buy low / sell high" strategy is often employed. Indeed, it is a prominent category of strategy groups discussed below.

We then sought to group participants into strategy clusters. To do so, we computed the similarity of strategies using sentence embeddings and then performed a k-means clustering algorithm, after dimensionality reduction via principal component analysis (PCA). [36] Next, we stemmed [37] the words with `PorterStemmer` from Python's Natural Language Toolkit (`nltk`). This step combines similar words,

[36] We preprocessed strategies by removing all stop words, except for the word "no". Thirteen participants responded with a single-word answer, "no", which we still wanted to classify.

[37] Removed the prefixes and suffixes of the words.



Figure 3.34: **Word cloud of trading strategy** Larger words indicate higher use frequency in described strategies.

increasing their impact on the subsequent analysis. We then vectorized the individual strategies using sentence embeddings and ran a k-means cluster analysis on the results to classify the strategies into five groups. Figure 3.36 presents the clustering results, categorizing the reported strategies into five distinct groups. The first two principal components explained 20.8 percent of the variance.

The first group, which we designated “No Strategy", comprises participants who explicitly state that they did not follow a specific trading plan. Typical responses here were single-word answers of “no" or “none". The second group consisted of participants who stated that they did not follow a strategy, but further expressed that they at least attempted to follow market trends and/or acknowledged confusion (the word “guess" appeared three times in this group); effectively, Group 2 tried, while Group 1 gave up. We refer to this group as the “No / Trend" group. The third group stated that they tried to employ the strategy of “Buy Low / Sell High." Notably, most of the lowest median earners were assigned to one of these first three groups, while few of the highest earners were (see Figure 3.35).

Group 4 belongs to the “Low Risk" participant category. Participants assigned to this group were typically focused on the interest accrued on CASH holdings. Many participants in this group sold their shares as quickly as possible and chose to earn interest on their CASH holdings. Descriptions of strategies written by these individuals frequently referenced words such as “interest" and “cash." Group 5, we refer to as the “Buy Early" group. These participants typically attempted to capitalize on their forecast of market price increases by accumulating shares of stock in the early periods with the intention of selling them later. This is similar to the “Buy Low / Sell High" strategy in that participants tended to buy stocks early when

the price was low and sell later when the price was high, though more sophisticated. The “Buy Low / Sell High" group tended to look for market dips to buy throughout the market. In contrast, the “Buy Early" participants typically made most (or all) of their purchases in the early periods, seemingly anticipating the boom-and-bust nature of these markets. This group and their strategy were likely the drivers of the early surge in the average share path of high-earning median participants, as seen in Figure 3.11. The modal assignment to these two strategy groups was the high-earning median participants, while the lowest-earning median participants appeared in these groups the least.

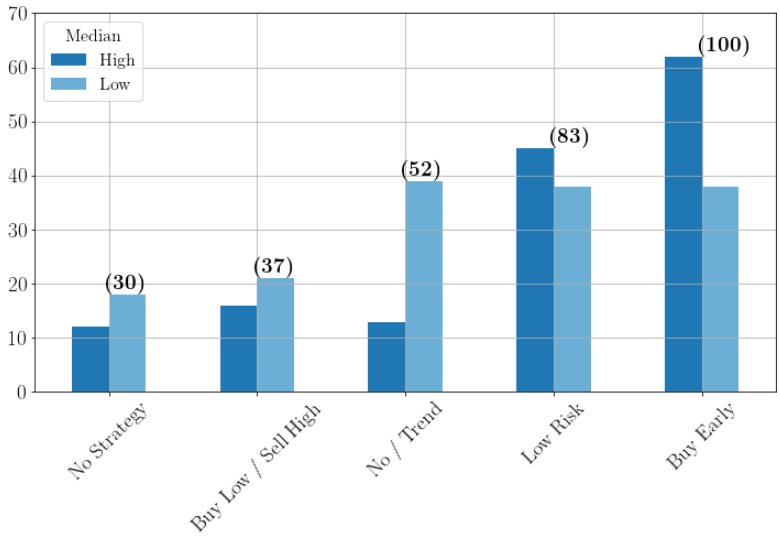


Figure 3.35: Strategy Group Assignments by Earning Median. Counts of strategy group assignment by earnings median. The count of all strategies assigned to a group is shown above the bar clusters. The strategy group `No Strategy` consists of those who responded with “No", “None", or the like. `No / Trend` mostly stated that they did not follow a strategy, but also indicated that they followed the market trend or guessed. `Buy Low / Sell High` are those that express that classic strategy. `Low-Risk` strategies focused on accruing interest and selling shares as soon as possible. `Buy Early` strategies involve acquiring the stock at an early stage with the intention of selling it later.

3.26 Time Varying Risk Premia

Our experimental design also enables us to investigate the widely held belief that bubbles are a rational consequence of changing individual risk premia over time, rather than some form of behavioral irrationality. To investigate this, we incorporated a separate task in which participants chose between multiple lotteries that utilized the Dynamically Optimal Sequential Experimentation (DOSE) methodology developed

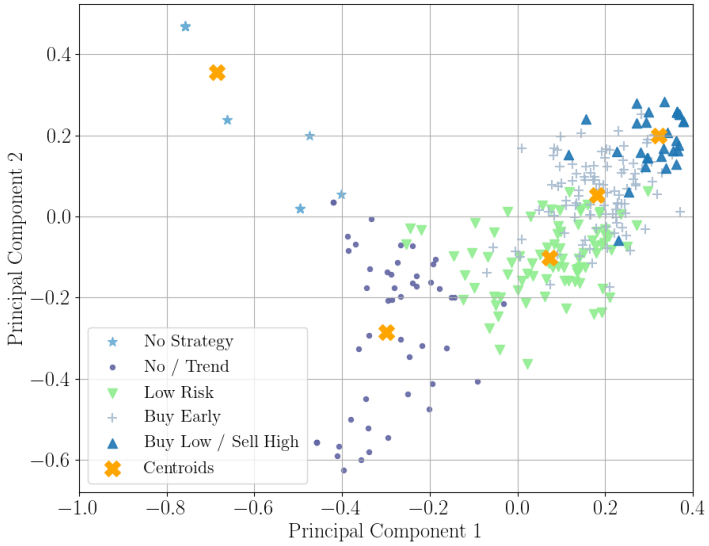


Figure 3.36: PCA of sentence embeddings with K-Means clustering. After vectorizing all strategies, we ran a k-means clustering algorithm and used principal component analysis (PCA) to reduce the vector space to two dimensions. The first principal component accounts for 13.7 percent of the variance, and the second component accounts for 7.1 percent.

by Chapman et al., 2024 to measure individual subject risk premia at each market period. The task is encompassed in the fourth window in the sequence of windows given to the participants (see section 4 for details). This measurement enabled us to obtain an exogenous measure of risk at each time period and directly compare it to price movements to determine if there is any relationship.

Figure 3.37 shows the plots of the average price for our risk measure ρ (risk-aversion) and Figure 3.38 shows the plots of the average price for our risk measure μ (risk-consistency), obtained from DOSE. We then also fit a regression, given the quadratic shape of our price curve, in order to see if the lack of relationship between ρ and average price may be due to a misspecification.

$$R_t = \beta_0 + \beta_1 \rho_t + \beta_2 \rho_t^2 + \epsilon_t$$

In our experimental context, despite the findings of (Guiso, Sapienza, and Zingales, 2018; Cohn et al., 2015), we find little evidence supporting the idea that there is a significant relationship between risk premia and the formation of bubbles. Despite some partial correlations with next-period returns in specific specifications, the effects are small and do not appear in the specifications we describe above..

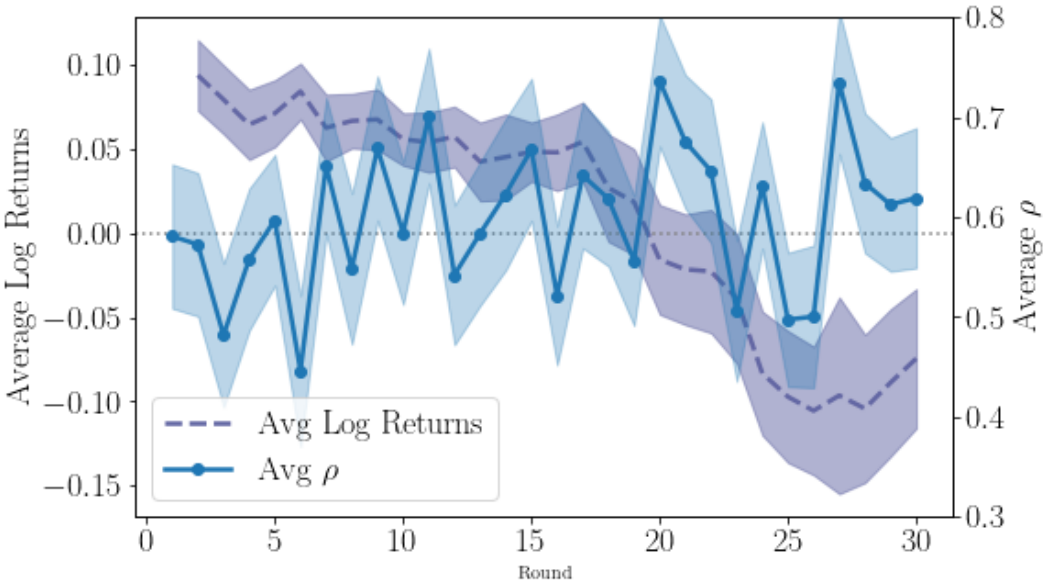


Figure 3.37: Log Return R_t vs ρ Across Periods. ρ refers to the risk-aversion coefficient estimated for each subject with the risk elicitation task (higher values indicate greater risk aversion). The parameter is estimated using the procedure described in (Chapman et al., 2024). There is no statistically significant relationship between ρ and R_t, suggesting that time-varying risk premias are not the drivers of bubble-like returns.

We find no evidence that time-varying risk premia explain bubble formation. The quadratic regression yielded an R^2 of 0.000, an adjusted R^2 of -0.003, and an F-statistic of 0.902, indicating no explanatory power. The linear and quadratic term estimates were ($\beta_{linear} = -0.004, p = 0.767$; $\beta_{quadratic} = 0.004, p = 0.876$) close to zero These findings provide contradictory evidence to the time-varying risk premia hypothesis of bubble formation.

Power to Detect Correlations Between Risk Preferences and Returns

We conducted a sensitivity-style power calculation to assess the smallest correlation between elicited risk preferences and market returns that our design could plausibly detect. This calculation provides context for interpreting the near-zero relationship between time-varying risk premia and returns. The raw panel contains 20,160 participant-round observations, corresponding to 672 participants each completing 30 exogenous risk-preference elicitations. These observations are nested within 28 independent markets.

Table 3.30: EDA Peak is correlated, at the individual-trial level, with risk-aversion ρ

	ρ
$EDA_{i,t}^{peak}$	-0.176*
	(0.080)
Constant	0.710***
	(0.051)
Observations	2219
R^2	0.007

Notes: Coefficients are estimated by OLS with tercile fixed effects; cluster-robust standard errors (clustered by subject) in parentheses. Panel A reports regression coefficients; the intercept shifts for Medium and High terciles are absorbed into the fixed effects and not shown. Panel B reports the total effect of $EDA_{i,t}^{peak}$ for each tercile, computed as the sum of the main effect and the relevant interaction term, with standard errors derived from the full variance-covariance matrix. 95% confidence intervals are based on the normal approximation. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

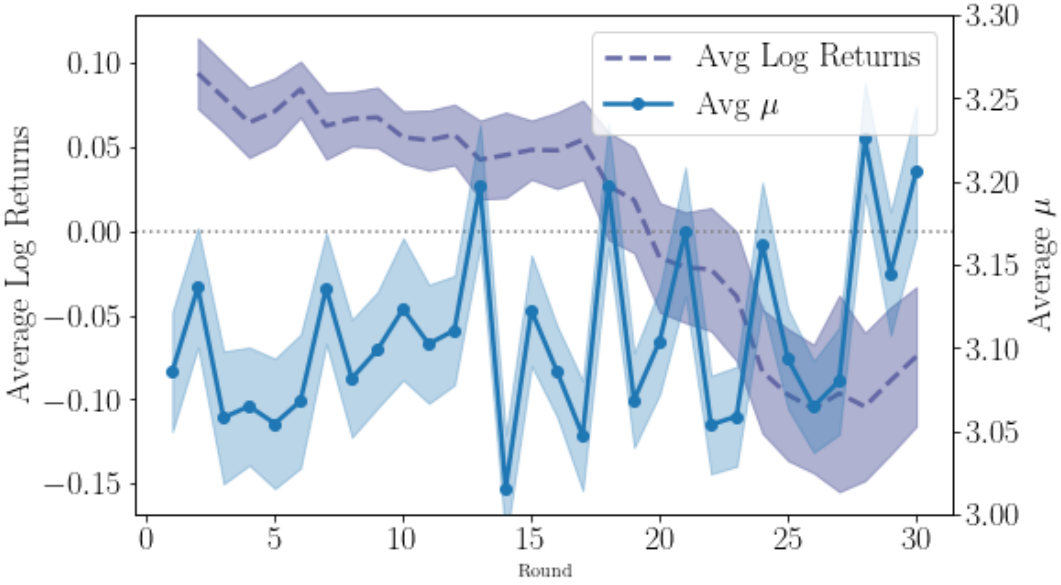


Figure 3.38: Log Return R_t vs μ Across Periods. μ refers to the choice consistency coefficient estimated for each subject during the risk elicitation task (higher values indicate greater consistency of choice). The parameter is estimated using the procedure described in (Chapman et al., 2024).

For a correlation test, we use Fisher's z transformation,

$$z_r = \operatorname{arctanh}(r),$$

with approximate standard error

$$SE(z_r) = \frac{1}{\sqrt{n_{\text{eff}} - 3}}.$$

For a two-sided test at 90% confidence, $\alpha = 0.10$ and $z_{1-\alpha/2} = 1.645$. The minimum detectable correlation is approximated by

$$r_{\min} = \tanh\left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\sqrt{n_{\text{eff}} - 3}}\right),$$

where $1 - \beta$ denotes statistical power.

Under a naive independent-observation calculation using all participant-round observations,

$$n_{\text{eff}} = 672 \times 30 = 20{,}160,$$

the design could detect very small correlations:

$$r_{\min}^{80\%} = \tanh\left(\frac{1.645 + 0.842}{\sqrt{20{,}160 - 3}}\right) \approx 0.018,$$

and

$$r_{\min}^{90\%} = \tanh\left(\frac{1.645 + 1.282}{\sqrt{20{,}160 - 3}}\right) \approx 0.021.$$

However, this calculation is best interpreted as an upper-bound precision benchmark. Risk preferences are repeatedly measured within participants, and market returns are common to participants within a given market-period. Therefore, the effective sample size is smaller than the raw number of participant-round observations. To account for repeated measurements of risk preferences, we compute design-effect-adjusted effective sample sizes:

$$n_{\text{eff}} = \frac{N}{1 + (m - 1)\rho_{\text{ICC}}},$$

where $N = 20{,}160$, $m = 30$ repeated measurements per participant, and ρ_{ICC} is the within-participant intraclass correlation of repeated risk-preference measures.

For plausible within-participant correlations, this yields:

ρ_{ICC}	n_{eff}	$r_{\text{min}}^{80\%}$	$r_{\text{min}}^{90\%}$
0.00	20,160	0.018	0.021
0.05	8,229	0.027	0.032
0.10	5,169	0.035	0.041
0.20	2,965	0.046	0.054
0.30	2,078	0.055	0.064
0.50	1,301	0.069	0.081
0.70	946	0.081	0.095
0.90	744	0.091	0.107

Thus, even after allowing for substantial within-participant dependence, the design is powered to detect small participant-level correlations between risk preferences and returns. For example, with within-participant correlations between $\rho_{\text{ICC}} = 0.10$ and $\rho_{\text{ICC}} = 0.50$, the effective sample size ranges from approximately 5,200 to 1,300 observations, corresponding to minimum detectable correlations of approximately $r = 0.04$ to $r = 0.08$ at 80–90% power.

A separate conservative benchmark is to treat only market-period returns as independent observations. With 28 markets and 30 periods, this gives

$$n_{\text{eff}} = 28 \times 30 = 840.$$

Under this benchmark,

$$r_{\text{min}}^{80\%} = \tanh\left(\frac{1.645 + 0.842}{\sqrt{840 - 3}}\right) \approx 0.086,$$

and

$$r_{\text{min}}^{90\%} = \tanh\left(\frac{1.645 + 1.282}{\sqrt{840 - 3}}\right) \approx 0.101.$$

Finally, under the most conservative benchmark, if inference is conducted only at the level of independent markets, then $n_{\text{eff}} = 28$. In that case,

$$r_{\text{min}}^{80\%} = \tanh\left(\frac{1.645 + 0.842}{\sqrt{28 - 3}}\right) \approx 0.46,$$

and

$$r_{\text{min}}^{90\%} = \tanh\left(\frac{1.645 + 1.282}{\sqrt{28 - 3}}\right) \approx 0.53.$$

Therefore, the interpretation of the null result depends on the level of inference. At the participant-round level, the design is powered to detect very small associations,

although this requires appropriate clustering for repeated observations within participants and common market-period returns. At the market-period level, the design is powered to detect correlations of approximately $r = 0.09$–0.10. At the fully conservative market level, the design is powered only to detect larger correlations, $r \approx .50$.

The most relevant empirical benchmarks are studies of counter-cyclical risk aversion rather than studies of static risk preferences alone. Cohn et al. (2015) show that financial professionals in an experiment primed with a bust invest 17–22% less in risky assets than those primed with a boom, corresponding to roughly a 10–12% reduction in risky investment. However, their result failed to replicate in two student and online samples (Alempaki, Starmer, and Tufano, 2019; König-Kersting and Trautmann, 2018).

In a lab experiment Guiso, Sapienza, and Zingales (2018) found that emotional priming with a scary film clip increased risk premia by 27%.[38] Guiso, Sapienza, and Zingales (2018) also found that measured risk premia among Italian bank clients rose sharply after the 2008 crisis: their survey measured the unincentivized certainty equivalent of a 50% chance of winning €10,000. The mean risk premium (the reduction from the expected value of €5,000) increased from 973 euros to 2,215 euros, which implies a doubling of the average absolute risk aversion coefficient.

Our design tests a more direct within-market implication of this mechanism: whether period-by-period variation in experimentally elicited risk premia explains contemporaneous or next-period market returns. Since our market-period design is powered to detect correlations of approximately $r = 0.09$–0.10, and participant-round specifications can detect even smaller associations under repeated-measures assumptions, the near-zero estimates provide evidence against the hypothesis that time-varying risk premia of the magnitude documented in this literature are a primary driver of returns in our experimental markets.

This interpretation is consistent with related experimental asset-market evidence showing that risk preferences may affect individual demand without necessarily

[38] Several other studies test whether frightening natural events change risk preferences (measured with lab-in-field simple financial risk choices). There is evidence for increased risk-aversion after floods in Indonesia (Cameron and Shah, 2015), political violence in Kenya (Jakiela and Ozier, 2019), and drug violence in Mexico (Brown et al., 2019). Hanaoka, Shigeoka, and Watanabe (2018) found decreased risk-aversion after an earthquake in Japan. Shum and Xin (2022) investigated the short-term effect of near-miss driving events– hard turns and hard braking– which happen every couple of days. They found that near misses increased measures of driving risk-aversion, such as reduced use of cell phones and highways, for about two weeks.

translating into aggregate price movements. Filippin et al. (2023) study risk aversion and information aggregation in experimental two-state (prediction) markets and find that risk aversion affects individual demands but not market prices. Similarly, Mantovani and Filippin (2026) reports no evidence that risk preferences significantly affect prices in prediction markets. Thus, even if risk preferences influence individual choices, their relationship with aggregate market prices or returns may be attenuated by market aggregation.

3.27 Physiological Arousal and Trading Behavior

We previously documented a reduced-form relationship between physiological arousal and de-risking behavior. Here, we decompose this effect into distinct trading actions to identify whether arousal predicts intentions to sell, successful execution, or net selling within a period.

We explored how physiological arousal (measured by in-period EDA peak EDA^{peak}) related to individuals' trading behavior, specifically de-risking behavior. We find that those who had heightened physiological arousal placed more aggressive sell orders. To uncover this, we conducted an exploratory analysis where we compared derisking on different levels, each represented by a binary variable that takes a value of 1 if the given behavior was observed. Logit regressions were then run on these variables with EDA^{peak} as the independent variable. Participants who attempted to sell more shares than they bought are referred to as Net Sellers. We find that EDA^{peak} is not predictive of Net Selling, though the relationship is also positive. Likewise, EDA^{peak} does not significantly predict an attempt to sell ($\beta = -0.0678; p = 0.663$). However, EDA^{peak}'s effect on participants executing a sell is significant ($\beta = 0.602; p = 0.005$). This dissociation suggests that the decision to attempt a sale may be driven by multiple factors — such as hedging or arbitraging expected returns against the fixed return of cash — whereas a strong motivation to derisk may lead participants to price the asset more aggressively, thereby increasing the likelihood of completing the trade.

Table 3.31: EDA Derisk Logistic Regressions

	Executed Sell	Attempted Sell	Net Sell
EDA Peak	0.602**	0.328*	0.0503
	(0.212)	(0.166)	(0.169)
Constant	-1.980***	0.564***	-0.945***
	(0.153)	(0.115)	(0.119)
Observations	2280	2280	2280
Pseudo R2	0.004	0.001	0.000

Notes: All are logit regressions with robust standard errors, clustered by subject (shown in parentheses). p-values are shown in square brackets. Executed Sell takes a value of 1 when a participant manages to sell shares. Attempted Sale is a boolean that takes a value of 1 when a participant places at least one sell order (regardless of quantity and price). Net Sell takes a value of 1 when a participant attempts to sell more shares than they bought. Attempted Trade takes a value of 1 if a participant in a given period places any order at all (buy or sell) and zero otherwise. Across the 2,280 round-level observations, the share of zeros is 82.98% for Executed Sell, 60.22% for Attempted Sell, and 71.36% for Net Sell.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Bubble Metrics

We use several common measures of market bubbles to compare price paths of our treatments. Most of the measures are intended for use in continuous auctions in which trades are executed instantly and the market price can fluctuate during a given period. With the call market setup that we employ, there is a singular price per period, which simplifies measures that involve the mean or median price for a period. Further, several of these measures are designed around a changing fundamental value. Since our fundamental value is constant (F = 14) for all periods (t=30), we can further simplify these measures.

The measures are outlined briefly below.

- Relative Absolute Deviation (RAD): A measure of the average price deviation from the fundamental value within a given market.
- Average Bias (AB): A measure of how far the market price is, on average, from the fundamental value.
- Total Dispersion (TD): A measure of the absolute difference between the market price and the fundamental value.

- Price Amplitude (PA): A measure of the largest difference in price deviations from the fundamental value.
- Duration (DUR): The longest number of periods in which the price deviation from the fundamental value is increasing.

Relative Absolute Deviation (RAD)

Relative Absolute Deviation was introduced by (Stöckl, J. Huber, and Kirchler, 2010) in an attempt to develop a measure that can be compared across market configurations. It is calculated by:

$$\text{RAD} = \frac{\sum_{t=1}^{T} v_t \left(\frac{|P_t - 14|}{14} \right)}{\sum_{t=1}^{T} v_t}$$

Where p_t is the market price in period t and v_t is the number of shares traded in period t.

Average Bias (AB)

Average Bias (Haruvy and Charles N. Noussair, 2006b) measures how far the market price is, on average, from the fundamental value. As given by Haruvy, the expression is:

$$\text{AB} = \frac{1}{T} \sum_{t=1}^{T} \text{Median}(P_t) - 14$$

Where $\text{Median}(P_t)$ is the median price realized during period t and (14) is the fundamental value for period t. Our market design allows us to simplify those terms to constants, and the expression becomes a simple average of the differences between price and fundamental value.

$$\text{AB} = \frac{1}{T} \sum_{t=1}^{T} P_t - 14$$

Total Dispersion (TD)

Haruvy and Noussair also introduced Total Dispersion, which is very similar to the AB measure except that it is a measure of the absolute difference between the market price and the fundamental value (14).

$$\text{TD} = \sum_{t=1}^{T} |P_t - 14|$$

Price Amplitude (PA)

Price Amplitude (D. P. Porter and V. L. Smith, 1995) is a measure of the largest difference in price deviations from the fundamental value (14).

$$\text{PA} = \max_{t \in T}(P_t - 14_t) - \min_{t \in T}(P_t - 14)$$

Because our constant fundamental value cancels out, this measure reduces to the difference between the maximum price and the minimum price achieved in a market.

$$\text{PA} = \max_{t \in T}(P_t) - \min_{t \in T}(P_t)$$

Duration (DUR)

Duration (D. P. Porter and V. L. Smith, 1995) is the longest number of periods in which the price deviation from the fundamental value (14) is increasing.

$$\text{DUR} = \max\left(D \ : \ P_t - 14 < P_{t+1} - 14 < \ldots < P_{t+D-1} - 14\right)$$

Since the fundamental value is constant, all the (14)'s cancel out, and the duration is simply the longest streak of increasing prices.

$$DUR_s = \max\left(D \ : \ P_t < P_{t+1} < \ldots < P_{t+D-1}\right)$$

Table 3.32: Summary statistics.

		All	Lab	Prolific	Experienced	High Stakes
N		514	90	424	93	59
Age	(mean)	35.77	27.05	37.53	37.31	40.64
Quiz Grade	(mean)	2.86	3.42	2.74	3.31	2.69
Quiz 1	(mean)	0.66	0.60	0.68	0.80	0.69
Quiz 2	(mean)	0.53	0.53	0.53	0.57	0.53
Quiz 3	(mean)	0.56	0.79	0.51	0.60	0.47
Quiz 4	(mean)	0.49	0.76	0.43	0.58	0.39
Quiz 5	(mean)	0.62	0.74	0.59	0.76	0.61
Market Bonus	(mean)	4.94	3.98	5.14	3.37	16.84
Forecast Bonus	(mean)	0.14	0.22	0.13	0.03	0.27
Risk Bonus	(mean)	0.99	0.98	0.99	1.05	0.94
Total Bonus	(mean)	6.66	5.76	6.85	5.28	18.73
Risk Gen (pre)	(mean)	3.97	3.69	4.03	4.04	3.76
Risk Gen (post)	(mean)	3.79	3.65	3.82	3.94	3.57
Invest 100 (pre)	(mean)	49.58	54.27	48.63	47.80	48.55
Invest 100 (post)	(mean)	48.55	53.80	47.48	48.59	48.77
Race	AI/AN	1	1	0	0	0
	Asian or Asian Indian	107	44	63	18	4
	B/AA.	90	1	89	21	7
	Other	13	4	9	1	1
	Prefer Not to Say	1	1	0	0	0
	White caucasian	266	29	237	51	46
Gender	Female	213	39	174	37	32
	I prefer not to say	1	1	0	0	0
	Male	259	39	220	54	25
	Non-binary	5	1	4	0	1

AI/AN stands for *American Indian or Alaska Native*. B/AA stands for *Black or African American*. Note 36 subjects decided not to answer the gender question/

Table 3.33: Individual Executed Trades and Bonuses for Experienced Participants

1				2				3				4			
Att	Exe	FB	MB	Att	Exe	FB	MB	Att	Exe	FB	MB	Att	Exe	FB	MB
31	4	3	9	20	1	14	10	9	2	8	12	7	1	10	15
42	5	6	3	44	5	6	11	18	7	17	21	35	7	3	7
64	38	13	5	89	24	6	7	60	8	17	1	84	22	6	10
8	4	18	18	9	4	16	13	12	8	8	13	15	5	12	18
18	9	14	15	19	5	3	2	10	6	10	12	-	-	-	-
57	12	14	12	68	5	8	4	84	2	16	13	-	-	-	-
24	7	20	3	21	10	14	10	25	11	19	6	-	-	-	-
19	10	7	23	17	3	8	8	9	5	3	19	-	-	-	-
63	9	10	3	63	7	1	5	52	3	6	16	-	-	-	-
11	2	2	15	19	5	8	13	23	4	16	9	-	-	-	-
38	4	3	3	20	11	12	3	10	4	22	21	-	-	-	-
33	3	16	6	35	12	6	17	24	8	9	8	-	-	-	-
49	4	7	2	48	8	8	3	35	5	10	4	-	-	-	-
23	10	18	14	29	6	10	8	3	2	16	20	-	-	-	-
17	3	18	21	13	8	20	20	19	6	18	9	-	-	-	-
54	11	6	10	46	9	6	1	58	13	2	2	-	-	-	-
13	4	6	11	18	3	12	10	17	2	20	19	-	-	-	-
32	10	6	4	24	11	3	4	19	7	10	16	-	-	-	-
63	5	16	1	47	5	8	2	33	6	12	4	-	-	-	-
30	4	8	14	30	6	8	17	29	2	18	13	-	-	-	-
40	4	14	13	38	14	2	3	15	6	9	11	-	-	-	-
29	5	14	14	22	6	3	11	8	3	14	18	-	-	-	-
37	13	9	5	25	9	8	6	7	5	3	21	-	-	-	-
58	25	14	13	48	23	6	6	33	13	14	12	-	-	-	-
21	9	6	2	21	4	22	17	-	-	-	-	-	-	-	-
29	10	4	4	21	7	11	5	-	-	-	-	-	-	-	-
18	10	19	5	42	14	6	2	-	-	-	-	-	-	-	-
18	5	4	15	22	7	9	10	-	-	-	-	-	-	-	-
20	1	14	14	20	2	8	16	-	-	-	-	-	-	-	-
22	11	13	19	44	4	10	5	-	-	-	-	-	-	-	-
27	5	2	5	40	9	3	7	-	-	-	-	-	-	-	-
43	10	8	8	52	17	10	3	-	-	-	-	-	-	-	-
35	8	6	10	28	6	20	5	-	-	-	-	-	-	-	-
54	13	19	19	23	7	18	22	-	-	-	-	-	-	-	-
41	8	8	9	26	4	8	14	-	-	-	-	-	-	-	-
29	12	13	7	32	12	2	3	-	-	-	-	-	-	-	-
53	7	2	21	-	-	8	11	-	-	-	-	-	-	-	-
39	17	5	5	37	13	3	14	-	-	-	-	-	-	-	-
16	8	16	11	5	4	6	15	-	-	-	-	-	-	-	-
31	8	21	18	40	6	16	3	-	-	-	-	-	-	-	-
39	5	4	4	67	12	2	5	-	-	-	-	-	-	-	-
24	8	14	18	9	2	13	14	-	-	-	-	-	-	-	-
8	6	18	19	14	8	13	7	-	-	-	-	-	-	-	-
41	10	2	12	41	16	2	1	-	-	-	-	-	-	-	-
34	6	6	12	62	6	6	9	-	-	-	-	-	-	-	-
24	4	7	14	19	6	6	14	-	-	-	-	-	-	-	-
38	14	12	8	31	15	10	6	-	-	-	-	-	-	-	-
32	1	3	8	22	8	8	10	-	-	-	-	-	-	-	-
17	9	14	20	4	1	13	12	-	-	-	-	-	-	-	-
26	6	6	15	7	4	13	15	-	-	-	-	-	-	-	-
77	20	7	8	30	9	10	2	-	-	-	-	-	-	-	-
32	6	11	1	23	12	21	20	-	-	-	-	-	-	-	-
31	10	17	8	19	4	10	11	-	-	-	-	-	-	-	-
59	9	14	11	54	12	10	7	-	-	-	-	-	-	-	-
15	4	9	10	8	3	6	13	-	-	-	-	-	-	-	-
25	9	7	22	14	6	13	9	-	-	-	-	-	-	-	-
14	5	5	15	9	3	8	15	-	-	-	-	-	-	-	-
39	6	21	19	68	22	18	22	-	-	-	-	-	-	-	-
35	10	7	17	21	5	10	18	-	-	-	-	-	-	-	-
27	4	14	4	27	6	6	6	-	-	-	-	-	-	-	-
31	7	6	16	18	7	8	9	-	-	-	-	-	-	-	-
30	7	2	7	14	6	18	10	-	-	-	-	-	-	-	-
28	7	12	12	10	5	17	19	-	-	-	-	-	-	-	-
7	6	5	10	38	23	2	4	-	-	-	-	-	-	-	-
26	8	4	8	21	8	3	1	-	-	-	-	-	-	-	-
27	6	14	11	18	9	22	11	-	-	-	-	-	-	-	-
28	9	14	8	6	3	17	18	-	-	-	-	-	-	-	-
32	12	12	8	31	12	18	8	-	-	-	-	-	-	-	-
85	36	9	1	83	13	6	8	-	-	-	-	-	-	-	-
16	5	13	16	29	9	12	12	-	-	-	-	-	-	-	-
33	6	4	5	40	8	17	7	-	-	-	-	-	-	-	-
64	13	2	3	92	31	1	1	-	-	-	-	-	-	-	-

Notes: Att: number of attempted trades during a market. Exe: number of executed trades during a market. FB: forecast bonus rank, with 1 indicating the lowest rank. MB: market bonus rank, with 1 indicating the lowest rank. Line breaks represent a participant's second, third, or fourth market.

3.28 Summary Stats

Table 3.34: Market level summary statistics.

Type	Date	*N*	Float	RAD	Bias	Disp	PA	Dur	Pk. Price	Pk. Round
Standard	2024-03-07	21	84	1.86	25.76	850	62	19	77	20
	2024-03-14	20	80	0.77	13.94	460	32	19	47	23
	2024-04-04	16	32	2.67	41.27	1362	87	23	102	24
	2024-06-04	22	88	1.23	18.30	604	32	18	47	19
	2024-06-20	22	88	0.51	7.12	235	14	3	28	17
	2024-07-02	21	84	0.59	8.09	273	21	16	34	19
	2024-07-03	17	68	0.95	14.33	473	34	13	48	14
	2024-07-08	18	72	0.75	11.58	384	31	12	44	13
	2024-07-09	16	64	0.56	7.76	260	13	6	26	11
	2024-07-10	23	92	1.16	15.85	525	38	14	51	19
	2024-07-12	24	96	2.12	31.79	1049	67	24	82	25
	2024-07-15	19	76	0.65	9.61	317	26	9	40	19
	2024-07-17	16	64	0.79	13.03	430	26	7	40	21
	2024-07-19	18	72	1.18	15.73	519	40	18	55	19
	2024-07-24	15	60	0.70	11.00	363	29	9	43	23
	2024-07-26	20	80	1.94	26.88	887	66	16	80	20
	2024-07-31	13	52	1.47	21.03	704	52	11	64	18
	2024-08-07	20	80	1.18	15.76	520	27	22	42	24
	2024-08-09	21	84	0.91	13.21	436	23	9	38	10
Experienced	2024-09-06	15	60	1.15	14.15	467	38	9	52	10
	2024-09-10	21	84	1.58	29.91	987	68	21	83	22
	2024-09-12	13	52	0.51	5.61	197	17	4	29	12
	2024-09-28	22	88	0.59	7.24	239	22	5	36	6
	2024-09-30	22	88	1.16	18.45	615	54	10	67	11
High Stakes	2024-09-19	17	68	0.25	3.52	116	7	2	21	15
	2024-09-20	12	48	1.23	18.06	596	33	10	48	21
	2024-09-24	13	52	0.98	14.06	464	34	15	49	27
	2024-09-25	17	68	0.41	6.21	209	12	6	25	25
Averages										
Standard		19.05	74.53	1.16	16.95	560.58	37.89	14.11	52.00	18.84
Experienced		18.60	74.40	1.00	15.07	501.00	39.80	9.80	53.40	12.20
High Stakes		14.75	59.00	0.72	10.46	346.25	21.50	8.25	35.75	22.00

Std means Standard market. Exp is the Experienced markets. HS is the High-Stakes markets.

Table (3.34) shows bubble metrics, number of participants, peak period, and price for all markets.

3.29 Forecast Rationality

Figure 3.39 shows histograms of current-period and 2-period forecast errors. Means are slightly positive for both and about the same. The distribution for current period errors is more leptokurtic and the 2-period errors. This makes senses because current period prices are easies to predict.

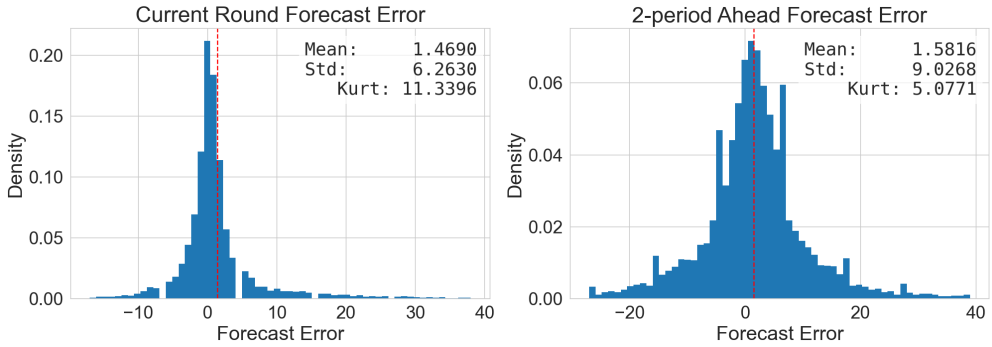


Figure 3.39: Histograms of forecast errors for current round and 2 periods ahead.

Table 3.35: Summary Statistics of Forecasts and Forecast Errors

Variable	Mean	SD	Min	Q1	Median	Q3	Max
Price	31.31	16.42	12	19.00	26.00	40.00	102
f_t^t	29.65	17.80	1	18.00	25.00	38.00	147
f_t^{t+2}	30.69	18.18	1	19.00	26.00	39.00	137
f_t^{t+5}	33.80	20.02	1	20.00	29.00	42.00	180
f_t^{t+10}	35.56	21.05	1	21.00	30.00	45.00	190
E_t^t	1.67	8.70	-84	-1.00	0.00	2.00	92
E_t^{t+2}	1.27	9.36	-68	-2.00	0.00	3.00	95
E_t^{t+5}	-1.23	12.01	-107	-6.00	-2.00	3.00	93
E_t^{t+10}	-3.63	15.14	-132	-10.00	-3.00	3.00	77

Tests of Competing Models of Forecasts (Continued)

As a more demanding joint test, we regress both experimental bonus measures on all five expectation-formation parameters simultaneously (Table 3.38). In the OLS specifications, these parameters explain a modest share of the variation in bonuses ($R^2 = 0.084$ for market bonus; $R^2 = 0.059$ for forecast bonus). Including session fixed effects increases explanatory power, slightly for the market bonus ($R^2 = 0.094$) and substantially for the forecast bonus ($R^2 = 0.219$). Corresponding

Table 3.36: Summary Statistics of Parameter Estimates and Fit by Forecasting Model for Experienced Markets

Model (Parameter)	Mean	SD	Mean R^2	N
Adaptive Expectations (σ)	1.036	0.446	0.548	91
Extrapolation (ϕ)	0.389	0.447	0.386	90
Overreaction (β)	0.121	0.395	0.162	92
DAE: Level Error (η)	0.772	0.364	0.754	92
DAE: Trend Error (γ)	0.483	0.330		

Notes: Mean and SD denote the mean and standard deviation of the model parameter estimate across participants for each model, respectively. Mean R^2 reports the average of fit across participants, and N indicates the number of participants with non-missing estimates. The slight variation in sample size N across models is due to differences in model specifications (e.g., lag structures), which prevented estimation for a small subset of participants with insufficient observations. For the Double Adaptive Expectations model, results are reported separately for the level forecast error and level change error parameters.

Table 3.37: Summary Statistics of Parameter Estimates and Fit by Forecasting Model for High-Stakes Markets

Model (Parameter)	Mean	SD	Mean R^2	N
Adaptive Expectations (σ)	0.746	0.407	0.463	58
Extrapolation (ϕ)	0.273	2.520	0.162	58
Overreaction (β)	-0.214	0.337	0.175	59
DAE: Level Error (η)	0.608	0.369	0.572	58
DAE: Trend Error (γ)	0.644	0.569		

Notes: Mean and SD denote the mean and standard deviation of the model parameter estimate across participants for each model, respectively. Mean R^2 reports the average of fit across participants, and N indicates the number of participants with non-missing estimates. The slight variation in sample size N across models is due to differences in model specifications (e.g., lag structures), which prevented estimation for a small subset of participants with insufficient observations. For the Double Adaptive Expectations model, results are reported separately for the level forecast error and level change error parameters.

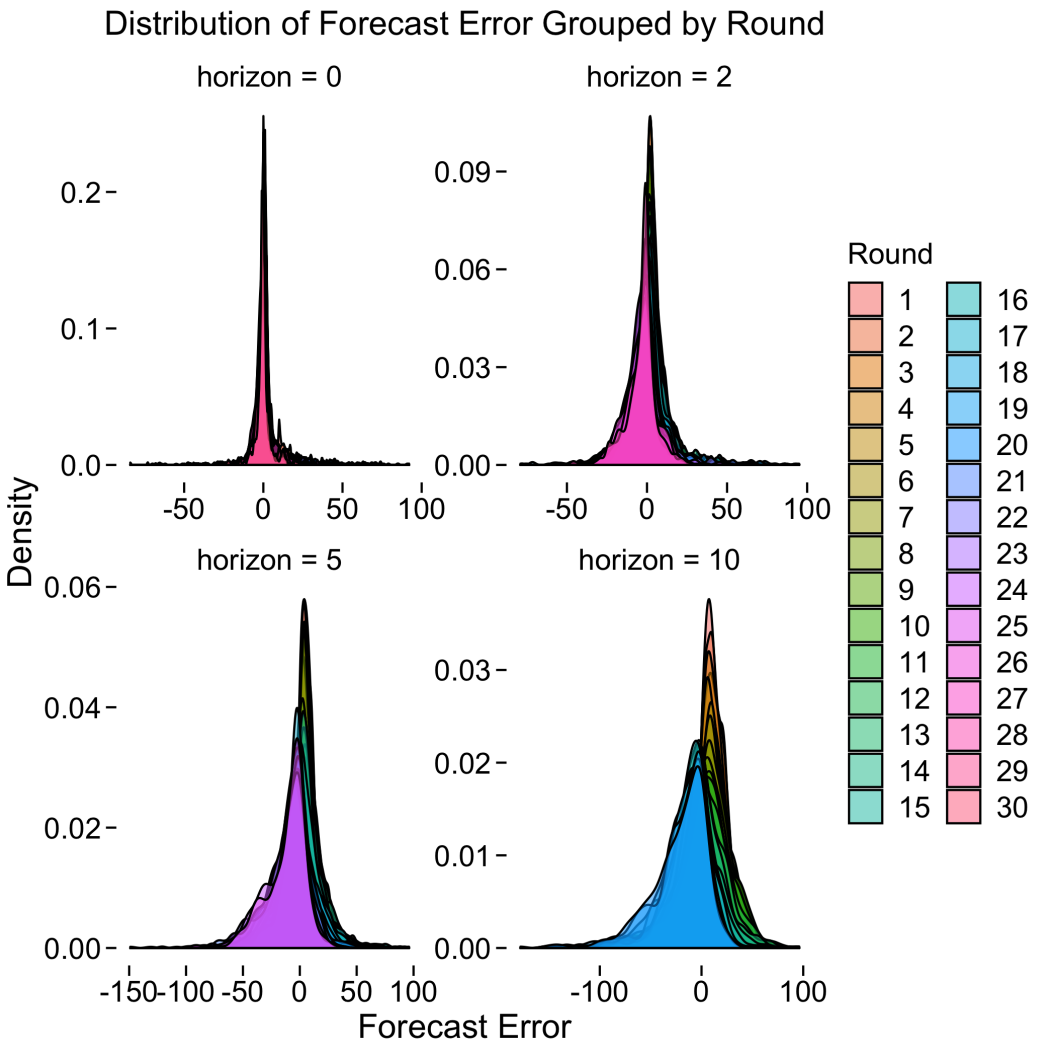


Figure 3.40: Distribution of Forecast Error Grouped by Period

tables for experiences and high-stakes sessions can be found in Table 3.39 and Table 3.40 respectively.

The adaptive parameter positively predicts the forecast bonus in both OLS and fixed-effects models, but lacks precision for the market bonus. Anchoring heavily on recent prices improves short-term prediction accuracy, yet it does not reliably translate into profitable trading decisions. Extrapolation positively correlates with both bonuses under OLS but becomes small and statistically insignificant once session fixed effects are included. This indicates that the apparent payoff to trend-chasing is driven by overall market conditions rather than individual skill.

The overreaction parameter (β) strongly predicts the market bonus across all specifications. A higher β corresponds to more cautious belief revision, which likely

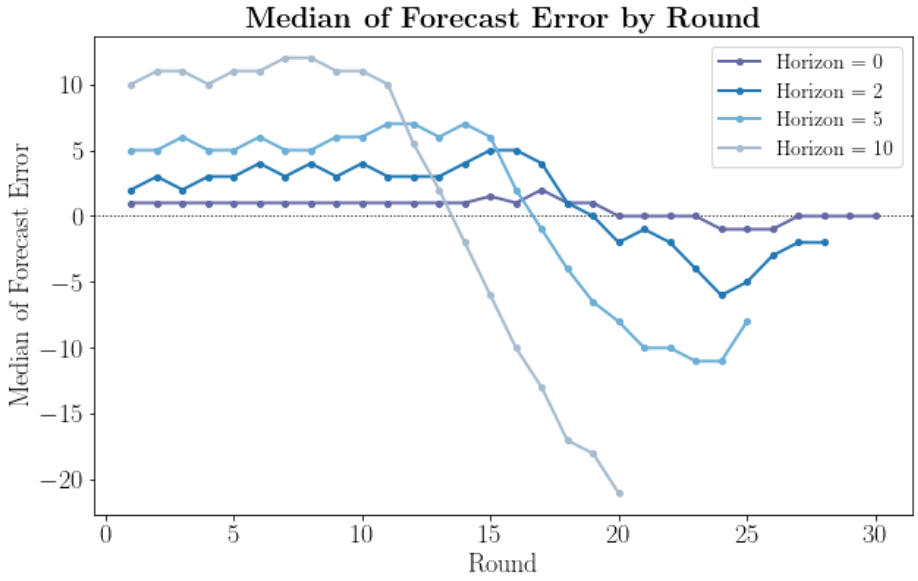


Figure 3.41: Median forecast error by period for each horizon. Forecast error is defined as $E^h_{i,t} = P_{t+h} - \tilde{f}^{t+h}_t$ similar to Eq 3.3. Here $\tilde{f}$ indicates the median forecast of all participants in the period t for horizon h. A positive forecast error indicates an underestimate while a negative error indicates an overestimate.

protects participants from suboptimal trades like buying at market peaks. Interestingly, its relationship with the forecast bonus flips from positive in OLS to significantly negative in the fixed-effects model. This reversal highlights a clear trade-off: while underreacting safeguards trading profits, it causes agents to lag behind in fast-moving markets, thereby hurting short-term forecast accuracy. The two DAE terms lack general precision, though the forecast-error component is significant for the market bonus under OLS, likely because the combined responses to price levels and trends offset each other in determining overall payoffs.

Table 3.38: OLS and Fixed-Effects Regression Results: Market Bonus vs. Forecast Bonus

	Market Bonus		Forecast Bonus	
	(1) OLS	(2) FE	(3) OLS	(4) FE
Adaptive (σ)	0.00252	0.0328	0.0759**	0.0690**
	(0.0486)	(0.0387)	(0.0310)	(0.0328)
Extrapolation (ϕ)	0.0514*	0.0451	0.0228**	0.00387
	(0.0282)	(0.0349)	(0.00882)	(0.00561)
Overreation (β)	0.109***	0.0983***	0.0270*	-0.0316**
	(0.0321)	(0.0330)	(0.0148)	(0.0129)
DAE: Forecast Err. (η)	0.136**	0.131	-0.00257	0.0158
	(0.0677)	(0.0886)	(0.0323)	(0.0294)
DAE: Change Err. (γ)	0.0723	0.108	-0.0227	0.0274
	(0.0550)	(0.0722)	(0.0174)	(0.0204)
Constant	1.303***	1.260***	0.0588**	0.0285
	(0.0709)	(0.0937)	(0.0289)	(0.0538)
Observations	341	341	341	341
R^2	0.084	0.094	0.059	0.219

Notes: Dependent variables are $\log(1 + \text{bonus})$. This table compares OLS and fixed-effects (FE) regression results for market bonuses and forecast bonuses. Adaptive expectations, extrapolation, overreaction, and double adaptive expectation parameters are individually estimated from their respective forecasting models and then used as regressors. Columns (1) and (3) report OLS models, robust standard errors in parentheses, and overall R^2, and columns (2) and (4) include session fixed effects, the within R^2 (the share of variation explained after removing session-level means), and standard errors clustered by session in parentheses.

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

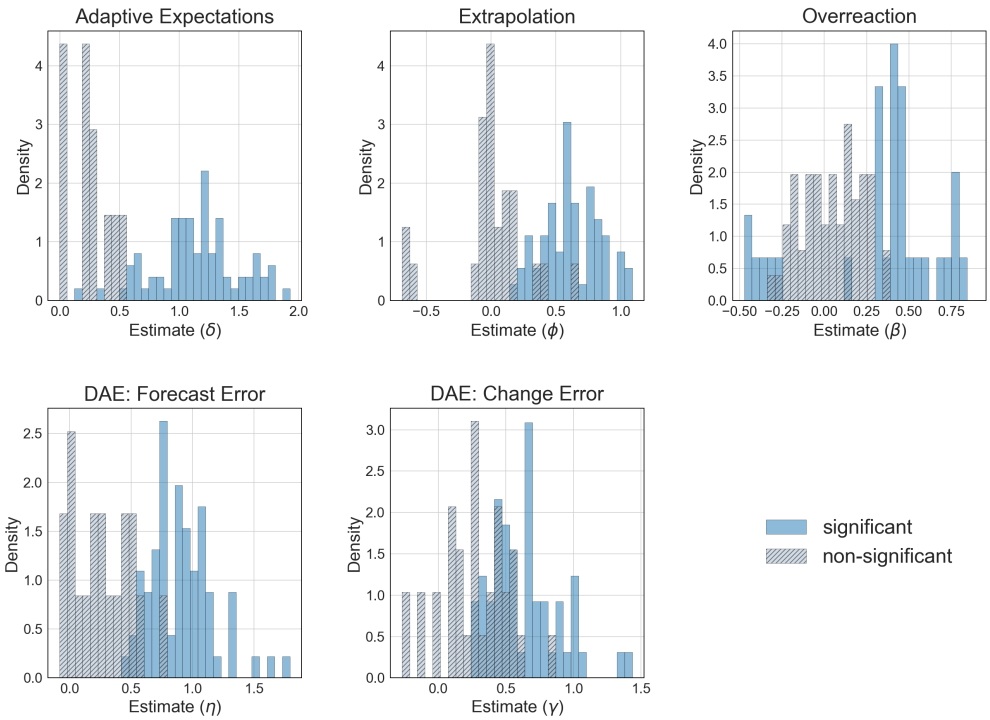


Figure 3.42: Histograms of Parameter Estimates for Different Expectation Models in the Experienced Markets

Notes: Different colors represent different levels of statistical significance. Light gray indicates that the estimate is not significantly different from zero ($p > 0.05$), while dark gray indicates that the estimate is significantly different from zero ($p < 0.05$). All models are estimated using OLS regression, with standard errors that are robust to heteroskedasticity.

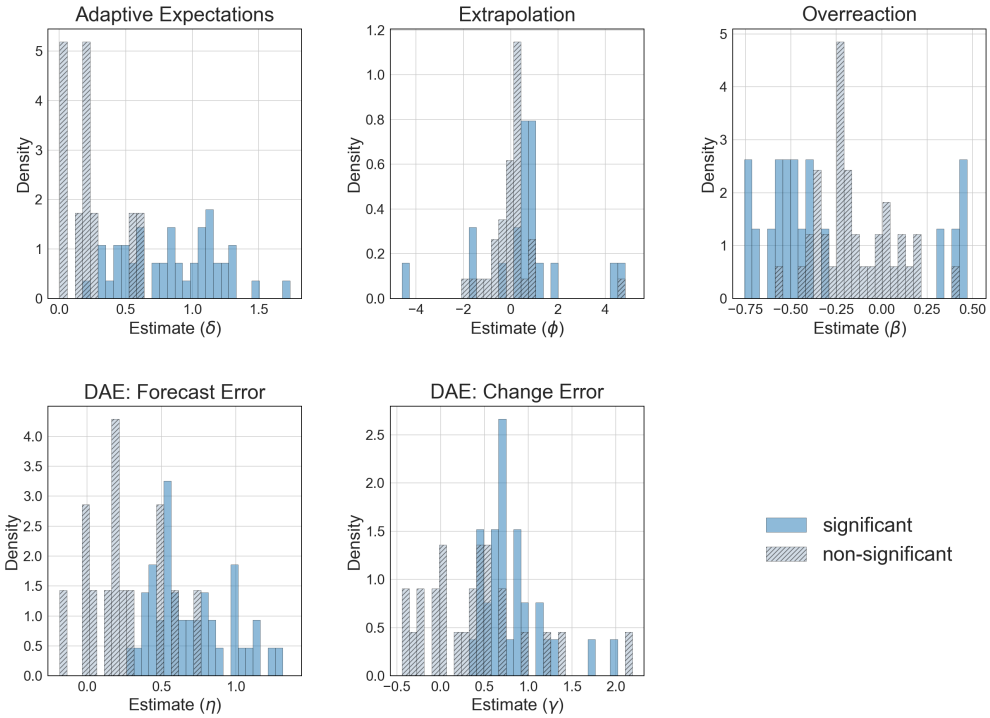


Figure 3.43: Histograms of Parameter Estimates for Different Expectation Models in High-Stakes Markets

Notes: Different colors represent different levels of statistical significance. Light gray indicates that the estimate is not significantly different from zero ($p > 0.05$), while dark gray indicates that the estimate is significantly different from zero ($p < 0.05$). All models are estimated using OLS regression, with standard errors that are robust to heteroskedasticity.

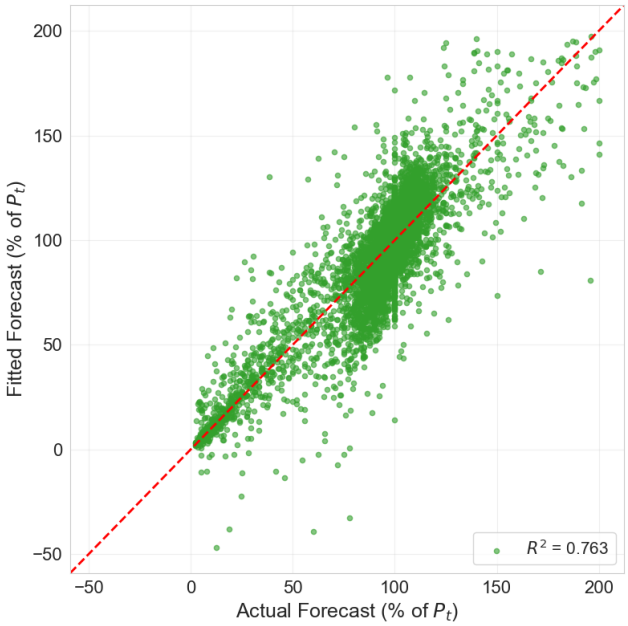


Figure 3.44: Double Adaptive Model: Actual vs. Fitted Forecasts: Forecasts expressed as a percentage of lagged price

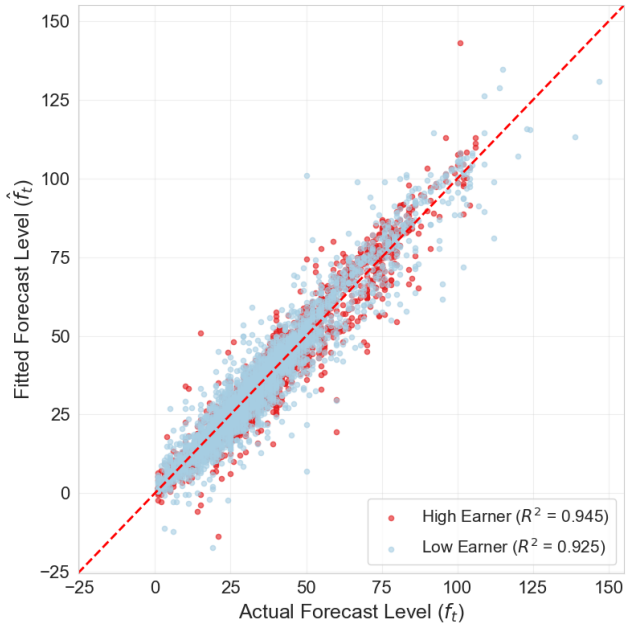


Figure 3.45: Double Adaptive Model: Actual vs. Fitted Forecasts: High vs Low Earners

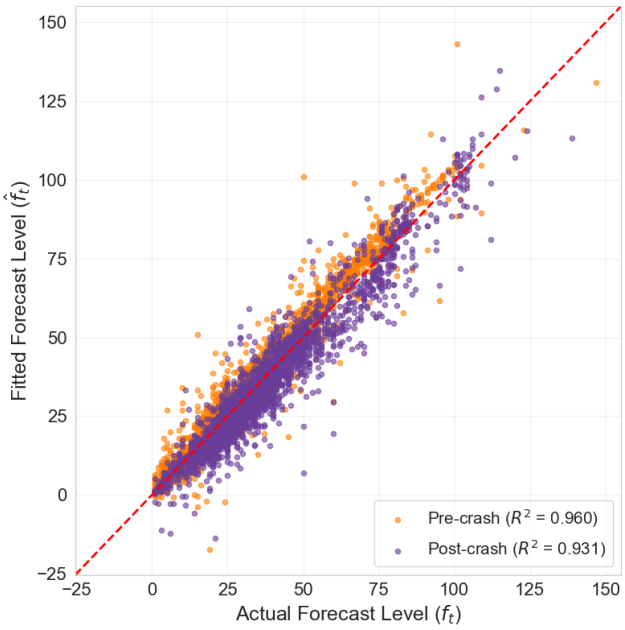


Figure 3.46: Double Adaptive Model: Actual vs. Fitted Forecasts: Pre- vs. Post-crash periods

Table 3.39: OLS and Fixed-Effects Regression Results: Market Bonus vs. Forecast Bonus - Experienced Markets

	Market Bonus		Forecast Bonus	
	(1) OLS	(2) FE	(3) OLS	(4) FE
Adaptive (σ)	0.216	0.226	0.0362***	0.0215
	(0.143)	(0.185)	(0.0101)	(0.0111)
Extrapolation (ϕ)	0.188	0.242**	0.0110	0.0176
	(0.138)	(0.0583)	(0.00826)	(0.0109)
Overreation (β)	-0.0471	-0.0365	-0.0300**	-0.0121**
	(0.123)	(0.119)	(0.0143)	(0.00322)
DAE: Level Err. (η)	-0.526**	-0.570**	-0.0489***	-0.0349
	(0.228)	(0.131)	(0.0175)	(0.0171)
DAE: Trend Err. (γ)	-0.171	-0.151*	-0.0426***	-0.0249
	(0.158)	(0.0607)	(0.0147)	(0.0141)
Constant	1.606***	1.598***	0.0523***	0.0436***
	(0.177)	(0.136)	(0.0164)	(0.00675)
Observations	90	90	90	90
R^2	0.084	0.103	0.204	0.131

Notes: Dependent variables are $\log(1 + \text{bonus})$. This table compares OLS and fixed-effects (FE) regression results for market bonus and forecast bonus. Adaptive expectations, extrapolation, overreaction, and double adaptive parameters are individually estimated from their respective forecasting models and then used as regressors. Columns (1) and (3) report OLS models, robust standard errors in parentheses, and overall R^2, and columns (2) and (4) include session fixed effects, the within R^2, and standard errors clustered by session in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 3.40: OLS and Fixed-Effects Regression Results: Market Bonus vs. Forecast Bonus - High-Stakes Markets

	Market Bonus		Forecast Bonus	
	(1) OLS	(2) FE	(3) OLS	(4) FE
Adaptive (σ)	0.0235	0.0521	-0.00634	0.0137
	(0.117)	(0.159)	(0.0747)	(0.0230)
Extrapolation (ϕ)	0.00454	0.00408	-0.0140***	-0.0102*
	(0.00615)	(0.00718)	(0.00503)	(0.00328)
Overreation (β)	-0.139	-0.164	-0.0656	0.0273
	(0.116)	(0.163)	(0.0878)	(0.0734)
DAE: Level Err. (η)	-0.157	-0.180	-0.0223	0.0254
	(0.135)	(0.0990)	(0.0894)	(0.0469)
DAE: Trend Err. (γ)	-0.0152	-0.00992	0.0533	0.0559
	(0.0558)	(0.0364)	(0.0350)	(0.0239)
Constant	2.926***	2.910***	0.205***	0.177**
	(0.0714)	(0.0969)	(0.0580)	(0.0315)
Observations	57	57	57	57
R^2	0.040	0.041	0.091	0.089

Notes: Dependent variables are $\log(1 + \text{bonus})$. This table compares OLS and fixed-effects (FE) regression results for market bonus and forecast bonus. Adaptive expectations, extrapolation, overreaction, and double adaptive parameters are individually estimated from their respective forecasting models and then used as regressors. Columns (1) and (3) report OLS models, robust standard errors in parentheses, and overall R^2, and columns (2) and (4) include session fixed effects, the within R^2, and standard errors clustered by session in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

3.30 Instructions

Instructions were given by a video which can be seen here:
$https://youtu.be/232inD_HChc$. Below is the transcript of those instructions.

Welcome.

This experiment is about economic decision-making in an experimental stock market. The experiment will consist of a series of 30 trading periods in which you will have the opportunity to buy or sell shares of an asset that can yield payments in the future. Understanding these instructions may help you earn money, and if you 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.

There are two assets in this experiment cash and stock you begin with 100 units of cash and four shares of stock. Stock is traded in a market each period among all the experimental subjects in units of cash. When you buy stock the price you agreed to pay is deducted from your amount of cash. When you sell stock the price you sold at is added to your amount of cash. The reward from holding stock is called dividend. Each period, every unit of stock earns a low or high dividend of either 0.4 cash per unit or 1.00 cash per unit with equal probability. These dividend payments are the same for everyone in all periods the dividend in each period does not depend on whether the previous dividend was low or high. The reward from holding cash is given by a fixed interest rate of 5each period.

Price conversion at the end of the 30 periods of trading each unit of stock is automatically converted to 14 cash. For example if the market price for round 30 is 20 and you have three stocks, you'll receive 42 cash not 60 cash. Then your experimental cash units are converted to US dollars at a rate of 200 cash per 1 US dollar to determine how much you will be paid at the end of the experiment. If you buy shares for more than 14 as you get near Round 30 it is possible those shares will be terminated at a value of 14 if you cannot sell them.

Let's take an example. Suppose one period you have 120 units of cash and 5 units of stock. The dividend is 0.4 per unit of stock. Then your new cash amount is going to be the sum of the initial 120 cash, a 5% increase in cash, and total dividends of 5 times 0.4. Your total cash would therefore be 120 + 6 + 2 resulting in 128 cash. Notice that keeping cash will earn a return 5% per period and using cash to buy units of stock will also yield dividend earnings.

Each trading period will have four main screens that you will see one after the

other. Order submission, forecast prices, round results, and choosing a lottery. This sequence is repeated 30 times. Let's look at each screen in a bit more detail.

On the order submission page in the upper right corner you will see the time left for this page and the trading period. In this example the timer was set on purpose for 5 minutes but in the actual trading task you'll only have 20 seconds. For this page also, the next button will not be displayed in the actual trading task so you cannot click it.

Below that you will see the order submission rectangle where you can buy or sell stocks. The price is displayed on the y-axis with the Bold horizontal line being the previous market price. The quantity of shares you can buy is displayed on the x-axis. The vertical bold line represents the threshold between selling and buying. You can can select the price level at which you want to buy or sell using the mouse or cursor. For example, I can enter a buy order at 15 and submit two sell orders at 16. You can see the orders you entered in the lower right corner. You can cancel orders by clicking on the X next to them. Orders are not carried from one period to the next. If I want to buy at 16 and sell at 15 I will get an error and the sell order will not go through similarly. If I place a sell order at 15 and then I want to submit a buy order at 16 the buy order will not go through. You can only buy stocks if you have enough cash, and you can only sell stocks you own. For example, if I want to sell five stocks I will get an error. Orders are limit orders. A limit order to buy one unit at a price of 15 means you would like to buy one unit at the price of 15 and at any lower price. Similarly a limit order to sell two units at 16 means you would like to sell two units at the price of 16 and at any higher price.

A single market price is determined by matching the number of people who would like to buy at that price with the same number of people who would like to sell. The lower left part of the screen displays your personal stats such as the number of shares you own, your current cash, and the total stock value. For the market stats you can see the current market price, the interest rate, the dividends value, and the buyback value of the stock at the end of the 30 trading periods; which is 14. The upper left corner displays the graph of the stock market from the beginning until the current trading period. We can also see the traded volume for each period.

In the Forecast Prices screen the upper right part of the screen will change. On this page, you will submit your best estimate of what you think the market price will be for this period, two periods in advance, five periods in advance, and 10 periods in advance. As you can see the slider can be dragged from left to right in this example

I am setting the Price Forecast for this current trading period to 14. For two periods ahead I will set it to 14 as well. You can set the price forecast as you want. During the trading task, if the forecast is within 2.5 units from the actual realized price for each of the forecasted periods, then you will receive 2.5 units of cash as reward for each correct forecast. For example, if the actual price for period 1 is going to be 15 then you will be rewarded for that forecast. If the actual price for period 3 is 18 then you will not receive the reward for that forecast. On the round results screen you will see what the market price was for this period, how many shares were traded, as well as the number of shares you traded. You will also see updates on the new current cash and stock value and the dividend you earned. For period one, as you had four shares with a dividend of 0.4 for that period you won $1.60 from holding four stocks and $5 from holding 100 cash.

On the last page you will see two pie charts with different monetary rewards and associated probabilities. You must select one of the pie charts. For example on the left side you have a lottery of winning either $7.25 with a probability of 74% or $9.00 with a probability of 26%. On the right side you have a lottery of winning either $19.00 with 26% probability or $11.50 with 74% probability. After you select one of the lotteries a new set of lotteries will appear. The next button will not appear in the actual trading task so you cannot click on it. There will be four iterations of different lotteries with different monetary rewards and probabilities. You will have 5 seconds for each pair of lotteries, so you must be quick. At the end, one Lottery will be selected at random from those that you chose throughout the experiment, and you will receive the outcome of the lottery divided by 10. It's in your interest to actually choose accordingly. Here are the key points. You will trade One stock for 30 trading periods using cash. You start with 100 units of cash and four stocks. Each period stock provides a dividend of either 0.4 or 1.00 while interest provides 5% reward each trading period. The experiment consists of four pages, order submission, forecasting prices, round results, and choosing a lottery. After the last trading round all of your Shares are converted to 14 cash units. If you buy shares for more than 14, as you get near Round 30 it is possible those shares will be terminated at a value of 14 if you cannot sell them. You will also take a five question quiz. After this, pay attention to the answers you give. You will earn a reward for each question you answer correctly on the first try. If you fail to answer at least one question correctly then you cannot continue.

The experiment will begin shortly.

3.31 Quiz

This quiz was administered at the end of the instructions. Participants answered the questions and submitted their responses. Feedback was given along with an explanation of any errors. Participants were asked to submit answers until they got all the questions correct. The quiz bonus was assessed based on their first submission.

Correct answers appear in **bold**

1. During order entry, you want to input a BUY order at a price of 8.52 and a SELL order at a price of 7.89. What will happen to your orders for that period?
    - They'll be entered as usual
    - **The SELL order will not go through**
2. During order entry, the highest price at which you submit a BUY order is 16.78, and the lowest price at which others SELL is 16.22. The market price is 16.56 for that period. You will:
    - BUY one unit at 16.78
    - **BUY one unit at 16.56**
    - SELL one unit at 16.56
    - SELL one unit at 17.22
    - Not Trade
3. You have 200 units of CASH at the start of a trading period and no STOCK. You do not trade that period. How much CASH do you have at the beginning of the next period?
    - 200
    - **210**
    - 205
    - 190
4. Your account has 5 STOCK and 100 CASH at the start of a trading period, and you do not BUY or SELL during that period. The dividend for that round is 1.00. How much CASH do you have at the start of the next round?

- 105
- **110**
- 100
- 114

5. After the final trading period, you have one remaining unit of STOCK. The market price in the final period is 29. How many units of experiment CASH do you receive in exchange for your STOCK?

   - 1
   - 29
   - **14**
   - 30

3.32 Surveys

Pre-Experiment Survey 1

Thank you for deciding to participate in the experiment. The answers you will provide herein will solely be used in academic research. The following is a pre-experimental consent form.

If you give your consent, you can participate in the experiment. If you do not give your consent, then please return the study on Prolific.

- Given the exclusion criteria, please confirm that you qualify to participate in the experiment:
  - Yes
  - No
- Print Prolific ID (as substitute for Name) *[text entry]*
- Please initial one option below to indicate whether or not you consent to us contacting you in the future:
  - I consent to having Caltech researchers contact me about other studies at Caltech that are funded by non-profit organizations, government agencies, universities, and other non-commercial entities.
  - I consent to having Caltech researchers contact me about other studies at Caltech that are funded by for-profit companies and other commercial entities.
  - I consent to having Caltech researchers contact me about any other studies at Caltech, regardless of who is funding the research.
  - I do not wish to be contacted about participating in future research studies.
- Print Participant's Initials (as substitute for signature) *[text entry]*
- Please indicate your consent
  - I consent
  - I do not consent

Pre-Experiment Survey 2

The following is a pre-experimental questionnaire.

The answers you will provide herein will solely be used in academic research.

After filling it in, you will perform the task in the actual experiment and then fill in the post-experimental questionnaire.

It is very important to be as honest as possible and to treat each question rigorously, as your participation will directly impact the subsequent research study.

The answers to this questionnaire are strictly confidential. Most of the questions are mandatory - you can distinguish them through the red asterisk (*) sign.

- Please enter your Prolific ID *[text entry]*
- Your age (in years) *[text entry]*
- Your gender
    1. Male
    2. Female
    3. Non-binary
    4. I prefer not to say
- Are you Hispanic, Latino/a, or of Spanish origin?
    - Yes
    - No
- What is your race?
    1. White caucasian
    2. Black or African American
    3. American Indian or Alaska Native
    4. Asian or Asian Indian
    5. Native Hawaiian or Other Pacific Islander
    6. What race do you consider yourself? Write here: *[text entry]*
- Studies background

  - Social sciences: economics, psychology
  - IT and software
  - STEM
  - Humanities and arts
  - I don't have higher education studies
  - Other... *[text entry]*

- Are you a native English speaker (i.e., is English your mother tongue?)
  - Yes
  - No

- On a scale from 0 to 5, how good is your knowledge of English? (If YES to previous question, please mark five here) *[0 through 5]*

- Are you currently under medication treatment for depression, anxiety, or other psychiatric or neurological conditions?
  - Yes
  - No

- How would you rate your risk-taking behavior in general, on a scale from 1-7?

  The scale goes from 1 = "Very risk-averse" to 7 = "Very risk-taking"

- How would you rate your financial risk-taking behavior, on a scale from 1-7?

  The scale goes from 1 = "Very risk-averse" to 7 = "Very risk-taking"

- You have been given $100. You can invest any part of this money in a risky asset. If the investment is successful, the risky asset returns 2.5 times the amount invested with a probability of one-half and nothing with a probability of one-half. You can also choose not to invest at all, in which case you will keep the money you have. How much would you like to invest? (write amount in dollars)

Post-experimental questionnaire

Thank you for deciding to fill in this questionnaire. It takes approximately 10 minutes to fill in. The answers you will provide herein will solely be used in academic research.

It is very important to be as honest as possible and to treat each question rigorously, as your participation will directly impact the subsequent research study.

The answers to this questionnaire are strictly confidential. Most of the questions are mandatory - you can distinguish them through the red asterisk (*) sign.

- Prolific ID *[text entry]*
- Please enter the 3-digit ID. If you do not have it, enter the Prolific ID. *[text entry]*
- Have you participated in similar trading experiments before?
  - Yes, multiple times
  - Yes, just once
  - No
- On a scale from 1-10, how well did you understand the task and were able to trade as required?
- Did you understand exactly what you had to do from the beginning?
  - Yes, entirely
  - Yes, but partially, I understood more as I continued trading
  - No, not at all
- Did you follow any strategy when trading? Please describe, be as detailed as possible *[text entry]*
- Do you have any suggestions to improve the experiment? i.e., the software, the trading instructions, time to enter orders, time to enter forecast, time to do the risk elicitation tasks, etc. *[text entry]*
- Do you own/trade stocks?
  - YES
  - NO
- Imagine you performed the tasks with another 99 persons. In terms of performance, in what place do you think you placed? How would you rate your outcome from 1-100? *[text entry]*

- How would you rate your risk-taking behavior in general, on a scale from 1-7?

  The scale goes from 1 = “Very risk-averse" to 7 = “Very risk taking"

- How would you rate your financial risk-taking behavior, on a scale from 1-7?

  The scale goes from 1 = “Very risk-averse" to 7 = “Very risk taking"

- You have been given $100. You can invest any part of this money in a risky asset. If the investment is successful, the risky asset returns 2.5 times the amount invested with a probability of one-half and nothing with a probability of one-half. You can also choose not to invest at all, in which case you will keep the money you have. How much would you like to invest? (write amount in dollars)

  [text entry]

- Choose what you prefer:
  - 100% chance of winning $10
  - 50% chance of winning $0, 50% chance of winning $10
- Choose what you prefer:
  - 100% chance of winning $10
  - 50% chance of winning $0, 50% chance of winning $12
- Choose what you prefer:
  - 100% chance of winning $10
  - 50% chance of winning $0, 50% chance of winning $14
- Choose what you prefer:
  - 100% chance of winning $10
  - 50% chance of winning $0, 50% chance of winning $16
- Choose what you prefer:
  - 100% chance of winning $10
  - 50% chance of winning $0, 50% chance of winning $18
- Choose what you prefer:

  - 100% chance of winning $10
  - 50% chance of winning $0, 50% chance of winning $20
- Choose what you prefer:
  - 100% chance of winning $10
  - 50% chance of winning $0, 50% chance of winning $22
- Choose what you prefer:
  - 100% chance of winning $10
  - 50% chance of winning $0, 50% chance of winning $24
- Choose what you prefer:
  - 100% chance of winning $10
  - 50% chance of winning $0, 50% chance of winning $26
- Choose what you prefer:
  - 100% chance of winning $10
  - 50% chance of winning $0, 50% chance of winning $28
- Choose what you prefer:
  - 100% chance of winning $10
  - 50% chance of winning $0, 50% chance of winning $30
- Choose what you prefer:
  - 100% chance of winning $10
  - 50% chance of winning $0, 50% chance of winning $32

References

Abarbanell, Jeffery S. and Victor L. Bernard (1992). "Tests of Analysts' Overreaction/Underreaction to Earnings Information as an Explanation for Anomalous Stock Price Behavior". In: *Journal of Finance* 47, pp. 1181–1207. DOI: 10.1111/j.1540-6261.1992.tb04010.x.

Ackert, Lucy F. et al. (2006). "Margin, Short Selling, and Lotteries in Experimental Asset Markets". In: *Southern Economic Journal* 73.2, pp. 419–436. DOI: 10.1002/j.2325-8012.2006.tb00779.x.

Afrouzi, Hassan et al. (2023). "Overreaction in expectations: Evidence and theory". In: *The Quarterly Journal of Economics* 138.3, pp. 1713–1764.

Alempaki, Despoina, Chris Starmer, and Fabio Tufano (2019). "On the Priming of Risk Preferences: The Role of Fear and General Affect". In: *Journal of Economic Psychology* 75, p. 102137. DOI: https://doi.org/10.1016/j.joep.2018.12.011..

Ambuehl, Sandro (2017). *An offer you can't refuse? Incentives change what we believe*. Tech. rep. No. 6296. CESifo. URL: https://www.cesifo.org/en/publikationen/2017/working-paper/offers-you-cant-refuse-incentives-change-what-we-believe.

Amromin, Gene and Steven A. Sharpe (2014). "From the Horse's Mouth: Economic Conditions and Investor Expectations of Risk and Return". In: *Management Science* 60, pp. 845–866. DOI: 10.1287/mnsc.2013.1806.

Andrade, Eduardo B., Terrance Odean, and Shengle Lin (May 2015). "Bubbling with Excitement: An Experiment *". In: *Review of Finance* 20.2, pp. 447–466. ISSN: 1572-3097. DOI: 10.1093/rof/rfv016.

Anufriev, Mikhail and Cars Hommes (2012). "Evolutionary Selection of Individual Expectations and Aggregate Outcomes in Asset Pricing Experiments". In: *American Economic Journal: Microeconomics* 4.4, pp. 35–64. DOI: 10.1257/mic.4.4.35.

Ball, Sheryl B and Charles A Holt (Feb. 1, 1998). "Classroom Games: Speculation and Bubbles in an Asset Market". In: *Journal of Economic Perspectives* 12.1, pp. 207–218. ISSN: 0895-3309. DOI: 10.1257/jep.12.1.207. URL: https://pubs.aeaweb.org/doi/10.1257/jep.12.1.207 (visited on 10/21/2023).

Bao, Te, Cars Hommes, and Jiaoying Pei (Dec. 2021). "Expectation formation in finance and macroeconomics: A review of new experimental evidence". en. In: *Journal of Behavioral and Experimental Finance* 32, p. 100591. ISSN: 2214-6350. DOI: https://doi.org/10.1016/j.jbef.2021.100591.. (Visited on 11/27/2022).

Bao, Te, Cars Hommes, Joep Sonnemans, et al. (2012). "Individual Expectations, Limited Rationality and Aggregate Outcomes". In: *Journal of Economic Dynamics and Control* 36.8, pp. 1101–1120. DOI: 10.1016/j.jedc.2012.03.014.

Bao, Te and Jichuan Zong (Oct. 2019). "The impact of interest rate policy on individual expectations and asset bubbles in experimental markets". en. In: *Journal of Economic Dynamics and Control* 107, p. 103735. ISSN: 0165-1889. DOI: 10.1016/j.jedc.2019.103735. URL: https://www.sciencedirect.com/science/article/pii/S0165188919301344 (visited on 11/27/2022).

Barberis, Nicholas (2013). "Psychology and the Financial Crisis of 2007–2008". In: pp. 15–28.

Barberis, Nicholas, Robin Greenwood, et al. (2015). "X-CAPM: An extrapolative capital asset pricing model". In: *Journal of Financial Economics* 115.1, pp. 1–24. ISSN: 0304-405X. DOI: https://doi.org/10.1016/j.jfineco.2014.08.007. URL: https://www.sciencedirect.com/science/article/pii/S0304405X14001822.

– (2018). "Extrapolation and Bubbles". In: *Journal of Financial Economics* 129.2, pp. 203–227. DOI: https://doi.org/10.1016/j.jfineco.2018.04.007.

Barrero, Jose Maria (2022). "The Micro and Macro of Managerial Beliefs". In: *Journal of Financial Economics* 143, pp. 640–647. DOI: 10.1016/j.jfineco.2021.06.007.

Beedie, Christopher J., Peter C. Terry, and Andrew M. Lane (2005). "Distinctions between emotion and mood". In: *Cognition & Emotion* 19.6, pp. 847–878. DOI: 10.1080/02699930541000057.

Berkman, Henk et al. (2017). "Trading While Sleepy? Circadian Mismatch and Excess Volatility in a Global Experimental Asset Market". In: *Review of Finance* 21.3, pp. 1111–1156. URL: https://ideas.repec.org/p/apl/wpaper/17-06.html%3E.

Blanchard, Olivier J and Mark W Watson (1982). "Bubbles, rational expectations and financial markets". In: ed. by Paul Wachtel. Lexington Books, pp. 295–315.

Bollerslev, Tim, George Tauchen, and Hao Zhou (2009). "Expected Stock Returns and Variance Risk Premia". In: *The Review of Financial Studies* 22.11, pp. 4463–4492.

Bordalo, Pedro, Nicola Gennaioli, Rafael La Porta, et al. (2019). "Diagnostic Expectations and Stock Returns". In: *Journal of Finance* 74, pp. 2839–2874. DOI: 10.1111/jofi.12833.

Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, et al. (2020). "Overreaction in Macroeconomic Expectations". In: *American Economic Review* 110, pp. 2748–2782. DOI: 10.1257/aer.20181219.

Bossaerts, Peter (July 2021). "How Neurobiology Elucidates the Role of Emotions in Financial Decision-Making". English. In: *Frontiers in Psychology* 12. Publisher: Frontiers. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.697375. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.697375/full (visited on 06/13/2024).

Bossaerts, Peter et al. (June 2024). "Emotional Engagement and Trading Performance". In: *Management Science* 70.6. Publisher: INFORMS, pp. 3381–3397. ISSN: 0025-1909. DOI: 10.1287/mnsc.2023.4883. URL: https://pubsonline.informs.org/doi/full/10.1287/mnsc.2023.4883 (visited on 06/16/2024).

Bostian, A. J., J. Goeree, and C. A. Holt (Sept. 2005). "Price Bubbles in Asset Market Experiments with a Flat Fundamental Value". Draft for the Experimental Finance Conference, Federal Reserve Bank of Atlanta.

Bouchaud, Jean-Philippe et al. (2019). "Sticky Expectations and the Profitability Anomaly". In: *Journal of Finance* 74, pp. 639–674. DOI: 10.1111/jofi.12734.

Boucsein, Wolfram (2012). *Electrodermal Activity*. 2nd ed. New York: Springer. DOI: 10.1007/978-1-4614-1126-0.

Brandt, Michael W. and Kevin Q. Wang (Oct. 1, 2003). "Time-varying risk aversion and unexpected inflation". In: *Journal of Monetary Economics* 50.7, pp. 1457–1498. ISSN: 0304-3932. DOI: 10.1016/j.jmoneco.2003.08.001. URL: https://www.sciencedirect.com/science/article/pii/S0304393203000886 (visited on 03/06/2025).

Breaban, Adriana and Charles N. Noussair (2018). "Emotional State and Market Behavior". In: *Review of Finance* 22.1, pp. 279–309. DOI: 10.1093/rof/rfx022. URL: https://doi.org/10.1093/rof/rfx022.

Brown, Ryan et al. (2019). "Impact of Violent Crime on Risk Aversion: Evidence from the Mexican Drug War". In: *Review of Economics and Statistics* 101.5, pp. 892–904. DOI: 10.1162/rest_a_00788.

Brunnermeier, Markus K and Stefan Nagel (2004). "Hedge funds and the technology bubble". In: *The Journal of Finance* 59.5, pp. 2013–2040.

Cagan, Phillip (1956). "The monetary dynamics of hyperinflation". In: *Studies in the Quantity Theory if Money*.

Caginalp, Gunduz, David Porter, and Vernon Smith (June 1, 2001). "Financial Bubbles: Excess Cash, Momentum, and Incomplete Information." In: *Journal of Psychology & Financial Markets* 2.2. Publisher: Taylor & Francis Ltd, pp. 80–99. ISSN: 1520-8834. DOI: 10.1207/S1532776OJPFM0202_03. URL: https://research.ebsco.com/linkprocessor/plink?id=36e5de29-6d8e-3e55-97a5-1fd0b1fd1b62 (visited on 02/18/2025).

Camerer, Colin (1989). "Bubbles and Fads in Asset Prices". en. In: *Journal of Economic Surveys* 3.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6419.1989.tb00056.x, pp. 3–41. ISSN: 1467-6419. DOI: 10.1111/j.1467-6419.1989.tb00056.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6419.1989.tb00056.x (visited on 11/27/2022).

Camerer, Colin F. and Robin M. Hogarth (1999). "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework". In: *Journal of Risk and Uncertainty* 19.1–3, pp. 7–42. DOI: 10.1023/A:1007850605129.

Camerer, Colin F. and Keith Weigelt (1993). "Convergence in Experimental Double Auctions for Stochastically Lived Assets". In: ed. by Daniel Friedman and John Rust. Redwood City, CA: Addison-Wesley, pp. 355–396.

Cameron, Lisa and Manisha Shah (2015). "Risk-Taking Behavior in the Wake of Natural Disasters". In: *Journal of Human Resources* 50.2, pp. 484–515. DOI: 10.3368/jhr.50.2.484.

Campbell, John Y. and John H. Cochrane (1999). "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior". In: *Journal of Political Economy* 107.2, pp. 205–251. DOI: 10.1086/250059.

Campbell, John Y. and Robert J. Shiller (1991). "Yield Spreads and Interest Rate Movements: A Bird's Eye View". In: *Review of Economic Studies* 58.3, pp. 495–514.

Chapman, Jonathan et al. (Oct. 2024). *Dynamically Optimized Sequential Experimentation (DOSE) for Estimating Economic Preference Parameters*. DOI: 10.3386/w33013. URL: https://www.nber.org/papers/w33013 (visited on 02/09/2025).

Chen, Daniel L., Martin Schonger, and Chris Wickens (Mar. 1, 2016). "oTree—An open-source platform for laboratory, online, and field experiments". In: *Journal of Behavioral and Experimental Finance* 9, pp. 88–97. ISSN: 2214-6350. DOI: 10.1016/j.jbef.2015.12.001. URL: https://www.sciencedirect.com/science/article/pii/S2214635016000101 (visited on 04/28/2022).

Coates, John M. and Joe Herbert (2008). "Endogenous steroids and financial risk taking on a London trading floor". In: *Proceedings of the National Academy of Sciences* 105.16, pp. 6167–6172. DOI: 10.1073/pnas.0704025105.

Cochrane, John H. (2011). "Presidential Address: Discount Rates". In: *The Journal of Finance* 66.4, pp. 1047–1108. DOI: 10.1111/j.1540-6261.2011.01671.x.

Cohn, Alain et al. (Feb. 2015). "Evidence for Countercyclical Risk Aversion: An Experiment with Financial Professionals". In: *American Economic Review* 105.2, pp. 860–885. ISSN: 0002-8282. DOI: 10.1257/aer.20131314. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20131314 (visited on 03/06/2025).

Cordes, Henning, Sven Nolte, and Judith C. Schneider (Sept. 1, 2023). "Dynamics of stock market developments, financial behavior, and emotions". In: *Journal of Banking & Finance* 154, p. 106711. ISSN: 0378-4266. DOI: 10.1016/j.jbankfin.2022.106711. URL: https://www.sciencedirect.com/science/article/pii/S0378426622002916 (visited on 02/19/2025).

Corgnet, Brice et al. (2022). "Forecasting Skills in Experimental Markets: Illusion or Reality?" In: *Management Science* 68.7, pp. 5216–5232. DOI: https://doi.org/10.1287/mnsc.2021.4160.

Cui, Xuegang, Nick Feltovich, and Kun Zhang (May 2022). "Incentive schemes, framing, and market behaviour: Evidence from an asset-market experiment". In: *Journal of Economic Behavior & Organization* 197, pp. 301–324. DOI: 10.1016/j.jebo.2022.03.007. URL: https://doi.org/10.1016/j.jebo.2022.03.007.

Dawson, Michael E., Anne M. Schell, and Diane L. Filion (2007). "The electrodermal system". In: ed. by John T. Cacioppo, Louis G. Tassinary, and Gary G. Berntson. 3rd ed. Cambridge, UK: Cambridge University Press, pp. 159–181.

De Bondt, Werner F. M. and Richard H. Thaler (1990). "Do Security Analysts Overreact?" In: *American Economic Review* 80, pp. 52–57. URL: https://www.jstor.org/stable/2006542.

Dufwenberg, Martin, Tobias Lindqvist, and Evan Moore (Dec. 2005). "Bubbles and Experience: An Experiment". In: *American Economic Review* 95.5, pp. 1731–1737. ISSN: 0002-8282. DOI: 10.1257/000282805775014362. URL: https://www.aeaweb.org/articles?id=10.1257/000282805775014362 (visited on 04/19/2022).

Enke, Benjamin et al. (2023). "Cognitive Biases: Mistakes or Missing Stakes?" In: *The Review of Economics and Statistics* 105.4, pp. 818–832. DOI: 10.1162/rest_a_01093. URL: https://doi.org/10.1162/rest_a_01093.

Fama, Eugene F. (1970). "Efficient Capital Markets: A Review of Theory and Empirical Work". In: *The Journal of Finance* 25.2, pp. 383–417. DOI: 10.2307/2325486.

– (1984). "The Information in the Term Structure". In: *Journal of Financial Economics* 13.4, pp. 509–528.

Fama, Eugene F. and Robert R. Bliss (1987). "The Information in Long-Maturity Forward Rates". In: *American Economic Review* 77.4, pp. 680–692.

Fenton-O'Creevy, Mark et al. (Nov. 1, 2012). "Emotion regulation and trader expertise: Heart rate variability on the trading floor." In: *Journal of Neuroscience, Psychology, and Economics* 5.4. Publisher: Educational Publishing Foundation, pp. 227–237. ISSN: 1937-321X. DOI: 10.1037/a0030364. URL: https://research.ebsco.com/linkprocessor/plink?id=b624240f-74ae-3ea9-a5fc-0c9131a1c11e (visited on 02/19/2025).

Filippin, Antonio et al. (2023). "Risk Aversion and Information Aggregation in Binary-Asset Markets". In: *Quantitative Economics* 14.2, pp. 753–798. DOI: 10.3982/QE1981.

Frino, Alex, Joel Grant, and David Johnstone (2008). "The House Money Effect and Local Traders on the Sydney Futures Exchange". In: *Pacific-Basin Finance Journal* 16.1–2, pp. 8–25. DOI: 10.1016/j.pacfin.2007.06.003. URL: https://doi.org/10.1016/j.pacfin.2007.06.003.

Gabaix, Xavier (2012). "Variable Rare Disasters: An Exactly Solved Framework for Ten Puzzles in Macro-Finance". In: *Quarterly Journal of Economics* 127.2, pp. 645–700. DOI: 10.1093/qje/qjs001.

Gennaioli, Nicola, Yueran Ma, and Andrei Shleifer (2016). "Expectations and Investment". In: *NBER Macroeconomics Annual* 30, pp. 379–431. DOI: 10.1086/685965.

Greenwood, Robin and Andrei Shleifer (2014a). "Expectations of Returns and Expected Returns". In: *Review of Financial Studies* 27, pp. 714–746. DOI: 10.1093/rfs/hht082.

– (2014b). "Expectations of Returns and Expected Returns". In: *Review of Financial Studies* 27.3, pp. 714–746. DOI: 10.1093/rfs/hht082.

Greenwood, Robin, Andrei Shleifer, and Yang You (2019). "Bubbles for Fama". In: *Journal of Financial Economics* 131.1, pp. 20–43. DOI: 10.1016/j.jfineco.2018.07.010.

Grosshans, Daniel and Stefan Zeisberger (2025). "Mind the Gap: Investor Beliefs vs. Trading Actions". Trading Actions, December 10, 2025.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales (June 2018). "Time varying risk aversion". In: *Journal of Financial Economics* 128.3, pp. 403–421. ISSN: 0304405X. DOI: 10.1016/j.jfineco.2018.02.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304405X18300461 (visited on 02/19/2025).

Hanaoka, Chie, Hitoshi Shigeoka, and Yasutora Watanabe (2018). "Do Risk Preferences Change? Evidence from the Great East Japan Earthquake". In: *American Economic Journal: Applied Economics* 10.2, pp. 298–330. DOI: 10.1257/app.20170048.

Haruvy, Ernan, Yaron Lahav, and Charles N Noussair (Dec. 1, 2007). "Traders' Expectations in Asset Markets: Experimental Evidence". In: *American Economic Review* 97.5, pp. 1901–1920. ISSN: 0002-8282. DOI: 10.1257/aer.97.5.1901. URL: https://pubs.aeaweb.org/doi/10.1257/aer.97.5.1901 (visited on 02/19/2025).

Haruvy, Ernan and Charles N. Noussair (2006a). "The Effect of Short Selling on Bubbles and Crashes in Experimental Spot Asset Markets". In: *The Journal of Finance* 61.3, pp. 1119–1157. DOI: 10.1111/j.1540-6261.2006.00868.x.

– (2006b). "The Effect of Short Selling on Bubbles and Crashes in Experimental Spot Asset Markets". In: *The Journal of Finance* 61.3, pp. 1119–1157. ISSN: 1540-6261. DOI: 10.1111/j.1540-6261.2006.00868.x. URL: http://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2006.00868.x (visited on 07/23/2021).

Hauser, Florian and Jürgen Huber (2012). "Short-Selling Constraints as Cause for Price Distortions: An Experimental Study". In: *Journal of International Money and Finance* 31.5, pp. 1279–1298. DOI: 10.1016/j.jimonfin.2012.02.001.

Hirshleifer, David, Jun Li, and Jianfeng Yu (2015). "Asset pricing in production economies with extrapolative expectations". In: *Journal of Monetary Economics* 76, pp. 87–106.

Holt, Charles A., Megan Porzio, and Michelle Y. Song (2017). "Price Bubbles, Gender, and Expectations in Experimental Asset Markets". In: *European Economic Review* 100, pp. 72–94. DOI: 10.1016/j.euroecorev.2017.05.005. URL: https://doi.org/10.1016/j.euroecorev.2017.05.005.

Hommes, Cars et al. (2005). "Coordination of expectations in asset pricing experiments". In: *Review of Financial Studies* 18.3, pp. 955–980.

– (2008). "Expectations and bubbles in asset pricing experiments". In: *Journal of Economic Behavior & Organization* 67, pp. 116–133.

Hommes, Cars H. (2011). "The Heterogeneous Expectations Hypothesis: Some Evidence from the Lab". In: *Journal of Economic Dynamics and Control* 35.1, pp. 1–24. DOI: 10.1016/j.jedc.2010.10.003.

Hong, Harrison and Jeremy C. Stein (2007). "Disagreement and the Stock Market". In: *Journal of Economic Perspectives* 21.2, pp. 109–128. DOI: 10.1257/jep.21.2.109.

Hsu, Yuan-Lin and Edward H. Chow (2013). "The House Money Effect on Investment Risk Taking: Evidence from Taiwan". In: *Pacific-Basin Finance Journal* 21.1, pp. 1102–1115. URL: https://doi.org/10.1016/j.pacfin.2012.08.005..

Huber, Christoph et al. (2024). "Do Experimental Asset Market Results Replicate? High-Powered Preregistered Replications of 17 Key Results". In: *Journal of Finance*. Forthcoming. DOI: 10.2139/ssrn.5048949. URL: https://ssrn.com/abstract=5048949.

Huber, Jürgen and Michael Kirchler (Mar. 1, 2012). "The impact of instructions and procedure on reducing confusion and bubbles in experimental asset markets". In: *Experimental Economics* 15.1, pp. 89–105. ISSN: 1573-6938. DOI: 10.1007/s10683-011-9290-8. URL: https://doi.org/10.1007/s10683-011-9290-8 (visited on 05/01/2025).

Hussam, Reshmaan N., David Porter, and Vernon L. Smith (2008). "Thar She Blows: Can Bubbles Be Rekindled with Experienced Subjects?" In: *American Economic Review* 98.3, pp. 924–937. DOI: 10.1257/aer.98.3.924.

Jakiela, Pamela and Owen Ozier (2019). "The Impact of Violence on Individual Risk Preferences: Evidence from a Natural Experiment". In: *Review of Economics and Statistics* 101.3, pp. 547–559. DOI: 10.1162/rest_a_00763.

Kandasamy, Narayanan et al. (Sept. 19, 2016). "Interoceptive Ability Predicts Survival on a London Trading Floor". In: *Scientific Reports* 6.1. Publisher: Nature Publishing Group, p. 32986. ISSN: 2045-2322. DOI: 10.1038/srep32986. URL: https://www.nature.com/articles/srep32986 (visited on 02/19/2025).

Kim, Don H. and Athanasios Orphanides (June 2007). "The Bond Market Term Premium: What Is It, and How Can We Measure It?" In: *BIS Quarterly Review*.

King, Ronald R., Vernon Smith, et al. (Jan. 1, 1993). "The Robustness of Bubbles and Crashes in Experimental Stock Markets". In: *Nonlinear Dynamics and Evolutionary Economics, edited by R.* Dall and P. Chen, Oxford, U.K.: Oxford University Press*, pp. 183–200.

King, Ronald R., Vernon L. Smith, et al. (1993). "The Robustness of Bubbles and Crashes in Experimental Stock Markets". In: *Nonlinear Dynamics and Evolutionary Economics*. Ed. by Richard H. Day and Ping Chen. New York: Oxford University Press, pp. 183–200.

Kirchler, Michael, Jürgen Huber, and Thomas Stöckl (Apr. 1, 2012). "Thar She Bursts: Reducing Confusion Reduces Bubbles". In: *American Economic Review* 102.2, pp. 865–883. ISSN: 0002-8282. DOI: 10.1257/aer.102.2.865. URL: https://pubs.aeaweb.org/doi/10.1257/aer.102.2.865 (visited on 10/11/2021).

Kohlhas, Alexandre N. and Ansgar Walther (Sept. 2021). "Asymmetric Attention". In: *American Economic Review* 111.9, pp. 2879–2925. DOI: 10.1257/aer.20191432. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20191432.

König-Kersting, Christian and Stefan T. Trautmann (2018). "Countercyclical Risk Aversion: Beyond Financial Professionals". In: *Journal of Behavioral and Experimental Finance* 18, pp. 94–101. DOI: 10.1016/j.jbef.2018.03.001.

Kopányi-Peuker, Anita and Matthias Weber (Aug. 2021). "Experience Does Not Eliminate Bubbles: Experimental Evidence". en. In: *The Review of Financial Studies* 34.9. Ed. by Lauren Cohen, pp. 4450–4485. ISSN: 0893-9454, 1465-7368. DOI: 10.1093/rfs/hhaa121. URL: https://academic.oup.com/rfs/article/34/9/4450/5926134 (visited on 12/23/2025).

Lee, Charles M. C., Andrei Shleifer, and Richard H. Thaler (1991). "Investor Sentiment and the Closed-End Fund Puzzle". In: *The Journal of Finance* 46.1, pp. 75–109. DOI: 10.1111/j.1540-6261.1991.tb03746.x.

Lo, Andrew W. and Dmitry V. Repin (2002). "The psychophysiology of real-time financial risk processing". In: *Journal of Cognitive Neuroscience* 14. Place: US Publisher: MIT Press, pp. 323–339. ISSN: 1530-8898. DOI: 10.1162/089892902317361877.

Lovell, Michael C (1986). "Tests of the rational expectations hypothesis". In: *The American Economic Review* 76.1, pp. 110–124.

Ma, Yueran et al. (2020). *A Quantitative Analysis of Distortions in Managerial Forecasts*. Tech. rep. 26830. National Bureau of Economic Research. DOI: 10.3386/w26830.

Makowski, Dominique et al. (Feb. 2021). "NeuroKit2: A Python toolbox for neurophysiological signal processing". In: *Behavior Research Methods* 53.4, pp. 1689–1696. DOI: 10.3758/s13428-020-01516-y. URL: https://doi.org/10.3758%2Fs13428-020-01516-y.

Mantovani, Marco and Antonio Filippin (2026). "When Do Prediction Markets Return Average Beliefs? Experimental Evidence". In: *Games and Economic Behavior* 156, pp. 135–148.

McCarthy, C. et al. (2016). "Validation of the Empatica E4 wristband". In: IEEE, pp. 1–4.

McKelvey, Richard D and Thomas R Palfrey (1992). "An experimental study of the centipede game". In: *Econometrica* 60.4, pp. 803–836.

Milstein, N. and I. Gordon (2020). "Validating measures of electrodermal activity and heart rate variability derived from the Empatica E4 utilized in research settings that involve interactive dyadic states". In: *Frontiers in Behavioral Neuroscience* 14, p. 148.

Mohr, Peter N. C., Guido Biele, and Hauke R. Heekeren (2010). "Neural Processing of Risk". In: *Journal of Neuroscience* 30.19, pp. 6613–6619. DOI: 10.1523/JNEUROSCI.0003-10.2010.

Morriss, Jayne, Martin Gell, and Carien M. van Reekum (2019). "The Uncertain Brain: A Co-ordinate Based Meta-analysis of the Neural Signatures Supporting Uncertainty During Different Contexts". In: *Neuroscience and Biobehavioral Reviews* 96, pp. 241–249. ISSN: 0149-7634. DOI: 10.1016/j.neubiorev.2018.12.013. URL: https://centaur.reading.ac.uk/81223/.

Nerlove, Marc (May 1958). "Adaptive Expectations and Cobweb Phenomena". In: *The Quarterly Journal of Economics* 72.2, pp. 227–240. ISSN: 0033-5533. DOI: 10.2307/1880597. eprint: https://academic.oup.com/qje/article-pdf/72/2/227/5314908/72-2-227.pdf. URL: https://doi.org/10.2307/1880597.

Noussair, Charles, Stephane Robin, and Bernard Ruffieux (June 1, 2001). "Price Bubbles in Laboratory Asset Markets with Constant Fundamental Values". In: *Experimental Economics* 4.1, pp. 87–105. ISSN: 1573-6938. DOI: 10.1023/A:1011445522861. URL: https://doi.org/10.1023/A:1011445522861 (visited on 04/19/2022).

Palan, Stefan (2013). "A Review of Bubbles and Crashes in Experimental Asset Markets". In: *Journal of Economic Surveys* 27.3. Publisher: Wiley Blackwell, pp. 570–588. URL: https://econpapers.repec.org/article/blajecsur/v_3a27_3ay_3a2013_3ai_3a3_3ap_3a570-588.htm (visited on 11/27/2022).

Porter, David P. and Vernon L. Smith (1995). "Futures Contracting and Dividend Uncertainty in Experimental Asset Markets". In: *The Journal of Business* 68.4. Publisher: University of Chicago Press, pp. 509–541. ISSN: 0021-9398. URL: https://www.jstor.org/stable/2353144 (visited on 04/19/2022).

– (2010). "Stock Market Bubbles in the Laboratory". In: *Journal of Behavioral Finance* 4.1, pp. 7–20. DOI: 10.1207/S15427579JPFM0401_03. URL: https://doi.org/10.1207/S15427579JPFM0401_03.

Russell, James A. (2003). "Core affect and the psychological construction of emotion". In: *Psychological Review* 110.1, pp. 145–172. DOI: 10.1037/0033-295X.110.1.145.

Schmalensee, Richard (1976). "An Experimental Study of Expectation Formation". In: *Econometrica* 44.1, pp. 17–41. URL: https://www.jstor.org/stable/1911378.

Schuurmans, Angela A. T. et al. (Sept. 23, 2020). "Validity of the Empatica E4 Wristband to Measure Heart Rate Variability (HRV) Parameters: a Comparison to Electrocardiography (ECG)". In: *Journal of Medical Systems* 44.11, p. 190. ISSN: 1573-689X. DOI: 10.1007/s10916-020-01648-w. URL: https://doi.org/10.1007/s10916-020-01648-w (visited on 03/06/2025).

Shum, Matthew and Yi Xin (2022). "Time-Varying Risk Aversion? Evidence from Near-Miss Accidents". In: *Review of Economics and Statistics* 104.6, pp. 1317–1328. DOI: 10.1162/rest_a_01009.

Smith, Alec et al. (July 2014). "Irrational exuberance and neural crash warning signals during endogenous experimental market bubbles". In: *Proceedings of the National Academy of Sciences* 111.29. Publisher: Proceedings of the National Academy of Sciences, pp. 10503–10508. DOI: 10.1073/pnas.1318416111. URL: https://www.pnas.org/doi/10.1073/pnas.1318416111 (visited on 11/27/2022).

Smith, Vernon L. (Jan. 1, 2010). "Theory and experiment: What are the questions?" In: *Journal of Economic Behavior & Organization*. On the Methodology of Experimental Economics 73.1, pp. 3–15. ISSN: 0167-2681. DOI: 10.1016/j.jebo.2009.02.008. URL: https://www.sciencedirect.com/science/article/pii/S016726810900033X (visited on 02/19/2025).

Smith, Vernon L., Gerry L. Suchanek, and Arlington W. Williams (1988). "Bubbles, Crashes, and Endogenous Expectations in Experimental Spot Asset Markets". In: *Econometrica* 56.5. Publisher: [Wiley, Econometric Society], pp. 1119–1151. ISSN: 0012-9682. DOI: 10.2307/1911361. URL: https://www.jstor.org/stable/1911361 (visited on 11/27/2022).

Smith, Vernon L. and James M. Walker (1993). "Monetary Rewards and Decision Cost in Experimental Economics". In: *Economic Inquiry* 31.2, pp. 245–261. DOI: 10.1111/j.1465-7295.1993.tb00881.x.

Stöckl, Thomas, Jürgen Huber, and Michael Kirchler (Sept. 1, 2010). "Bubble measures in experimental asset markets". In: *Experimental Economics* 13.3, pp. 284–298. ISSN: 1573-6938. DOI: 10.1007/s10683-010-9241-9. URL: https://doi.org/10.1007/s10683-010-9241-9 (visited on 04/13/2022).

Sui, Pengfei and Baolian Wang (2023). *Stakes and Investor Behaviors*. Tech. rep. 4219861. SSRN. DOI: 10.2139/ssrn.4219861. URL: https://ssrn.com/abstract=4219861.

Sunder, Shyam (1995). "Experimental Asset Markets: A Survey". In: *The Handbook of Experimental Economics*. Ed. by John H. Kagel and Alvin E. Roth. Princeton University Press, pp. 445–500. ISBN: 978-0-691-04290-9. DOI: 10.2307/j.ctvzsmff5.10. URL: https://www.jstor.org/stable/j.ctvzsmff5.10 (visited on 02/19/2025).

Tashjian, Sarah M. et al. (Feb. 1, 2022). "Physiological Responses to a Haunted-House Threat Experience: Distinct Tonic and Phasic Effects". In: *Psychological Science* 33.2. Publisher: SAGE Publications Inc, pp. 236–248. ISSN: 0956-7976. DOI: 10.1177/09567976211032231. URL: https://doi.org/10.1177/09567976211032231 (visited on 03/06/2025).

Thaler, Richard H. and Eric J. Johnson (1990). "Gambling with the House Money and Trying to Break Even: The Effects of Prior Outcomes on Risky Choice". In: *Management Science* 36.6, pp. 643–660. DOI: 10.1287/mnsc.36.6.643. URL: https://doi.org/10.1287/mnsc.36.6.643.

Tirole, Jean (1985). "Asset bubbles and overlapping generations". In: *Econometrica: Journal of the Econometric Society*, pp. 1499–1528.

Toma, Filip-Mihai, Cosmin-Octavian Cepoi, et al. (Jan. 2023). "Gazing through the bubble: an experimental investigation into financial risk-taking using eye-tracking". In: *Financial Innovation* 9.1, p. 28. ISSN: 2199-4730. DOI: 10.1186/s40854-022-00444-4. URL: https://doi.org/10.1186/s40854-022-00444-4 (visited on 06/16/2024).

Toma, Filip-Mihai and Makoto Miyakoshi (June 2021). "Left Frontal EEG Power Responds to Stock Price Changes in a Simulated Asset Bubble Market". en. In: *Brain Sciences* 11.6. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute, p. 670. ISSN: 2076-3425. DOI: 10.3390/brainsci11060670. URL: https://www.mdpi.com/2076-3425/11/6/670 (visited on 11/28/2022).

Van Boening, Mark V., Arlington W. Williams, and Shawn LaMaster (Jan. 1993). "Price bubbles and crashes in experimental call markets". In: *Economics Letters* 41.2, pp. 179–185. ISSN: 01651765. DOI: 10.1016/0165-1765(93)90194-H. URL: https://linkinghub.elsevier.com/retrieve/pii/016517659390194H (visited on 08/19/2021).

Weitzel, Utz et al. (2020). "Bubbles and financial professionals". In: *The Review of Financial Studies* 33.6, pp. 2659–2696.

Williams, Arlington W. (2008). "Price Bubbles in Large Financial Asset Markets". In: ed. by Charles R. Plott and Vernon L. Smith. Vol. 1. Amsterdam: Elsevier, pp. 242–246. DOI: 10.1016/S1574-0722(07)00029-7.

Williams, Arlington W. and James M. Walker (1993). "Computerized Laboratory Exercises for Microeconomics Education: Three Applications Motivated by Experimental Economics". In: *Journal of Economic Education* 24.4, pp. 291–315. DOI: 10.1080/00220485.1993.10844835.

Yu, Rongjun (Feb. 2015). "Choking under pressure: the neuropsychological mechanisms of incentive-induced performance decrements". In: *Frontiers in Behavioral Neuroscience* 9, p. 19. DOI: 10.3389/fnbeh.2015.00019. URL: https://doi.org/10.3389/fnbeh.2015.00019.

Zarnowitz, Victor (1985). "Rational expectations and macroeconomic forecasts". In: *Journal of Business & Economic Statistics* 3.4, pp. 293–311.

Chapter 4

LLM AGENTS DO NOT REPLICATE HUMAN MARKET TRADERS: EVIDENCE FROM EXPERIMENTAL FINANCE

Thomas Henning, Siddhartha Ojha, Ross Spoon, Jiatong Han, and Colin F. Camerer

ABSTRACT

In this study, we compare Large Language Models (LLMs) with human traders in a classic experimental-finance paradigm where prices are determined endogenously. Using a well-established asset-trading design, we run homogeneous (mono-agent) markets with single-model LLM agents and heterogeneous "Battle Royale" markets with multiple LLM models. Our findings reveal that LLMs generally exhibit a "textbook-rational" approach, pricing the asset near its fundamental value and showing only a muted tendency toward bubble formation, whereas humans deviate substantially and consistently generate bubbles. Additional treatments, including dividend shocks and repeated-exposure ('experienced') runs, show that these differences persist across various experimental settings. Further analyses of LLM-generated strategy text indicate lower variance, reduced bias, and stronger reliance on fundamentals relative to humans' more heuristic-driven trading. These results highlight the risk of using LLM-agents to model human-driven market phenomena, as key behavioral features such as large, emergent bubbles are not reproduced.

4.1 Introduction

Financial markets are complex, dynamic coordination systems shaped by individual decisions made under uncertainty. While traditional financial theory assumes rational behavior, decades of behavioral finance research has revealed persistent biases, such as herding and overconfidence, that can give rise to mispricings and asset bubbles. These deviations underscore the challenges of achieving optimal outcomes in uncertain and evolving market environments.

The rise of agentic systems, particularly those driven by Large Language Models (LLMs), introduces a novel and previously unexplored dimension to this problem. As LLMs increasingly participate in domains traditionally navigated by humans, including financial systems and markets, understanding their behavioral tendencies has become critical. Questions such as; Do LLM-agents exhibit similar behavioral biases, behave more rationally, or fall victim to entirely different biases not seen in humans?

This paper presents the first comprehensive study comparing "out-of-the-box" LLM and human behavior within a single experimental finance paradigm. We adopt a classical trading environment where participants trade a risky asset with a controlled fundamental value and examine the behavior of LLM-agents in both homogeneous (mono-agent) markets, populated entirely by instances of the same model, and heterogeneous "Battle Royale" markets, where agents based on different base LLMs compete. These experimental markets are entirely endogenous, with prices determined solely by the buy-and-sell orders submitted by participants, and without any external price anchor or market maker. This enables us to dive deeper into how the underlying agent behavior affects the market, as we can control all market parameters.

Previous research offers conflicting perspectives on AI-human behavioral alignment. Some works suggest a significant difference between human and LLM interactions (Anlló et al., 2024), while others indicate that LLM behavior could supplement or even substitute for human studies (Leng, 2024; Horton, 2023; Jia et al., 2024). We address this tension in the literature through rigorous experimental comparison in experimental finance, a domain that has long served as a testbed for understanding financial decision-making under risk and uncertainty.

Contributions Our findings reveal clear behavioral differences between LLM and human-driven markets. LLM markets demonstrate substantially more “textbook-

rational" behavior than their human counterparts. Human markets consistently generate significant price bubbles and deviations from the asset's fundamental value, while LLM-agents generally trade much closer to fundamental value and exhibit markedly reduced trading variance. We also find that LLMs appear to incorporate less bias into their future price forecasts and employ strategies more focused on fundamental value, rather than strategies commonly used by humans. These results challenge the notion that "out-of-the-box" LLMs can faithfully replicate human market dynamics, particularly in terms of the emergence of market-level phenomena such as bubbles and crashes. As agentic systems increasingly participate in and influence real financial markets, understanding these behavioral divergences becomes essential for effective regulation, risk management, and market design.

4.2 Background

Numerous works have studied LLMs in strategic environments such as social dilemmas and iterated games (Brookins and DeBacker, 2024; Horton, 2023; Akata et al., 2023). In finance, most existing studies focus on empirical applications—LLMs trading in real markets or analyzing financial data (e.g., (Xiao et al., 2024; Y. Li et al., 2024; Y. Yu, H. Li, et al., 2024; Yun, 2024; K. M. Liu and Lo, 2025; Ding et al., 2024; Abdelsamie and H. Wang, 2024; Q. Chen, 2025; L. Zhou et al., 2025)). However, to our knowledge, no prior work has analyzed LLM behavior in experimental markets where prices are endogenously generated by agent interaction.

In finance, LLMs have regularly been applied to real-world scenarios. Papers such as; X.-Y. Zhang et al. (2023), Bond (2023), and Kirtac and Germano (2024) explored LLMs' ability to process financial sentiment and predict market movements. Fatouros et al. (2024) introduced MarketSenseAI, which leverages GPT-4 for stock selection, illustrating the utility of LLMs in navigating complex financial data. While these studies demonstrate the promise of LLMs in financial settings, there is still a need to understand how these agents perform in experimental financial settings where strategic interactions drive outcomes.

The study of market dynamics in controlled experimental environments has revealed critical insights into behavioral biases and inefficiencies. For instance, V. L. Smith, Suchanek, and Williams (1988) demonstrated the formation of bubbles and crashes in experimental markets, even under conditions of perfect information. Subsequent work by A. A. Bostian and Charles A. Holt (2009) highlighted how individual behaviors, such as overconfidence and herding, contribute to these phenomena. In-

corporating AI agents into such paradigms presents a unique opportunity to evaluate whether these systems replicate, mitigate, or exacerbate market inefficiencies.

Several studies have analyzed whether LLMs replicate human-like behaviors in canonical game-theory tasks. Brookins and DeBacker (2024) examine GPT-3.5 in the Prisoner's Dilemma, showing tendencies toward cooperation that exceed typical human baselines. Horton (2023) study LLMs in public goods games, highlighting context-dependent altruism. Akata et al. (2023) explore GPT-4 in iterated games, where it adapts between cooperative and competitive strategies, suggesting emergent long-term planning.

Empirical finance applications. A large and rapidly growing literature investigates LLMs as tools for real-world financial prediction and decision-making. Studies include LLM agents for trading and market making Xiao et al., 2024; Y. Li et al., 2024; X. Li et al., 2025; Y. Yu, Yao, et al., 2024, financial data analysis Y. Yu, H. Li, et al., 2024; Yun, 2024; K. M. Liu and Lo, 2025; Ding et al., 2024, multimodal integration W. Zhang et al., 2024; X. Yu et al., 2023; Kim, 2025, and forecasting tasks Q. Chen, 2025; L. Zhou et al., 2025; C. Zhang et al., 2024; Kang et al., 2024; Siddique et al., 2025; Saqur, 2024; Bhat and Jain, 2024; Kuruvilla and Mythily, 2025; Oprea and Bâra, 2024; Shi and Hollifield, 2024; Cheng and Chin, 2024; H. Zhao et al., 2024. This literature demonstrates the usefulness of LLMs in financial contexts but focuses on exogenous markets and benchmarks.

Experimental and behavioral finance. Our work differs from the above by situating LLMs in endogenous experimental markets, where prices emerge solely from agent interaction. This connects more directly to the tradition of laboratory asset markets V. L. Smith, Suchanek, and Williams, 1988; A. A. Bostian and Charles A. Holt, 2009, which have long been used to isolate human biases such as herding, overconfidence, and bubble formation. To our knowledge, ours is the first study to apply this paradigm systematically to LLM-agents.

Evaluation Frameworks for LLMs Conventional benchmarks for LLMs focus on static environments, emphasizing either knowledge-oriented tasks (e.g., Hendrycks et al., 2021) or reasoning-oriented challenges (Cobbe et al., 2021). While these benchmarks provide valuable insights, they often overlook strategic reasoning, a key facet of human intelligence. Recent efforts, such as Chatbot Arena (Zheng et al., 2023), introduce pairwise comparison frameworks to assess LLM interactions dynamically.

4.3 Experiment

We focus on a classic experimental finance paradigm with a constant fundamental value, the mathematical price of an asset; see Equation 4.B in the Appendix. We adopt the experimental design from A. Smith et al. (2014) and Charles A. Holt, Porzio, and Song (2017), which involves an endogenous experimental asset market where subjects trade a risky asset (STOCK) in an open-call market for 30 periods operated using the oTree experiment platform (D. L. Chen, Schonger, and Wickens, 2016). Participants are given 4 shares of STOCK and 100 units of CASH. Participants have two options: either invest CASH in STOCK, which in each trading period pays out dividends of either 0.4 or 1 units of CASH with equal probability, or hold CASH, which generates interest at 5% each trading period. The price is determined solely by the trading activity of the experimental market participants.

After the experiment, the risky asset is redeemed for a fixed value of 14 units of CASH, representing the fundamental value of the STOCK (See Appendix 4.B for full formulation).

Because of these parameter decisions (i.e., dividend schedule, CASH interest rate, and redemption price), the fundamental value remains constant throughout the entirety of the experiment, enabling us to objectively classify when the market is in a “bubble" state A. Bostian, Goeree, and Charles A Holt, 2005. This differs starkly from applied financial studies, where the fundamental value of a real-world asset is unknown. Further discussion of the experimental details can be found in Appendices 4.A.

This task was assigned to both LLM agents and human traders in separate markets. In adapting the task to LLM-agents, our primary goal was to best emulate the experiment presented to human subjects. The description of the task setting mirrors the original instruction video, shown to human participants (See Appendix 4.6). In every round of the experiment, each agent simultaneously submits orders to the market as well as expectations regarding current and future market prices.

A common challenge in adapting multi-round games to LLM-agents is the persistence of memory between rounds, which is critical to learning and adapting strategies. Building off previous works (Fish, Gonczarowski, and Shorrer, 2024), we allow agents to communicate with future versions of themselves by allowing them to read and write their own “Insights" and “Thoughts" at each round. This structure incorporates Chain-of-Thought (CoT) reasoning (Wei, X. Wang, et al., 2022), which has been shown to help elicit the reasoning ability of LLMs, providing

an interpretable window into how agents perceive and interact with the market. Further prompting details are presented in Appendix 4.7.

To frame our analysis, we categorized potential outcomes into three groups: **rational (R)**, where models do not generate bubbles and trade close to the fundamental value; **human (H)**, where models generate bubbles in a manner similar to humans; and **erratic (E)**, where models perform inconsistently, failing to adhere to any consistent strategy, whether rational or human-like.

4.4 Single-Model LLM Market Performance

In this section, we evaluate how single-model LLM markets behave along two critical dimensions: rational adherence to fundamental asset value and similarity to human market dynamics. Table 4.1 and Figure 4.1 summarize our key findings.

Experimental Setup

We conducted three markets using agents based on six different API-based commercial LLMs: Anthropic's Claude-3.5-Sonnet Anthropic, 2024, OpenAI's GPT-3.5 and GPT-4o OpenAI, 2025, xAI's Grok-2 xAI, 2024, Mistral-Large AI's Mistral-Large AI, 2024, and Google's Gemini-1.5 Pro DeepMind, 2024. Each market consisted of 20 identical instances of the same LLM agent. The agents were given the same prompt and were unaware of the identities of their counterparts. Each instance was instructed to maximize its profit.

For comparison, we also ran nineteen human-only markets, with identical experimental conditions (though the humans also conducted a short risk-elicitation task in between each round). Participants were a mix of Caltech students and staff, as well as online participants recruited through `prolific.com`. The human participants were given instructions that mirror the LLM prompts; full instructions are shown in Appendix section 4.6. The average plot from these experiments is shown alongside the LLM-only Battle Royale market results in Figure 4.2.

Single-Model Market Analysis

As shown in Table 4.1, Claude-3.5 Sonnet and GPT-4o yield the lowest MSEs and nearly zero correlation with human price paths, keeping prices close to fundamentals and mapping to the Rational (R) hypothesis. Both also show low portfolio variance, suggesting consistent strategies across instances.

By contrast, Grok-2 and GPT-3.5 show the largest deviations from fundamentals and

the highest correlations with human markets. Grok-2 exhibits bubble–crash cycles, while GPT-3.5 generates a monotonic run-up without correction (Figure 4.1), a pattern occasionally seen in humans.

Mistral-Large also produces bubbles but with smaller peaks and sharper crashes, dipping below fundamentals late in the session. While more rational in absolute terms (MSE $\approx$ 5.7 vs. 429.8 for humans), its dynamics align with the Human (H) hypothesis.

Finally, Gemini-1.5 Pro departs from both rational and human-like behavior, trading well below fundamentals before correcting upward. This so-called "reverse bubble" (negative PCC) suggests erratic dynamics rather than coherent strategy, mapping most closely to the Erratic (E) hypothesis.

Table 4.1: Market performance metrics

Model	MSE (Fundamental)	PCC (Human Avg)	Closest Hypothesis	PV Variance
Claude-3.5-Sonnet	0.536	0.001	R	26.15
GPT-4o	0.789	0.050	R	22.76
Grok-2	17.325	0.558	H	39.43
Mistral-Large	5.694	-0.112	H	49.96
GPT-3.5	26.367	0.490	H	30.44
Gemini-1.5 Pro	3.103	-0.401	E	45.58

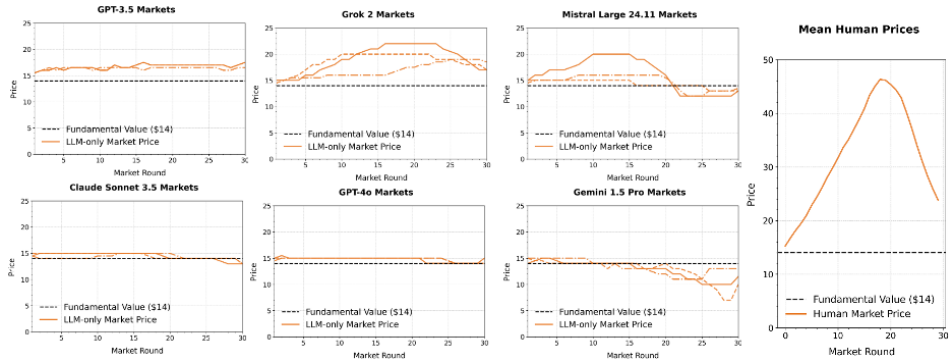
MSE vs. Fundamental Value measures rationality by assessing how closely the market prices align with the fundamental value. **PCC vs. Human Market Average** quantifies the similarity of LLM-driven markets to typical human market price trajectories using the Pearson correlation coefficient (PCC). **Portfolio Value (PV) Variance** reflects strategy diversity among agents, indicating whether they adopt varied approaches or behave uniformly. The **Closest Hypothesis** column maps each model's behavior to the closest hypothesized market behavior pattern, based on a mix of quantitative and qualitative comparisons. R if MSE around fundamental was below 1, and H if we observed a bubble formation and crash, E if the market behavior was nonsensical (reverse bubble).

4.5 Battle Royale: Comparing LLM Market Performance

To evaluate how various LLM agents fare when competing in the same market, we conducted *Battle Royale* market experiment with a heterogeneous mixture of the models. By introducing diverse LLMs into a single environment, we aimed to explore:

- Which LLM is most effective at maximizing profits in a competitive, multi-agent setting? Which Agent is the best trader?

Figure 4.1: Mono-Agent Price Charts compared to Humans



None of the LLM-agents generate market bubbles similar in magnitude to humans.

- How a mixed-LLM market behaves, and whether it will exhibit emergent phenomena when models with distinct trading strategies interact?
- Whether heterogeneous LLM interpretation and strategy differences are enough to generate financial bubbles?

Experimental Setup

We conducted markets involving the same six LLM models as in the Single-Model section. Each market included a total of 24 agents (four agents per model). The agents operated under identical conditions and were tasked with maximizing their profits over 30 trading periods. The risky asset in these markets maintained a constant fundamental value of 14 units, consistent with prior experiments. Figure 4.2 shows the three 'Battle Royale' market results.

Model	**Avg Portfolio Value ± 1 STD**
Mistral-Large	689.56 ± 4.11
Gemini-1.5 Pro	688.49 ± 4.48
Claude-3.5 Sonnet	680.36 ± 5.08
GPT-3.5	674.42 ± 3.48
GPT-4o	671.36 ± 17.30
Grok-2	668.51 ± 27.01

Table 4.2: Battle Royale Agent Performance

Market-Level Summary Statistics

MSE (Fundamental): **5.86** PCC (Human Avg): **0.27**

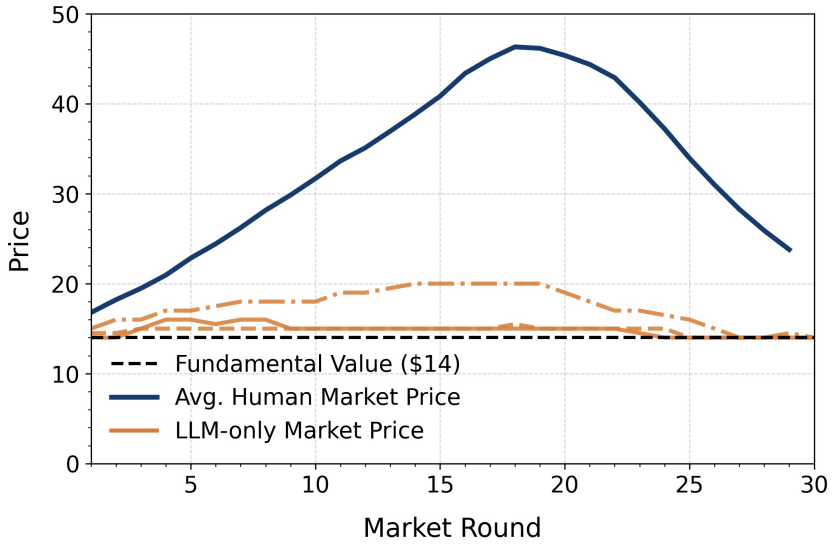


Figure 4.2: Battle Royale Market Performance with Human Comparison

Battle Royale Market Analysis

Across three heterogeneous markets, two converged on prices near fundamentals, while one produced a modest bubble resembling human markets but with a lower peak. Trading volume spiked at the end of this bubble run, echoing the late-round rush seen in human experiments (Figure 4.2). Overall, these dynamics align most closely with the Rational (R) hypothesis, although they demonstrate that competitive interaction can occasionally generate bubble-like patterns.

At the agent level (Table 4.2), portfolio values cluster tightly, with no model consistently dominating. The spread between best and worst traders remains small, suggesting convergence on broadly similar strategies and limited inequality of outcomes in mixed markets. This shows that none of these models are able to consistently best the others in competitive trading.

4.6 Trading Strategy Analysis

After each round, LLM agents were prompted to provide details of the strategy they used to drive their trading behavior. This was done both to allow us to analyze their behavior more deeply and to enable the LLMs to develop long-running strategies, utilizing Chain-of-Thought-like reasoning. In contrast, human subjects were asked what approach they employed only at the end of the experiment.

We first examined the word frequency of both humans and LLMs in the mono-agent runs. Of the top six words for each group, five are the same, though in a different order ("sell", "buy", "stock", "price", and "market"; see Table 4.3 in Appendix for

details). The top ten most popular words for the human subjects include "low" and "high", which do not appear among the top ten words used by the LLMs. Conversely, the words "continue" and "buyback" are among the most common for the LLMs (see also word clouds in Figure 4.3 in the Appendix). This pattern suggests that human traders follow a "buy low, sell high" strategy in an attempt to outperform the market. In contrast, LLM agents adopt a longer-term approach that emphasizes the asset's intrinsic value.

Text Mining and Strategy Comparison to Humans

To assess whether LLMs engaged in coordination or converged on shared narratives, we ran a text–mining analysis of the "insight" and "plan" files. We found no explicit coordination language (0 hits on terms such as "coordinate" or "work together"), and only a handful of generic references to "anticipating other agents." By contrast, we observed several instances of near–orthogonal language between agent pairs (cosine similarity ≤ 0.1; e.g., "buy aggressively" vs. "sell off"), suggesting divergence rather than cooperation. Overall, mixed–model markets appear to consist of independent traders reacting to shared price signals rather than colluding.

We then directly compared LLM and human strategies. A classifier trained on the insights/plans text achieved >99% accuracy in identifying which model produced which text. Extending this to a seven–way classifier including humans (using their end of experiment reflections), we found that ~33% of Gemini-1,5 Pro and GPT-4 reflections were misclassified as human, while all other models were below 10%. Binary classifiers confirmed this pattern: GPT-4 text was misclassified as human nearly 50% of the time (Mistral-Large 36%; GPT-3.5 28%), while other models were below 3%. Despite superficial stylistic overlap, cluster analysis of strategy words showed almost no clusters containing both humans and LLMs (in one run, only 8/497 humans clustered with any model), suggesting that the underlying reasoning remains distinct.

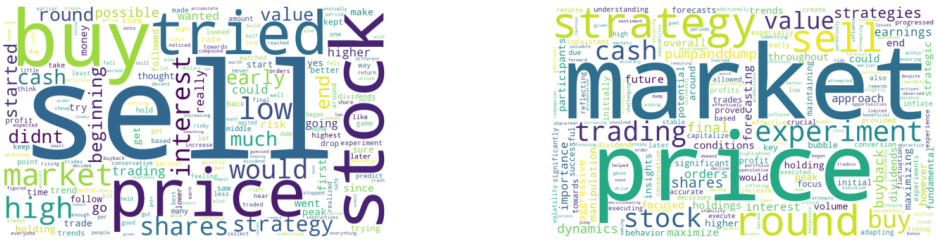
To probe this distinction further, we ran a Latent Dirichlet Allocation (LDA) Blei, Ng, and Jordan, 2003 on all strategies combined. Using the Python `gensim` library Řehůřek and Sojka, 2010, we assigned each strategy to one of two topics. The first topic, identified by the words *orders, sell, buy, shares, stock*, denoted strategies aligned with a "buy low; sell high" practice. The second topic, identified by the words *stock, adjust, buyback, closely, current*, reflected a more fundamental value focused approach. The LDA assigned 89.3% of human strategies to the first topic

but only 36.6% of the LLM strategies; conversely, 63.4% of LLM strategies (10.7% of humans) fell into the second category. A t-test confirmed this difference is highly significant ($p < .001$). Figure 4.4 shows this distribution.

Table 4.3: Most popular words used by human participants and by the LLM Agents

	Human			LLM-Agent		
Rank	Word	Count	Freq	Word	Count	Freq
1	**sell**	230	5.60%	**price**	88707	6.04%
2	**buy**	202	4.92%	**market**	63445	4.32%
3	**stock**	177	4.31%	**buy**	38996	2.66%
4	**price**	141	3.43%	**sell**	38141	2.60%
5	tried	129	3.14%	orders	31335	2.14%
6	**market**	79	1.92%	round	28385	1.93%
7	low	75	1.83%	**stock**	28130	1.92%
8	high	70	1.70%	**cash**	26469	1.80%
9	would	57	1.39%	continue	23318	1.59%
10	interest	53	1.29%	adjust	19250	1.31%
11	**shares**	51	1.24%	value	18669	1.27%
12	strategy	48	1.17%	**shares**	18476	1.26%
13	**cash**	44	1.07%	buyback	17555	1.20%
14	early	41	1.00%	monitor	16582	1.13%
15	much	41	1.00%	current	14595	0.99%

Common words are in bold text. The Count columns are the number of times a word is used by each group (humans and LLMs). Frequency is the count divided by the number of individual words used by that group.



(a) Human Strategies

(b) LLM Strategies

Figure 4.3: Word cloud representation of the most commonly used words in the strategies for human and LLM agents.

Linguistic Audit of Model Strategies

We also wanted to identify why specific models generate larger, more human-like bubbles, while others adhere relatively closely to fundamentals. To do so, we ran a finer-grained linguistic audit of the "insight" and "plan" outputs. Each statement was tagged as *speculative*, *fundamental*, or *generic* based on keyword dictionaries.

Table 4.4: Distribution (%) of speculative, fundamental, and generic text in agent strategy statements. The behavior column reflects observed market outcomes.

Model	Speculative	Fundamental	Generic	Behavior
Grok-2	68.4	16.0	15.6	Bubble-prone
GPT-3.5	28.4	18.0	53.6	Bubble-prone (No Crash)
GPT-4o	17.9	59.6	22.5	Bubble-dampening
Claude-3.5 Sonnet	14.5	82.8	2.7	Bubble-dampening
Gemini-1.5 Pro	20.5	58.7	20.8	Bubble-dampening
Mistral-Large	35.8	52.0	12.2	Mixed

Speculative keywords: "trend", "momentum", "price rising", "price falling", "buy more", "bubble", "sell off", "rally".
Fundamental keywords: "undervalued", "intrinsic", "dividend", "fair value", "discount", "fundamental", "14", "buy back".

Grok-2 and GPT-3.5 seem to create bubbles because speculative narratives dominate their reasoning, whereas GPT-4o, Claude-3.5, and Gemini-1.5 Pro seem to dampen exuberance by emphasizing intrinsic value anchors. Mistral-Large oscillates between the two, producing smaller, transient bubbles with a trajectory of value → momentum → uncertainty. These linguistic signatures provide a possible mechanistic explanation for the divergent market outcomes observed across models.

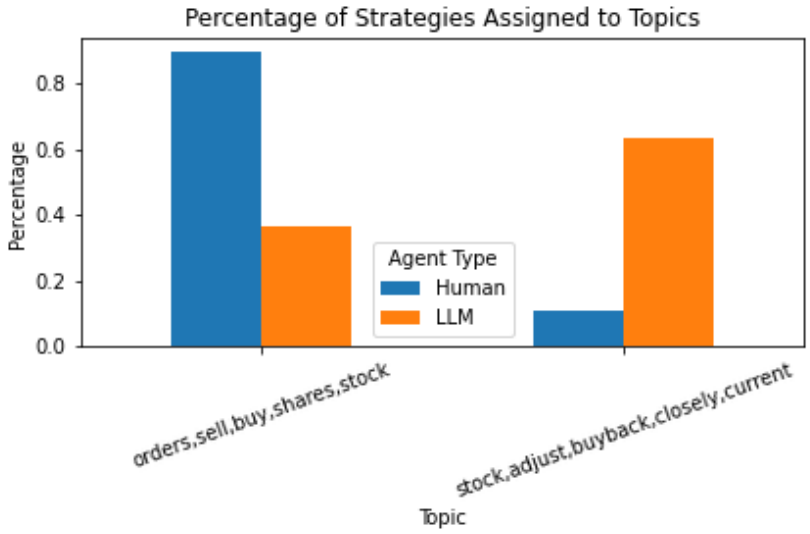
Individual Token counts

Method. For every agent, we concatenate the round-by-round `Plan` and `Insight` fields, convert them to lowercase, strip punctuation, and tokenize.

Four interpretable features are then computed:

1. **Fundamental token rate** – occurrences of *fundamental, intrinsic, value, mean-revert, FV, 14, discounted cash-flow* (14 keywords).
2. **Speculative token rate** – *pump, breakout, leverage, surge, moon, momentum, Fomo* (18 keywords).
3. **Trade-verb density** – *buy, sell, long, short, liquidate*.
4. **Risk / Profit ratio** – risk-related tokens divided by profit-related tokens (*risk, volatility, drawdown* vs. *profit, gain, return*).

Figure 4.4: Percentage of human and LLM strategies assigned to one of two topics.



The topics were generated and assigned by Latent Dirichlet Allocation. The topic *price, round, orders, sell, buy* indicates strategies that attempt to beat the market (buy low; sell high). While the *price market continue, stock, adjust* topic suggests a more measured and long-term strategy.

Counts are normalized per 1,000 tokens

Table 4.5: Language–based strategy fingerprints (plans+insights).

Model	**Fund. /1000**	**Spec. /1000**	**Trade verbs/1000**	**Risk /Profit**
Claude-3.5 Sonnet	**61.9**	4.3	21.7	1.98
GPT-4o	26.0	4.1	**31.5**	**0.90**
Gemini-1.5-Pro	40.1	7.0	34.8	1.01
Grok-2	29.9	5.0	17.8	2.30
Mistral-Large	25.7	6.5	34.2	1.12
GPT-3.5	19.6	**10.1**	**3.8**	**7.55**

Findings.

- **Value-anchored rationalists.** Claude-3.5 (62 fundamental tokens) and GPT-4o (lowest Risk/Profit) talk most about intrinsic value and trade in balanced lots, keeping price within $1–2 of FV.
- **Momentum bubble–builders.** Grok-2 and Mistral-Large use twice the speculative language of Claude and issue 18–34 trade verbs per 1,000 → collective buying pushes price up to $20–22 before a late crash.

- **Risk-obsessed inflator.** GPT-3.5 has the leanest trade language but a Risk/Profit ratio ≈8 ×. Agents buy cautiously, hold, and never realize gains, so the market inflates to $18.8 yet never panics.
- **Net-seller undervaluation.** Gemini's speculative share is moderate, but its sell verbs outnumber buys (net-buy = –0.19), driving the only "reverse bubble" that trades *below* FV.

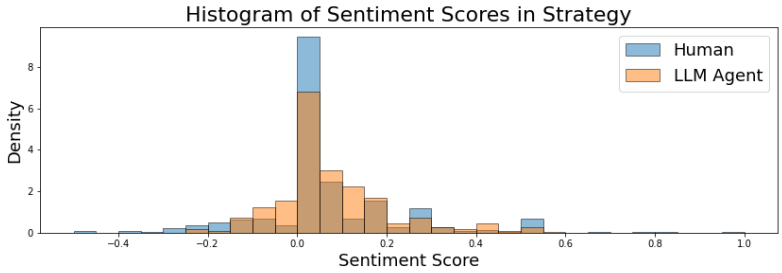
A simple multinomial logistic regression on these four features classifies the six models with 83% 5-fold CV accuracy, confirming that each family follows a linguistically—and behaviorally—distinct playbook.

Sentiment Analysis

Sentiment analysis (or opinion mining) is a natural language processing technique that measures the emotional tone of a passage of text. The aim is to assign a score that is either negative or positive to a passage to express whether it conveys a negative or positive affect. We use `nltk` (Bird, Klein, and Loper, 2009) to perform our analysis. We did not word stem or remove stopwords after Pang, Lee, and Vaithyanathan (2002).

We ran a sentiment analysis on the strategies and found that human and LLM agents produce a similar distribution of sentiment (Figure 4.5). This similarity might be due to the relatively banal nature of the subject. However, the human subjects did produce a slightly wider variance of sentiment ($\sigma_h = 0.167 > \sigma_b = 0.131$).

Figure 4.5: Distributions of sentiment for human and LLM agents



Average human sentiment was $\mu_h = 0.061$ ($\sigma_h = 0.167$). Average LLM sentiment was $\mu_b = 0.072$ ($\sigma_b = 0.131$)

A t-test revealed no significant difference (t=-1.842; p=0.063).

The strategies of the LLM agents were collected after each round, whereas the human subjects were asked to provide a strategy once after the experiment. This difference

	Sentiment x 100
Mispricing	0.294**
	(0.0674)
Round	-0.136**
	(0.0275)
Constant	8.684***
	(0.358)
Observations	18600

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.6: Fixed effects regression of sentiment score on round number and level of mispricing. Standard errors are clustered on LLM and shown in parentheses. Sentiment was multiplied by 100 to reduce the significant digits in the estimates.

Figure 4.6: Sentiment scores are plotted against the mispricing of the market compared to the fundamental value of $14



(Mispricing = Market Price - FV) The shaded region represents the 95% confidence interval.

allows us to examine the changes in sentiment over time and market conditions. We noticed that sentiment is significantly different when the market price is below the fundamental value (Figure 4.6). We define mispricing as $M_{sr} = MP_{sr} - FV_s$ where M_{sr} is the mispricing for session s in round r, MP_{sr} is the market price in session s round r, and FV_s is the fundament value for session s (which is 14 for all sessions). We ran a fixed effects regression of sentiment scores on round number and mispricing with standard errors clustered on the LLM (table 4.6). The results show that the models do tend to adjust the tone of their strategy based on market conditions, even when controlling for the specific LLMs.

4.7 Market Treatments

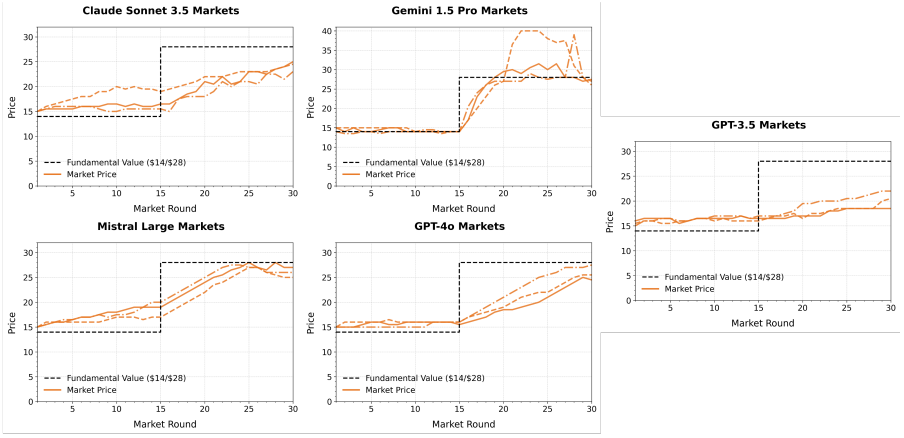
Beyond the baseline constant–fundamental design, we tested whether LLMs' rationality profiles held under more complex and dynamic market conditions, particularly when payoffs and fundamental values shift unexpectedly over time. In real financial markets, prices do not evolve in a static environment: traders must continually respond to new information about dividends, earnings, policy changes, macroeconomic conditions, or firm-specific events. Experimental asset markets have long used shock treatments of this kind to study whether traders update beliefs appropriately when fundamentals change, and whether prices remain tethered to intrinsic value or instead display inertia, overreaction, or bubble-like behavior.

To test whether our findings extend beyond a constant fundamental value environment, we conducted two separate mono-agent market experiments (each with three sessions) featuring a mid-experiment dividend shock. Beginning in Round 15, we either doubled (4.7) or halved (4.7) the dividend and redemption value, which correspondingly doubled or halved the fundamental value of the asset. The agents were not informed of this change in advance; instead, at Round 15 they were told that the dividend and redemption value had changed (either doubled or halved) and would remain at the new level for the rest of the experiment. This design mirrors a stylized unexpected news shock, allowing us to observe whether LLMs revise their forecasts and trading behavior in a manner consistent with the new fundamentals. More broadly, these treatments help distinguish between models that merely learn a stable pricing heuristic and models that can flexibly incorporate new information into valuation. Note we were unable to run Grok 2 as the public endpoint has been deprecated.

Dividend ("News") Shock Doubling

Results show that all models except GPT-3.5 quickly converged toward the new fundamental value. Claude-3.5 Sonnet, GPT-4o, Gemini-1.5 Pro, and Mistral-Large all incorporated the new information, though the process was not instantaneous, stabilizing prices near $28. GPT-3.5, however, lagged behind: its forecasts and orders remained anchored near the pre-shock equilibrium. Notably, Gemini-1.5 Pro occasionally overshoots the FV, creating minor bubbles and crashes. This treatment is especially informative because upward shocks require agents not only to recognize that the asset has become more valuable, but also to move away from a previously learned lower-price anchor. As such, it provides a test of whether models can re-anchor expectations when news implies a higher valuation regime.

Figure 4.7: Mono-Agent Market Price Charts Dividend Shock Doubling



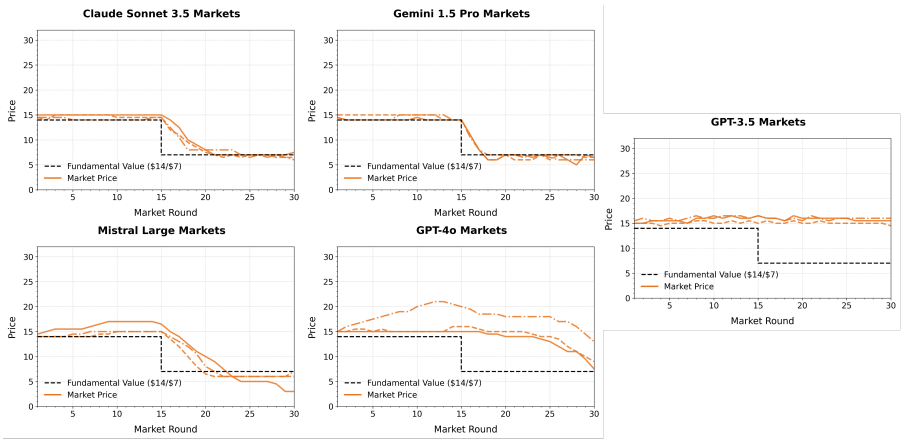
Models slowly adjust to the new FV price, with the exception of GPT-3.5. We observe that Gemini-1.5 Pro occasionally creates a modest bubble.

Dividend ("News") Shock Halving

As shown in Figure 4.8, all models except GPT-3.5 converged toward the new fundamental value. Claude-3.5 Sonnet, Gemini-1.5 Pro, and Mistral-Large all incorporated the new information within 2–3 rounds, stabilizing prices near $7. GPT-3.5 and GPT-4o, however, lagged behind, with GPT-3.5's prices remaining anchored near the pre-shock equilibrium, and GPT-4o incorporating the information slowly. This downward shock is likewise important because negative news often produces different adjustment dynamics than positive news. In both human and algorithmic settings, downward revisions can reveal whether agents are reluctant to lower valuations, whether they overcorrect, or whether they rapidly incorporate deteriorating fundamentals into price expectations.

These treatments reinforce our main result: most LLMs exhibit a fundamental-value-oriented adjustment process, while GPT-3.5 is uniquely prone to inertia and bubble-like dynamics. More generally, the shock treatments suggest that many of the stronger models are not simply reproducing a fixed pricing rule from the baseline environment, but are capable of revising valuations when new payoff-relevant information arrives. We do notice a slight increase in the propensity for the models to create bubbles under these conditions, though such episodes still occur infrequently. It may appear that the models converge more quickly to downward shocks than upward shocks; however, upward shocks are larger in absolute magnitude, and our experimental design limits orders to ±$3 of the current price. This restriction was originally implemented to keep the experiment accessible to

Figure 4.8: Mono-Agent Market Price Charts Dividend Shock Halving



All of the markets, with the exception of GPT-3.5, quickly adjust to the new fundamental value after the dividend shock. We observe that GPT4 operates within a modest bubble before converging to its redemption value towards the end of the market session.

human participants and was carried over to the LLM experiments for parity.

We also conduct multiple robustness checks in the Appendix, ensuring that slight experimental differences (such as removing the risk-elicitation task for LLMs 4.C) or LLM trading experience (4.C) do not materially alter our conclusion. While these treatments represent only a small subset of the numerous possible experimental modifications, they demonstrate that our conclusion regarding LLM rationality remains robust to various changes in market conditions.

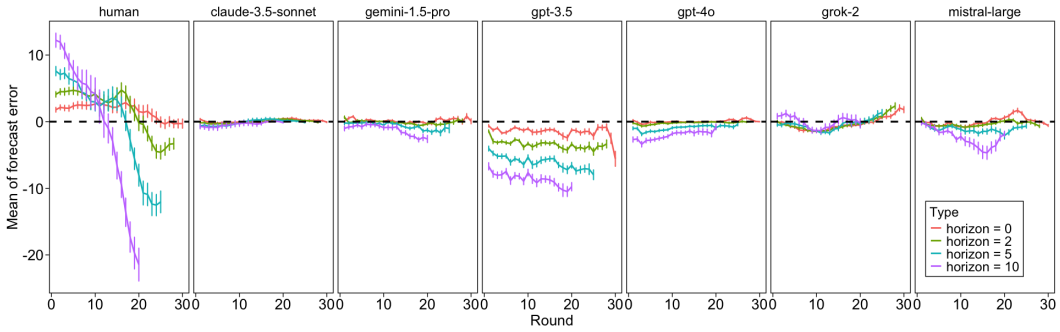
4.8 Rationality Properties of Individual LLM Forecasts

In addition to analyzing the pricing behavior of LLMs, we were also interested in how these models generate forecasts for future price periods. We captured each model's forecasts, which allowed us to assess not only the rationality of their forecasts, whether they align with fundamental valuation principles, but also whether the LLMs themselves behave rationally in response to their expectations. We compare this analysis to that of our human traders to understand whether LLMs, in the mono-agent market setting, make similar or different forecasting mistakes compared to their human counterparts. The mathematical formulation for calculating the mean and median forecasts for any given session is provided in the Appendix (Section 4.13).

Forecast errors are defined as $E^h_{i,t} = P_{t+h} - f^{t+h}_{i,t}$, where P_{t+h} denotes the actual price at round $t + h$ and $f^{t+h}_{i,t}$ is agent i's forecast formed at round t for horizon

h; positive values indicate underprediction. Full distributions and horizon-specific test results are reported in Appendix 4.13. Humans exhibit substantial systematic underprediction: the mean one-step error ($h = 1$) is 1.67. By contrast, all LLMs' mean errors are close to zero. Consistent with this, Appendix Figure 4.26 shows that LLM forecast-error distributions are more tightly centered around zero than those from human markets, indicating higher forecast precision.

Figure 4.9: Mean Forecast Error by Round



Humans exhibit substantial systematic underprediction, with a mean one–step forecast error of 1.67. By contrast, all LLMs are much closer to zero.

We also applied three additional diagnostics to compare the forecast-rationality profiles of LLM agents and human subjects: (i) **unbiasedness**—whether forecast errors are centered on zero (no systematic bias); (ii) **zero autocorrelation**—whether errors in one round fail to predict errors in later rounds; and (iii) **errors uncorrelated with forecasts**—whether the forecast itself fails to predict the size of the error (no optimism/pessimism bias).

Results are summarized in Table 4.7 and expanded upon in Section 4.13 of the Appendix. Results suggest that LLMs generally forecast more accurately than humans in the short run, but their rationality profiles differ between models. GPT-3.5 is highly biased and correlated; Claude-3.5 and GPT-4o are the most consistently rational; Gemini-1.5 Pro, Grok-2, and Mistral-Large fall in between. Importantly, no model fully reproduces the human error pattern.

4.9 Potential Limitations and Broader Impacts

Our results should be interpreted in light of several caveats. First, LLM agents do not experience incentives, emotions, or online learning in the human sense; their "strategies" emerge from the prompt structure and frozen weights, rather than endogenous reasoning or reinforcement. Second, we study stylized markets with

Table 4.7: Forecast Accuracy and Rationality Properties of Humans and LLMs.

Model	Forecast Error	Unbiasedness	Zero Autocorr.	Errors ⊥ Forecasts
Humans	1.67	0.56	0.17	0.31
Claude-3.5 Sonnet	0.09	0.46	0.08	0.16
Gemini-1.5 Pro	0.14	0.41	0.18	0.41
GPT-3.5	–1.54	0.00	0.60	0.00
GPT-4o	0.04	0.40	0.60	0.03
Grok-2	–0.31	0.56	0.18	0.27
Mistral-Large	0.00	0.44	0.14	0.36

"Forecast Error" is the mean one–step error ($E^0_{i,t}$; positive = underprediction). Other columns report the proportion of agents passing each rationality test, averaged across horizons $h \in \{0, 2, 5, 10\}$. Full horizon-specific results are reported in the Appendix.

a single asset and a short horizon. How LLM traders respond to longer horizons, short selling, or richer strategic interactions remains an open question. Finally, we evaluated only a handful of foundation models without fine-tuning. Model choice, prompting tactics, or retrieval-augmented memory may materially alter behavior.

Despite these limitations, our work has several areas of immediate impact. First, in live markets, institutional investors have already begun integrating LLM-based systems into research, execution, and portfolio processes, while regulators are increasingly examining the systemic risks that AI may pose to financial markets. As machine agents come to represent a measurable share of economic decision-making, empirical baselines are needed to understand the dynamics they generate, especially when multiple autonomous systems interact with one another in partially automated environments. What was initially framed as a question about isolated LLM traders is now also a question about agent-to-agent market ecology: how autonomous systems process news, anchor on prior valuations, and potentially amplify or dampen one another's errors.

Second, almost all major findings in behavioral finance—including loss aversion, herding, and mental accounting—were first identified in laboratory settings (Kahneman and Tversky, 2013, Thaler, 1985). As researchers seek to understand how LLM agents trade, update beliefs, and exhibit bias, we believe the same experimental approach remains the correct starting point. Controlled laboratory markets allow us to isolate how these agents respond to changing fundamentals, strategic environments, and novel information before such systems are deployed at larger scale.

Third, our findings matter for experimental methodology itself. Recent work has explored whether LLMs can substitute for human participants in laboratory finance experiments (Cui, N. Li, and H. Zhou, 2024). However, the rapid rise of autonomous and multi-agent AI systems makes the validation problem more urgent, not less. If off-the-shelf LLM agents differ systematically from humans in their pricing behavior, inertia, or susceptibility to bubble dynamics, then replacing human subjects with such agents without careful benchmarking could produce misleading inferences about economic behavior. We therefore advise caution in treating general-purpose LLMs as interchangeable with human subjects, particularly in settings involving strategic interaction, adaptation to news, and dynamic price formation.

We thus view this study as a *starting point rather than an endpoint* for incorporating LLM agents into financial and behavioral research. More broadly, we see it as an early empirical contribution to the study of autonomous economic agents: systems that may increasingly trade, negotiate, forecast, and interact with one another with limited or no human intervention.

4.10 Discussion and Conclusions

In this study, we compare the behavior of Human and Large Language Models (LLMs) in an experimental finance paradigm that has historically elicited asset bubbles and crashes when conducted with human participants. We found that LLM-agents typically converge on prices near the fundamental value, thereby exhibiting more "textbook-rational" behavior than human subjects. Although some single-model markets did display mild bubble-like phenomena, particularly GPT-3.5, which inflated prices without a clear crash, and Mistral-Large and Grok 2, which produced modest run-ups, these tendencies were weaker than what is commonly observed in human-only markets. Moreover, individual outcomes in LLM-driven markets tended to cluster more tightly around similar final portfolio values, indicating less strategic heterogeneity than seen among humans. We found that these findings persisted even under changing market conditions, news/dividend shocks. These findings collectively suggest that although LLMs may serve as rational traders, they seldom replicate the pronounced boom-and-bust patterns that emerge from human behavioral biases. We also analyze the forecasting behavior of these LLM agents and find that the majority of the LLMs generate more accurate forecasts of future price behavior than human traders; however, we observe some heterogeneity among LLM agent models in their ability to forecast rationally. Across the board, the LLM agents performed differently from humans on most tasks; therefore, our results

caution against using LLMs as replacements or for piloting for human participants in experimental finance studies. Future work may explore whether new prompting strategies or incentives can induce more human-like trading patterns from LLM-agents. However, for now, it remains clear that these models trade with a consistency and rationality that diverge significantly from the behaviors displayed by human counterparts.

4.11 Ethics Statement

We do not believe there are any ethical concerns with this work, all human data was collected under review of the IRB at Caltech. The experiments with large language models involved only synthetic agents and did not expose participants or sensitive populations to potential harm. All analyses follow standard practices in experimental economics and machine learning. We believe this work adheres fully to the ICLR Code of Ethics and raises no issues regarding privacy, fairness, security, or potential misuse.

4.12 Reproducibility Statement

All code will be released alongside the camera-ready submission. Complete experimental details, including instructions provided to agents and human participants, as well as the full set of prompts, are documented in Appendix 4.A and Appendix 4.7. These materials, together with the statistical procedures and robustness checks reported in Section 4 and Appendices E–G, ensure that all results can be readily reproduced. Our goal is not only to enable replication of the experiments presented here, but also to provide a foundation that future work can extend and build upon.

4.13 LLM Disclosure

Large Language Models (LLMs) were used in a limited capacity to assist with LaTeX formatting and debugging. They were not used for research ideation, experimental design, data analysis, or substantive writing. The authors take full responsibility for the content of this paper.

4.A Additional Experiment Details

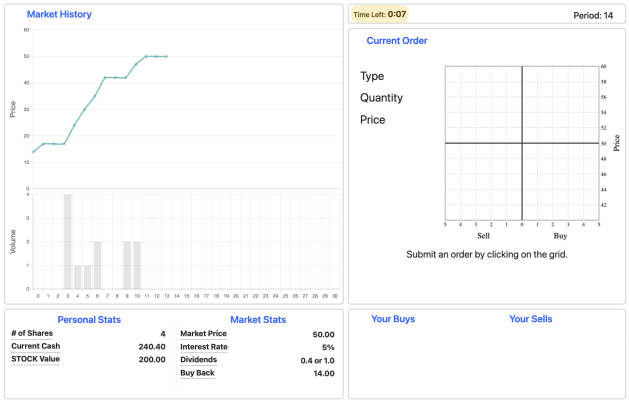
We now elaborate on the details of the experimental setup. There are 3 practice trading rounds followed by 30 rounds of trading in the main experiment. The practice rounds have no impact on the main experiment rounds and serve only to familiarize participants with the market setting. During each round, participants interact with the market through a sequence of screens.

Our experimental design involves an asset market where subjects trade a risky asset in an open-call market for 30 periods (Smith et al., 2014; Holt et al., 2017). Participants are given 4 shares of stock and 140 units of cash. Participants have two options: either invest cash in the risky asset, which pays out dividends of either 0.4 or 1 with equal probability, or hold cash, which accrues interest at a rate of 5% each trading period. The dividend payment is drawn independently for each round.

The experiment consists of a sequence of four screens in this order, as depicted in the Figures below:

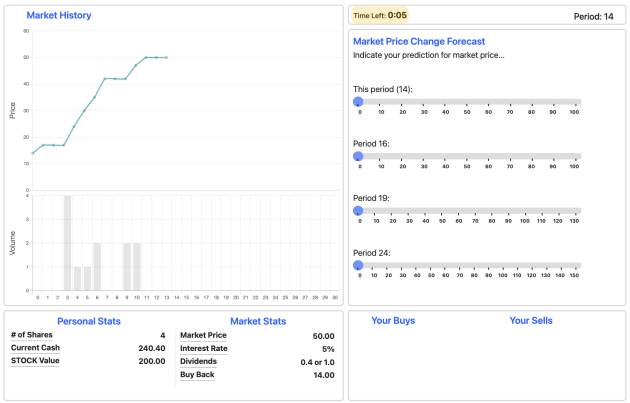
1. Order Submission Screen (20 seconds)(Figure 4.10): Participants submit orders to buy or sell the asset

Figure 4.10: Order Submission Page



2. Forecast Screen (30 seconds)(Figure 4.11): Participants submit their expectations regarding current and future market prices.

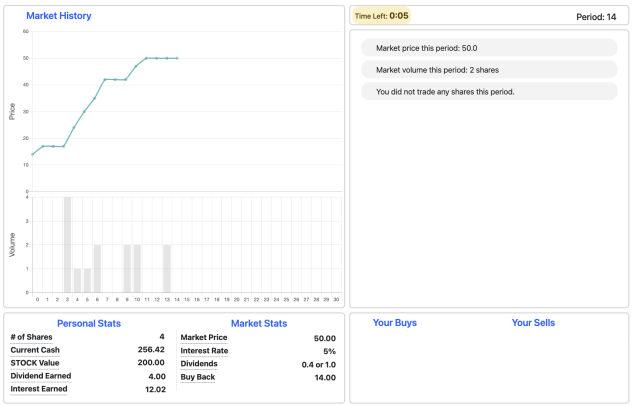
Figure 4.11: Forecast Page



Participants are asked to submit their estimation for the current market period and periods 2, 5, and 10 rounds ahead.

3. Round Results Screen (10 seconds)(Figure 4.12): Participants see the actual realized price for that period, how many shares were traded in total, and how many shares they traded.

Figure 4.12: Round Results Page



After the experiment, the risky asset is redeemed for a fixed value of 14 units, representing the fundamental price, calculated using the formula presented in Equation 4.7. Note that our experiment paradigm was designed such that this fundamental value remains constant throughout the entirety of the experiment. This enables us to objectively classify when the market is in a “bubble" state, given we always have a constant fundamental value to compare the market price to.

Price Formation

We now describe the procedure used in computing the market-clearing price for the call auctions in our experiment. For every session s and round r, the market price is determined as the price that maximizes volume Y. Sun et al., 2010. A bid submitted by subject i in round r of session s is a tuple consisting of price and quantity, (p, q). Here, $i = \{1, 2, 3\}$ indicates that a subject may submit up to 3 bids per round. The set of all bids submitted in a given market session and round is represented by $B_{s,r}$. Likewise, the set of all asks is denoted by $A_{s,r}$, with similar notation for individual asks, (p, q).

Let $P_{s,r}$ be the set of all prices submitted as bids or asks,

$$P_{s,r} = \{p \mid (p, q) \in B_{s,r} \cup A_{s,r}\}, \tag{4.1}$$

where $p((p, q))$ is a function that returns the price portion of a given order tuple (bid or ask).

Define the cumulative buy quantity of a given price p as the total quantity of all bids at or above that price:

$$Q^B(p) = \sum_{(p',q)\in B_{s,r}} q \cdot \mathbb{1}_{p' \geq p}, \tag{4.2}$$

where $q((p, q))$ is a function that returns the quantity portion of a given order tuple (bid or ask). This represents the total quantity of shares that all participants are willing to purchase at a given price.

The cumulative sell quantity is likewise defined as:

$$Q^A(p) = \sum_{(p',q)\in A_{s,r}} q \cdot \mathbb{1}_{p' \leq p}, \tag{4.3}$$

and is the total quantity of shares that all participants are willing to sell at a given price.

For a given price p, volume is determined by

$$V(p) = \min(Q^B(p), Q^A(p)). \tag{4.4}$$

The market price is then

$$p^*_{s,r} = \arg\max_{p\in P_{s,r}} V(p). \tag{4.5}$$

All prices are rounded down to the nearest integer value, a consideration made to simplify the user interface.

If there is no intersection between the bid and ask, the mid-point between the highest bid and the lowest ask is reported as the market price for that period.

Compute Details For Reproducibility

For our experiments, we leverage commercially available API endpoints from Anthropic's Claude-3.5-Sonnet (Anthropic, 2024), Google's Gemini 1.5 Pro (DeepMind, 2024), Mistral's Mistral-Large 24.11 (AI, 2024), OpenAI's GPT-3.5 and GPT-4o (OpenAI, 2025), and xAI's Grok-2 (xAI, 2024) to underpin the agents trading in our markets.

We note that our results are reproducible only insofar as these models remain accessible via their respective APIs. To contextualize the compute requirements for our experiments, Table 4.8 reports the approximate token usage per model. These figures include tokens consumed during pilot runs, which primarily involved the OpenAI and xAI models. Token usage for Gemini 1.5 Pro could not be accessed through the API console.

Provider	Model	Approx. Token Usage
Anthropic	Claude-3.5-Sonnet	17.9M
Google DeepMind	Gemini 1.5 Pro	Not Available
Mistral	Mistral-Large 24.11	18.7M
OpenAI	GPT-3.5-turbo-0125	30.1M
OpenAI	GPT-4o-2024-08-06	24.7M
xAI	Grok-2	33.9M

Table 4.8: Approximate token usage across models.

4.B Fundamental Value Calculation

For completeness, we present a derivation of the fundamental value (FV) of the asset trading in our experiments. We show that (1) the FV is constant and (2) the FV of the asset matches exactly the redemption value. We assume minimal familiarity with financial economics terminology.

In each period, a unit of stock pays a stochastic dividend drawn i.i.d. from a Bernoulli distribution over $\{0.4, 1.0\}$, with equal probability. Therefore, the expected dividend is:

$$\mathbb{E}[D_t] = 0.5 \cdot 0.4 + 0.5 \cdot 1.0 = 0.7$$

Cash holdings earn a constant risk-free return of $r = 5\%$ per period. Consequently,

the present value of receiving a dividend payment D_t (or any transfer in general) some t rounds in the future is:

$$\frac{D_t}{(1+r)^t}$$

Let T be the final trading period, and suppose the asset is redeemed at time T for a fixed amount $V_T = 14$. Then the fundamental value of the asset in any round $1 \leq t < T$ is the expected discounted value of all future dividends plus the redemption value:

$$P_t^* = \frac{V_T}{(1+r)^{T-t}} + \sum_{k=1}^{T-t} \frac{\mathbb{E}[D_k]}{(1+r)^k} \tag{4.6}$$

While Equation 4.6 is elegant in that it builds only off of the present values of expected dividend payments and redemption value, it is not immediately clear why P_t^* is constant in our setting. We now show this.

$$P_t^* = \frac{V_T}{(1+r)^{T-t}} + \mathbb{E}[D_k] \sum_{k=1}^{T-t} \frac{1}{(1+r)^k}$$

We may now compute the standard finite geometric sum.

$$P_t^* = \frac{V_T}{(1+r)^{T-t}} + \mathbb{E}[D_k] \cdot \frac{1}{1+r} \left(\frac{1 - (\frac{1}{1+r})^{T-t}}{\frac{r}{1+r}} \right)$$

Manipulating the expression, we obtain the following:

$$P_t^* = \frac{V_T}{(1+r)^{T-t}} + \mathbb{E}[D_k] \cdot \left(\frac{1 - (\frac{1}{1+r})^{T-t}}{r} \right)$$

$$P_t^* = \frac{V_T}{(1+r)^{T-t}} + \frac{\mathbb{E}[D_k]}{r} - \frac{\mathbb{E}[D_k]}{r(1+r)^{T-t}}$$

Recall that $V_T = \$14 = \frac{\$0.7}{0.05} = \frac{\mathbb{E}[D_k]}{r}$. Thus, we can complete our simplification to obtain:

$$P_t^* = \frac{\mathbb{E}[D_k]}{r} \tag{4.7}$$

Notice that Equation 4.7 is time independent. Hence, we have shown (1). Further, given the experiment parameters $P_t^* = \$14$, which is exactly the redemption value, demonstrating (2).

4.C Additional Treatments/Robustness

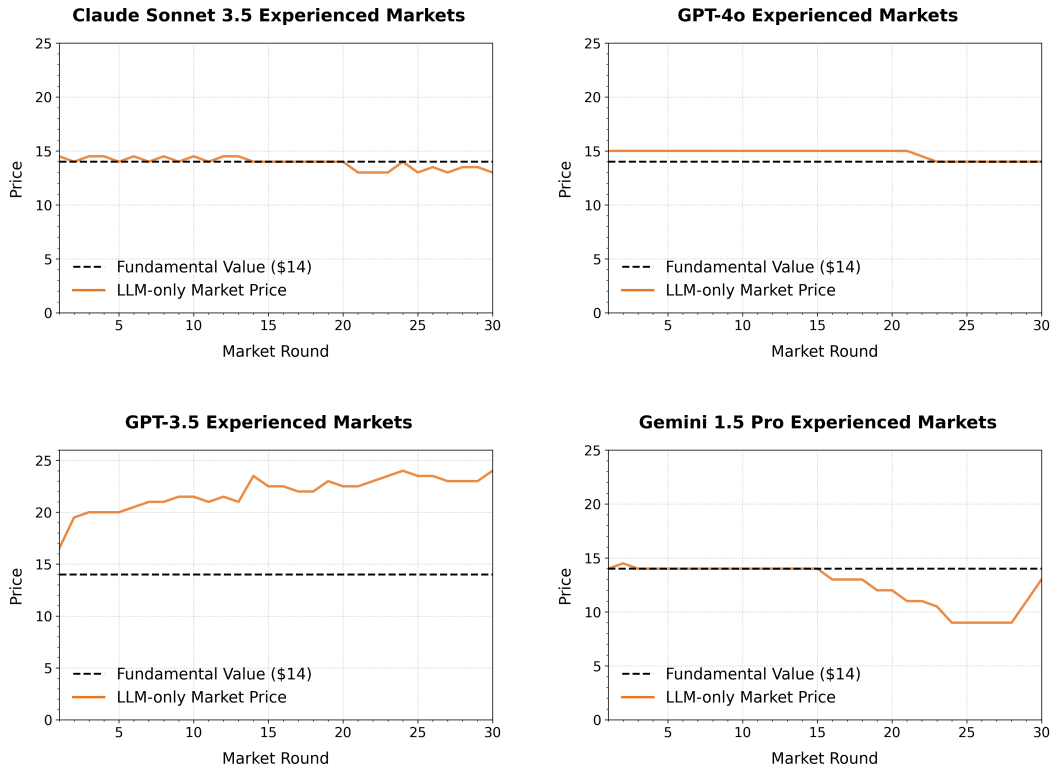
Experienced Sessions

We tested whether prior exposure changes agent behavior via an experienced-session variant of the mono-market. Each model completed two full 30-round sessions with

the standard environment (i.i.d. dividends, 5% cash interest, redemption at 14; prices fully endogenous to submitted orders). In Session 2, the prompt additionally provided the complete market history from Session 1 (prices/volumes/clearing outcomes) and the agent's own PLANS/INSIGHTS files from that session. Chain-of-thought was enabled in both sessions. Dividend paths were re-sampled; agent sampling (e.g., temperature) was not fixed across runs.

Performance in the "experienced" Session 2 was essentially unchanged relative to Session 1: mean absolute deviation from fundamental value, realized profit, and forecast accuracy remained at similar levels for all models. Figure 4.13 shows the price paths for the experienced runs. The incidence and magnitude of bubbles did not show a consistent increase or decrease. In short, exposure to the entire prior market, together with perfect recall of that history, did not materially alter trading outcomes. This suggests that substantive learning (if any) may require changing objectives (e.g., fine-tuning/reward shaping) or the environment (e.g., information asymmetries, cross-asset substitution, transaction costs). Note: Grok 2 has been deprecated, and thus we were unable to run this treatment with that model. Mistral-Large had technical errors that will be rerun for camera-ready submission.

Figure 4.13: Experienced Market Trajectories for different LLMs.



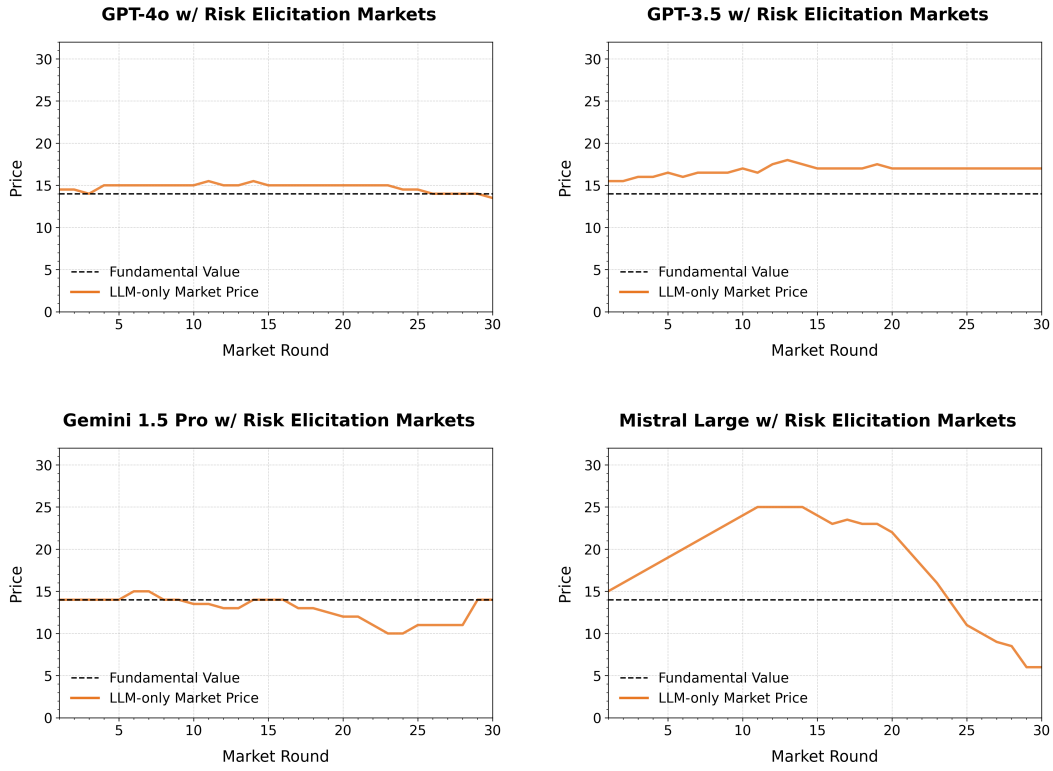
Risk Elicitation

We tested whether inserting a standard risk-elicitation task primes LLM trading. For each model, we ran one mono-market session (30 rounds, i.i.d. dividends, 5% cash interest, redemption value at 14) and added Holt–Laury–style lotteries at the same cadence used in our human studies (interleaved prompts between rounds). The task was exogenous: lottery choices did not affect cash balances, inventory constraints, or market clearing. Prompts otherwise matched the baseline; prices remained fully endogenous to submitted orders.

Adding risk-elicitation did not materially change outcomes relative to the baseline per model, as shown in Figure 4.14. Mean absolute deviation from fundamental value, bubble incidence/magnitude, realized profit, turnover, and forecast accuracy were all at similar levels; no consistent directional shift appeared in any model class.

In this environment, brief risk-preference queries do not prime LLM agents toward systematically different trading behavior. This aligns with our human-protocol rationale (the task is orthogonal to trading) and supports the symmetry claim: omitting the lotteries for LLMs in the main experiments does not explain the LLM–human differences we report. Note: Grok 2 has been deprecated, and thus we were unable to run this treatment with that model. Gemini-1.5 had technical errors that will be rerun for camera-ready submission.

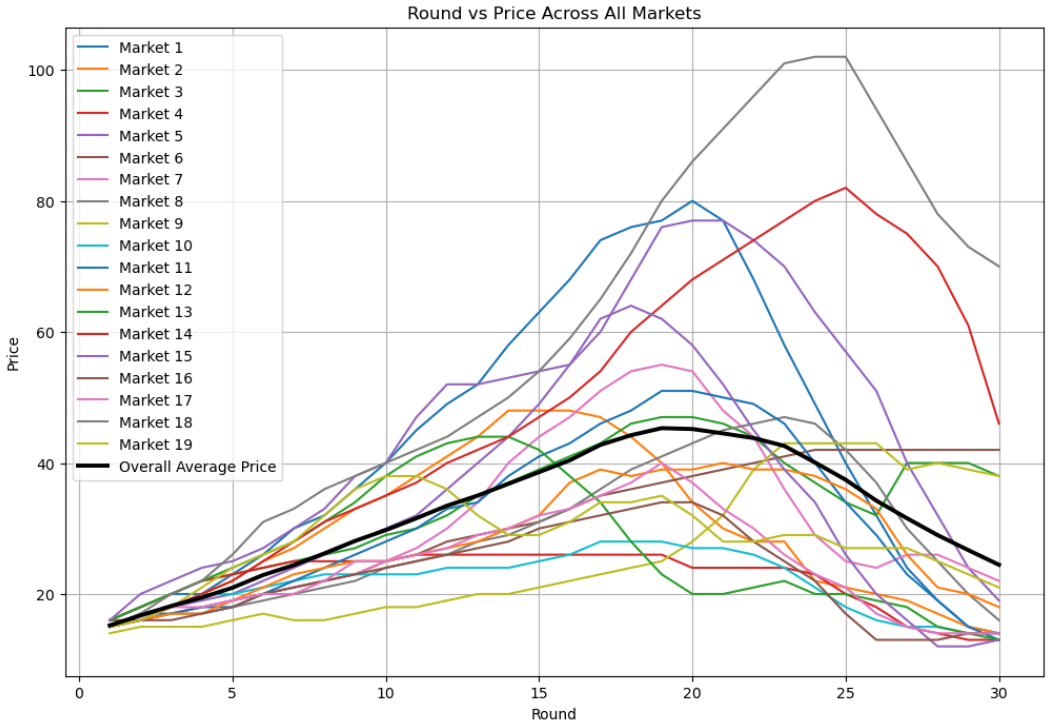
Figure 4.14: Risk-Elicitation Priming Market Trajectories for different LLMs.



4.D Additional Analysis

Human Market Trajectories

Figure 4.15: Human Market Trajectories

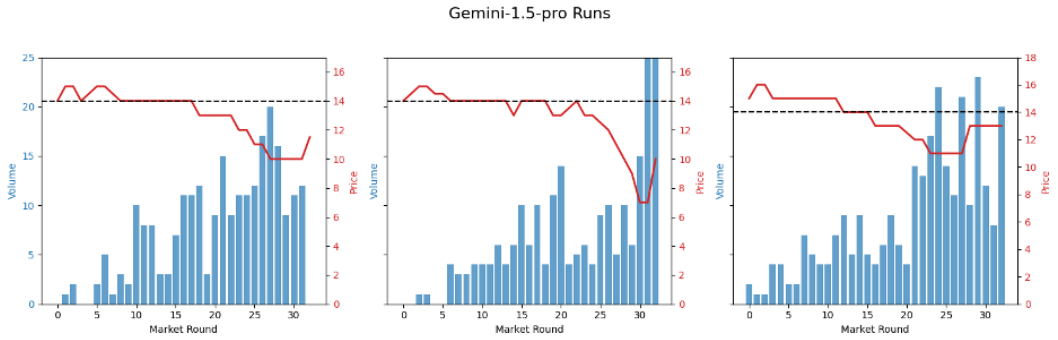


Individual market trajectories from human-only experiments

Detailed individual market analysis

- **Gemini 1.5 Pro:** Gemini markets seem to generate a reverse bubble trend, where the asset is initially sold off below fundamental and then bid back up as redemption grows closer (Figure 4.16). While overall rationality given by the MSE (~ 3.1) from the fundamental metric is low, the behavior is the exact opposite of what we have seen in the majority of human experiments and indicates that the model likely does not truly understand the task nor model any biases that humans may exhibit.

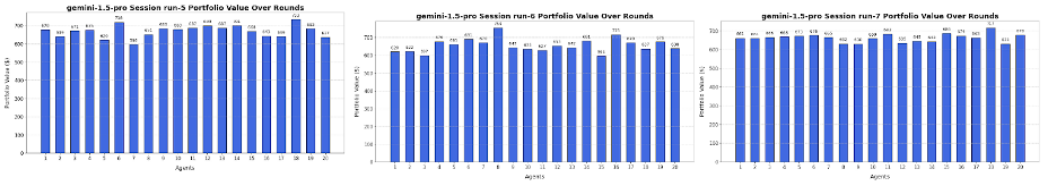
Figure 4.16: Gemini 1.5 Pro Markets



This market seems closest to the Erratic (E) hypothesis. This is also supported by the Pearson correlation coefficient (PCC) as compared to the average human market price of -0.4009, indicating that there are large differences between this model's performance and that of humans.

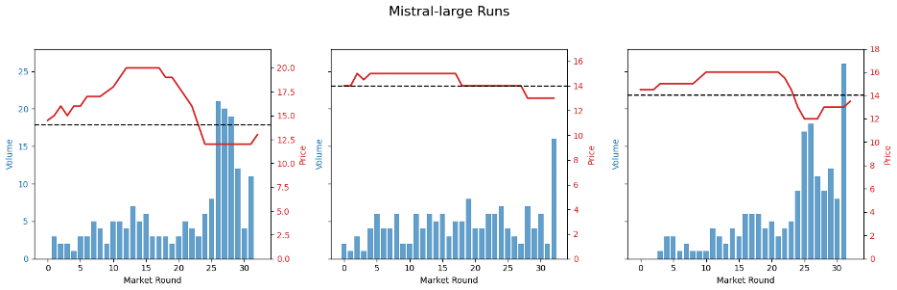
We can also analyze the uniformity of the market, as you can see in Figure 4.17, compared to other markets, the traders do seem to have a more varied earning profile. This is also shown in the Portfolio Value variance, which is 45.58, higher than all but one of the other single-model markets.

Figure 4.17: Gemini 1.5 Pro Portfolio Variance



- **Mistral-Large 24.11** Mistral seems to generate markets that appear mostly like the type of bubbles we've seen in our human markets, though the size of the asset bubble is greatly reduced from our median human market and the price crashes below fundamental during the final rounds (Figure 4.18).

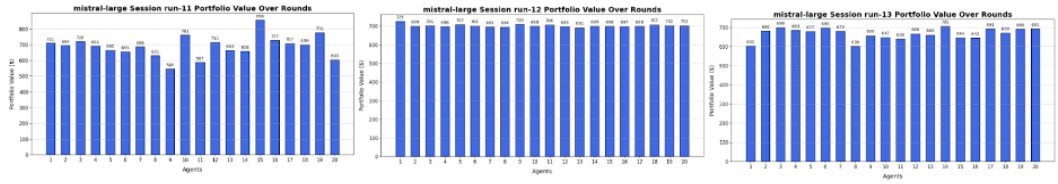
Figure 4.18: Mistral-Large 24.11 Portfolio Markets



This behavior, while not economically rational (purely rational agents would only buy a price below 14 and would sell any price above 14) does correspond somewhat with the behavior we've seen out of human players, though the Pearson correlation coefficient (PCC) as compared to the average human market price of -0.1118 are large differences. This negative PCC likely pertains to the behavior in these markets after the bubble crashes. This market seems closest to the Human (H) hypothesis.

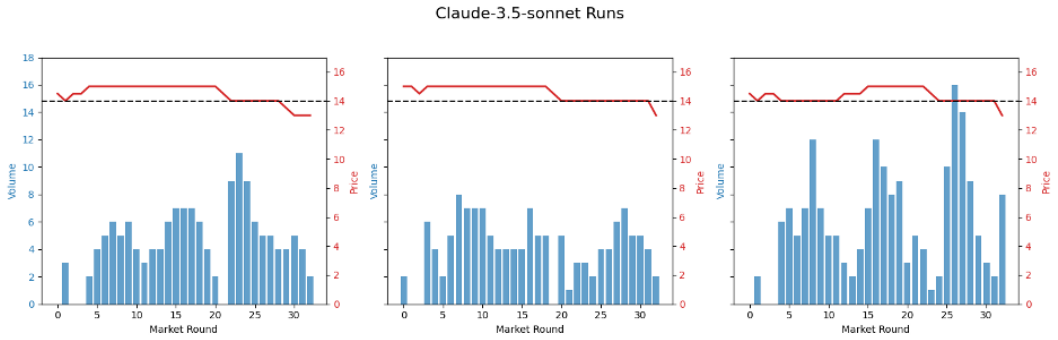
We can also analyze the uniformity of the market, as can be seen in the below graph (Figure 4.19), compared to other markets, the traders do seem to have a more varied earning profile. This is also shown in the mean Portfolio Value variance, which is 49.96, the highest of the single-model markets.

Figure 4.19: Mistral-Large 24.11 Portfolio Variance



- **Claude-3.5-Sonnet:** Claude markets seem to demonstrate highly 'rational' behavior, the price remains very close to the fundamental value throughout the market (Figure 4.20). This indicates that the model likely understands the "rational" strategy. This market is the "most" rational based on the MSE metric with the lowest MSE from fundamentals of all the markets (~ 0.53).

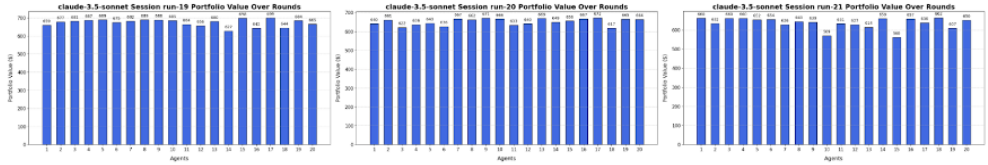
Figure 4.20: Claude-3.5-Sonnet Markets



This market seems closest to R. Which is further supported by the Pearson correlation coefficient (PCC) as compared to the average human market price of being $\sim$ 0.0013, showing that there are large differences between this model's performance and that of humans.

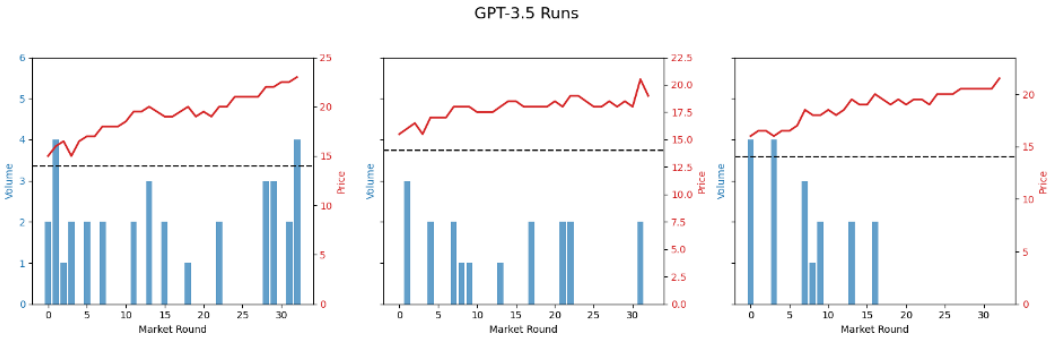
We can also analyze the uniformity of the market, as can be seen in the below graph (Figure 4.21), compared to other markets, the traders do seem to have a more uniform earning profile. This is also shown in the mean Portfolio Value variance, which is 26.15, lower than most of the other single-model markets.

Figure 4.21: Claude-3.5-Sonnet Portfolio Variance



- **GPT-3.5:** Demonstrates bubble behavior without the crash, where prices rise steadily with no indication of a return to fundamental value (Figure 4.22). Unlike human markets, trading volume almost disappears toward the end of the experiment, whereas humans generally trade more as the end approaches. This model is the least "rational" by the MSE metric, having a far and away the highest MSE of all the models ($\sim$ 26.36).

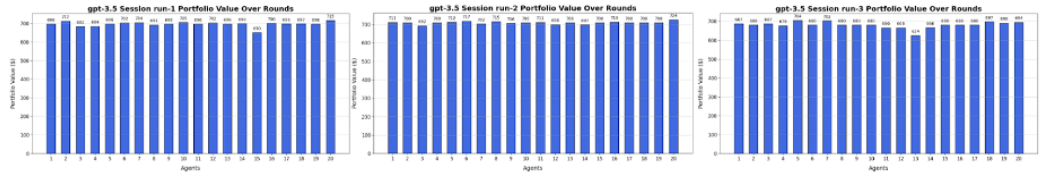
Figure 4.22: GPT-3.5 Markets



This market seems closest to H, as some bubbles without crashes were observed in human experiments. Interestingly, this model has a relatively high Pearson correlation coefficient (PCC) as compared to the average human market price of ~ 0.4897, but it is clear that there are still large differences between this model's performance and that of humans.

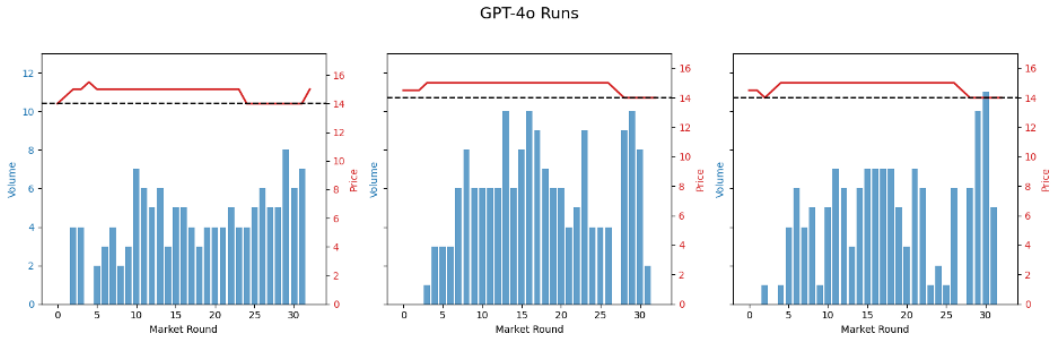
The Portfolio Value variance for this market is 30.44 (Figure 4.23), one of the lower values among single-model markets, indicating generally uniform behavior.

Figure 4.23: GPT-3.5 Portfolio Variance



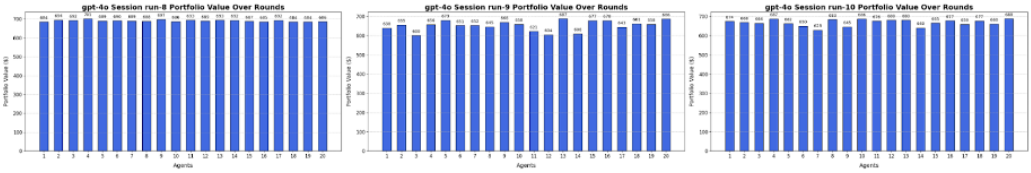
- **GPT-4o:** Demonstrates mostly rational behavior, with prices remaining very close to the fundamental value throughout the market (Figure 4.24). This indicates that the model likely understands the "rational" strategy. This is emphasized quantitatively by the model having the second lowest MSE compared to fundamental at (~ 0.79). This market seems closest to the Rational (R) hypothesis.

Figure 4.24: GPT-4o Markets



Additionally, unlike human markets, almost all the agents performed very similarly to each other, indicating that there was not a large amount of behavior variance among the agents. This is also shown in the mean Portfolio Value variance of 22.76 (Figure 4.25), which is the lowest amongst the single-model markets.

Figure 4.25: GPT-4o Portfolio Variance



4.E Rationality Properties of Individual LLM Forecasts

In addition to analyzing the pricing behavior of LLMs, we were also interested in how these models generate forecasts for future price periods. We captured each model's forecasts, which allowed us to assess not only the rationality of their forecasts, whether they align with fundamental valuation principles, but also whether the LLMs themselves behave rationally in response to their expectations. We compare this analysis to that of our human traders to understand whether LLMs, in the Single-LLM market setting, make similar or different forecasting mistakes compared to their human counterparts. The mathematical formulation for calculating the mean and median forecasts for any given session is provided in the appendix (Appendix 4.13).

Forecasting Framework

Fix a session (we will suppress session notation). For each period t, participants provide forecasts before the price revelation. The forecast made in period t, denoted $f_{i,t}^{t}$, predicts the price P_t. The forecast for future periods, made in period t, is denoted as $f_{i,t}^{(t+h)}$, where $h \in \{0, 2, 5, 10\}$ is the prediction horizon (or how far into the future agents are forecasting).

The forecast error for an individual i in period t, forecasting P_{t+h}, is given by: $E_{i,t}^{h} = P_{t+h} - f_{i,t}^{(t+h)}$ A positive value indicates that the agent underforecasted the price, while a negative value indicates an overforecast. In the appendix, Table 4.10 presents the full descriptive statistics of prices and forecasts, while Table **??** reports the forecast errors for each model and for our human participants.

For next-round forecasts, the Mistral-Large model achieves the smallest average error (0.00), highlighting its most rational performance. GPT-4o follows with an error of 0.04, Claude-3.5-Sonnet with 0.09, and Gemini-1.5-Pro with 0.14. Grok-2 and GPT-3.5 record errors of –0.31 and –1.54, respectively. By comparison, human participants exhibit a one-round forecast error of approximately 1.67. Thus, all AI models outperform human forecasters in terms of forecast accuracy.

Distribution of Forecast Errors

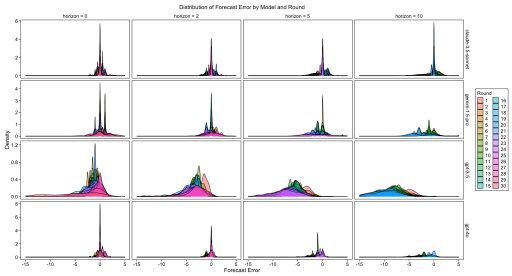
We investigate the distribution of forecast errors for each AI model across each horizon, grouped by experimental rounds. Figure 4.26 demonstrates that these distributions vary both by model and by prediction horizon.

At the model level, all the models exhibit more tightly centered error distributions around 0 when compared to human markets. While GPT-3.5's errors are the most dispersed, it has notably more concentrated distributions than human participants. GPT-3.5's distribution also shows a leftward shift, most notably at the five- and ten-period horizons, indicating an increasing underforecast bias over successive rounds, a pattern that somewhat parallels the trend among human forecasters. The other models do not show this shift.

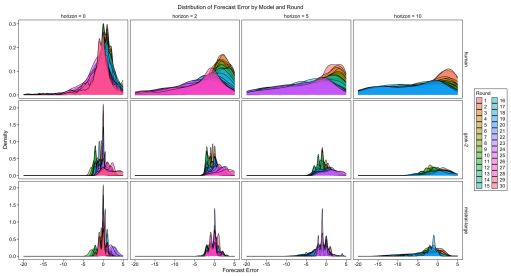
Next, we compute the mean forecast error for each round and horizon across all AI models and human participants. Figure 4.9 shows how this average error evolves over successive rounds. Among the models, Claude-3.5-Sonnet performs best—its mean error remains closest to zero at every horizon. GPT-4o initially overforecasts but steadily converges toward zero in later rounds, reflecting increasing accuracy. By contrast, GPT-3.5 consistently overforecasts as rounds progress, making it the

least accurate model. Notably, none of the AI models replicate the human pattern of early underforecasting and then subsequent overforecasting in later rounds.

Figure 4.26: Comparison of forecast errors across rounds for human participants and AI models



(a) Gemini-1.5-Pro, Claude-3.5-Sonnet, GPT-3.5, and GPT-4o models



(b) Human participants, Grok-2, and Mistral-Large models

Unbiasedness

At the individual agent level, we test whether the mean forecast error $E_{i,t}^{h}$ is zero across t for each horizon h. This yields four test statistics per agent, corresponding to the four horizons. For each case, we conduct t-tests to determine whether the mean forecast error is significantly different from zero, i.e., whether it is biased. For each horizon, we then count the number of agents whose mean forecast error is not significantly different from zero, indicating unbiased forecasts, and compute the percentage of unbiased agents for each model. The first block of Table 4.9 presents, for each model and horizon, the percentage of agents whose forecasts are unbiased. The corresponding results are also shown in Figure 4.28. Among the models, GPT-3.5's forecasts are the most biased, as all forecast errors for the four horizons differ from 0. Claude-3.5-Sonnet, Gemini-1.5-Pro perform most similarly to humans; the bias measure decreases as the horizon increases. GPT-4o performs best at horizon = 0, indicating the most rational forecasting and the least bias.

Zero Autocorrelation

To test for zero autocorrelation, we examine whether forecast errors from period t should be independent of forecast errors in subsequent periods. The guiding principle is that future forecast errors should not depend on the information available at the time the forecast was made.

Table 4.9: Rationality test results

Test	Horizon	Human	Claude-3.5 Sonnet	Gemini-1.5 Pro	GPT-3.5	GPT-4o	Grok-2	Mistral-Large
Unbiasedness	0	0.619	0.833	0.778	0	1.000	0.657	0.972
	2	0.623	0.389	0.583	0	0.597	0.554	0.528
	5	0.570	0.306	0.194	0	0.000	0.516	0.139
	10	0.438	0.292	0.069	0	0.000	0.547	0.111
Zero autocorrelation	0	0.420	0.069	0.208	0.819	0.458	0.031	0.194
	2	0.092	0.236	0.208	0.847	0.648	0.016	0.208
	5	0.068	0.000	0.083	0.806	0.403	0.062	0.139
	10	0.086	0.028	0.208	0.861	0.486	0.109	0.097
Forecast error uncorrelated with forecast levels	0	0.625	0.028	0.389	0	0.056	0.438	0.403
	2	0.349	0.308	0.431	0	0.069	0.203	0.472
	5	0.145	0.175	0.586	0	0.000	0.141	0.306
	10	0.117	0.123	0.643	0	0.000	0.125	0.250

To test this, we estimate the regression: $E_{t,0} = \beta_0 + \beta_1 E_{t-1,0} + \epsilon_t$, where the hypothesis is $\beta_1 = 0$. This zero-autocorrelation condition extends similarly to longer forecast horizons.

We divide the data at horizon $h = 0$ by agent and use ordinary least squares regression to estimate β_1 for each agent at different horizons. For each model and horizon, we then count the number of agents whose β_1 is not significantly different from zero—indicating zero autocorrelation and thus rationality—and calculate the corresponding percentage.

The second block of Table 4.9 presents the results of the autocorrelation test. We see some heterogeneity among the models; GPT-3.5 appears to be the most rational, with the highest percentage of agents showing no autocorrelation. In contrast, Claude-3.5-Sonnet and Grok-2 exhibit the strongest autocorrelation, with the lowest percentage of agents passing the test. Humans show autocorrelation when the horizon is 0, a pattern similar to GPT-4o (humans: 42%, GPT-4o: 45.8%).

Correlation Between Forecast Errors and Forecasts

Forecast errors, $P_{t+h} - f_{i,t}^{(t+h)}$, should be uncorrelated with forecasts: $\text{cor}(P_{t+h} - f_{i,t}^{(t+h)}, f_{i,t}^{(t+h)}) = 0$ This principle is grounded in two key intuitions. First, the current forecast is part of an agent's information set. If any variable within that set can predict future forecast errors, a rational agent should incorporate it to reduce those errors. Second, a negative correlation between forecast errors and forecasts suggests a systematic bias: when an agent is optimistic, making a high forecast, they tend to overestimate the actual outcome, resulting in a negative forecast error.

To test this, we compute these correlations separately for each forecast horizon and agent by estimating the regression: $E_{i,t}^{h} = \beta_0 + \beta_1 f_{i,t}^{(t+h)} + \epsilon_{i,t}^{h}$

For the four values of h, we count the number of agents whose β_1 is not significantly different from zero for each model and horizon, indicating no correlation between forecast errors and forecasts, which is consistent with rationality.

The third block of Table 4.9 reports the results of the correlation test across different horizons (see also Figure 4.30). When the horizon is 0, humans perform the best, followed by Grok-2. In contrast, GPT-3.5 and GPT-4o perform the worst, as their forecast errors are consistently correlated with their forecasts across nearly all horizons.

In summary, LLMs generally produce more stable and centered forecast errors than humans. GPT-3.5 shows high bias and strong forecast-error correlation but low autocorrelation. Claude-3.5-Sonnet and Gemini-1.5-Pro are closest to human behavior, while GPT-4o performs well at short horizons. Each model has its strengths and weaknesses, but no model fully replicates human behavior across all tests.

Additional Forecast Information

Mean and Median Forecasts

The mean forecast for session s at period $t + h$ is:

$$m_t^{(t+h)} = \frac{\sum_{i=1}^{N_s} f_{i,t}^{(t+h)}}{N_s}$$

Where N_s is the number of participants in session s.

The median forecast is calculated as follows: Arrange the values of $f_{t,t+h}^{(i)}$ in ascending order:

$$f_{t,t+h}^{(1)}, f_{t,t+h}^{(2)}, \ldots, f_{t,t+h}^{(N_s)}$$

where

$$f_{t,t+h}^{(1)} \leq f_{t,t+h}^{(2)} \leq \cdots \leq f_{t,t+h}^{(N_s)}$$

If N_s is odd:

$$\text{Median} = f_{t,t+h}^{\left(\frac{N_s+1}{2}\right)}$$

If N_s is even:

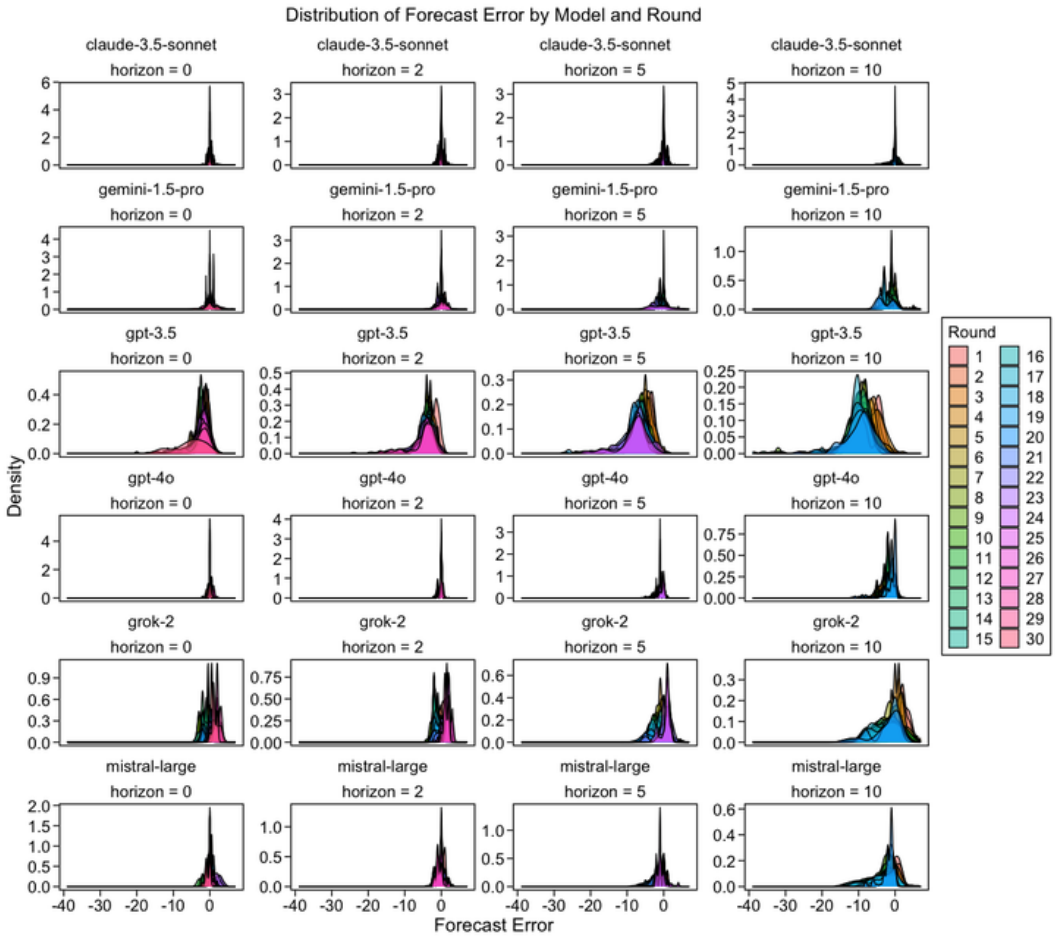
$$\text{Median} = \frac{f_{t,t+h}^{(N_s/2)} + f_{t,t+h}^{(N_s/2+1)}}{2}$$

Table 4.10: Descriptive statistics for Price and Forecasts by source.

Source	Variable	Mean	SD	Min	Max	Q1	Q3
claude-3.5-sonnet	Price	14.65	1.00	13.00	20.00	14.00	15.00
	Forecast current	14.67	0.98	13.00	20.00	14.00	15.00
	Forecast 2 rounds later	14.78	1.12	14.00	21.00	14.00	15.00
	Forecast 5 rounds later	14.90	1.49	14.00	22.00	14.00	16.00
	Forecast 10 rounds later	15.13	2.04	14.00	25.00	14.00	16.00
gemini-1.5-pro	Price	13.43	2.02	7.00	20.00	13.00	14.50
	Forecast current	13.68	1.91	6.00	21.00	13.00	15.00
	Forecast 2 rounds later	13.87	1.77	7.00	22.00	14.00	15.00
	Forecast 5 rounds later	14.26	1.48	8.00	23.00	14.00	15.00
	Forecast 10 rounds later	14.55	1.11	9.00	22.00	14.00	15.00
gpt-3.5	Price	16.48	0.92	14.00	20.00	16.00	17.00
	Forecast current	17.98	2.15	14.00	35.00	17.00	18.00
	Forecast 2 rounds later	19.92	2.06	14.00	38.00	19.00	21.00
	Forecast 5 rounds later	22.69	2.42	15.00	45.00	21.00	24.00
	Forecast 10 rounds later	25.28	2.76	17.00	55.00	24.00	27.00
gpt-4o	Price	14.94	0.87	14.00	20.00	15.00	15.00
	Forecast current	14.95	0.91	14.00	21.00	15.00	15.00
	Forecast 2 rounds later	15.20	1.10	13.00	22.00	15.00	15.00
	Forecast 5 rounds later	15.98	1.45	14.00	25.00	15.00	16.00
	Forecast 10 rounds later	17.06	2.10	14.00	28.00	16.00	18.00
grok-2	Price	18.17	2.13	14.00	22.00	16.00	20.00
	Forecast current	18.15	2.45	14.00	23.00	16.00	20.00
	Forecast 2 rounds later	18.39	2.64	14.00	24.00	16.00	20.00
	Forecast 5 rounds later	18.90	2.76	14.00	26.00	17.00	21.00
	Forecast 10 rounds later	19.07	3.02	14.00	28.00	17.00	21.75
mistral-large	Price	15.17	2.09	12.00	20.00	14.00	16.00
	Forecast current	15.34	2.21	10.00	22.00	14.00	16.00
	Forecast 2 rounds later	15.77	2.45	9.00	23.00	14.00	17.00
	Forecast 5 rounds later	16.60	2.75	9.00	25.00	14.00	18.00
	Forecast 10 rounds later	17.76	3.23	14.00	27.00	15.00	19.00
Human	Price	31.31	16.42	12.00	102.00	19.00	40.00
	Forecast current	15.50	8.66	1.00	30.00	8.00	23.00
	Forecast 2 rounds later	16.50	8.08	3.00	30.00	9.75	23.25
	Forecast 5 rounds later	18.00	7.21	6.00	30.00	12.00	24.00
	Forecast 10 rounds later	20.50	5.77	11.00	30.00	15.75	25.25

Additional Distribution of Forecast Errors Graph

Figure 4.27: Additional Forecast Error Distribution Graph



Median Forecast Error

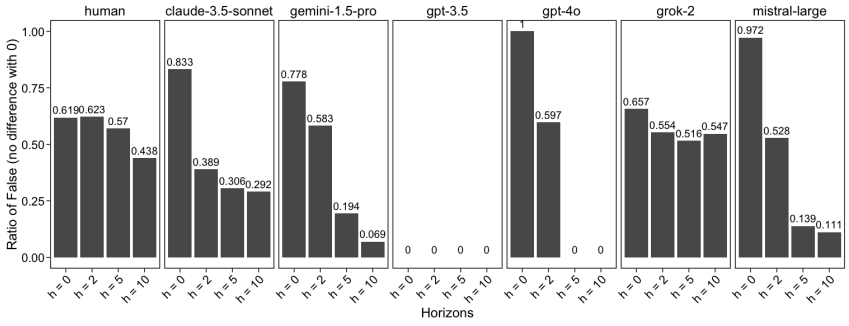
We calculate the median forecast error for each round across different horizons. Figure 5 in the main text illustrates how the median forecast error evolves over time for each horizon. Claude-3.5-Sonnet performs the best, as its mean forecast for each horizon remains closest to 0 compared to other models. GPT-4o shows a trend of initially overforecasting prices before gradually shifting toward 0 in later rounds, indicating a movement toward greater accuracy. In contrast, GPT-3.5 performs the worst, consistently overforecasting prices as the rounds progress.

Unbiasedness

At the individual agent level, we test whether the mean forecast error $E_{i,t}^{h}$ is zero across t for each horizon h. This provides four test statistics for each value of h. We conduct t-tests to determine whether the mean forecast error is statistically different from zero.

Specifically, we test whether the mean forecast error for each model and horizon is significantly different from zero. We use the t test at agent level for four horizons and get the p value. Then we count the number of agents whose mean forecast error is significantly different from 0 at each horizon, and calculate the percentage of each model. Figure 4.28 presents the percentage of no difference with 0 of each model at each horizon. Among the models, gpt-3.5 performs worst, as all forecast error for four horizons are all different from 0, Claude-3.5-Sonnet, Gemini-1.5-Pro performs most similarly to humans, the ratio gets lower as horizon increase. GPT-4o performs best at horizon = 0, indicating rational forecasting.

Figure 4.28: The percentage of no autocorrelation bias of each model at each horizon



Zero Autocorrelation To test for zero autocorrelation, we examine whether forecast errors from period t should be independent of forecast errors in subsequent periods. The guiding principle is that future forecast errors should not depend on the information available at the time the forecast was made.

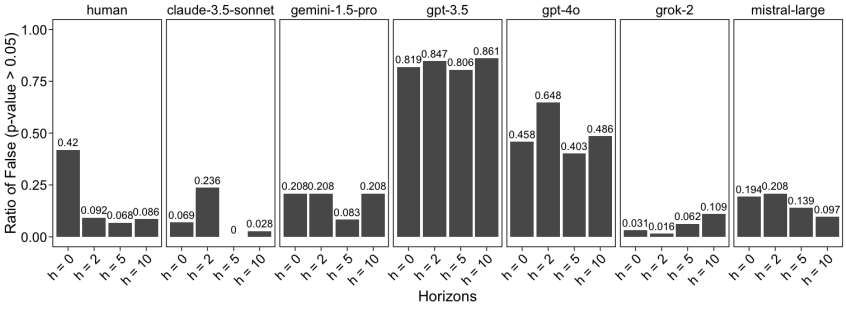
For example, the forecast error in period 8 is:

$$E_{i,8}^{0} = P_8 - f_{i,8}^{8}$$

This error should be uncorrelated with the forecast error in period 9:

$$P_9 - f_{i,9}^{9}$$

Figure 4.29: The results of the autocorrelation test.



A similar zero-autocorrelation principle applies for longer horizons. To test this, we estimate the regression:

$$E_{t,0} = \beta_0 + \beta_1 E_{t-1,0} + \epsilon_t$$

where the hypothesis is $\beta_1 = 0$.

We divide the data of horizon $h = 0$ by agent and use ordinary least squares (OLS) regression to estimate β_1 at the agent level for different horizons. We then count the number of agents whose β_1 is not significantly different from 0 (p-value > 0.05) for each model and each horizon and calculate its percentage.

Figure 4.29 presents the results of the autocorrelation test. Among the models, GPT-3.5 appears the most rational, as it has the highest percentage of agents showing no autocorrelation. In contrast, Claude-3.5-Sonnet and Grok-2 exhibit the most autocorrelation, with the lowest percentage of agents passing the test.

Correlation Between Forecast Errors and Forecasts Forecast errors, $P_{t+h} - f_{i,t}^{(t+h)}$, should be uncorrelated with forecasts:

$$\text{cor}(P_{t+h} - f_{i,t}^{(t+h)}, f_{i,t}^{(t+h)}) = 0$$

This principle is based on two intuitions: 1. The current forecast is part of an agent's information set. If something in their information set can predict future errors, they should use it to reduce the error. 2. A negative correlation between forecast errors and forecasts implies that when an agent is optimistic (high $f_{i,t}^{(t+h)}$), they tend to be too optimistic (negative $P_{t+h} - f_{i,t}^{(t+h)}$).

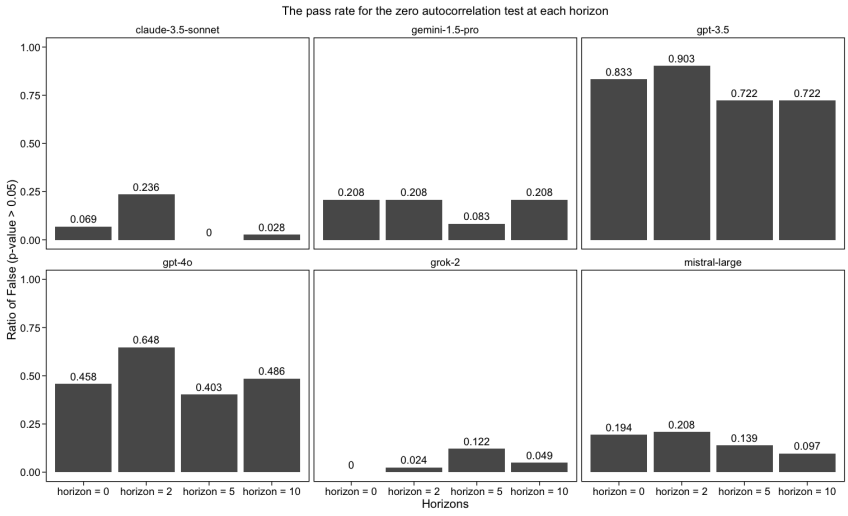
To test this, we compute these correlations separately for each forecast horizon and agent by estimating the regression:

$$E_{i,t}^h = \beta_0 + \beta_1 f_{i,t}^{(t+h)} + \epsilon_{i,t}^h$$

for the four values of h. We then count the number of agents whose β_1 is not significantly different from zero (p-value > 0.05) for each model and horizon, indicating no correlation between forecast errors and forecasts.

Figure 4.30 presents the results of the correlation test across different horizons and models. In this test, Gemini-1.5-Pro performs best, as it has the highest percentage across all four horizons compared to other models. In contrast, GPT-3.5 and GPT-4o perform worse, as most agents from these models exhibit forecast errors that are correlated with forecast levels.

Figure 4.30: Results of the correlation between forecast error and forecast



4.6 Human Experiment Details

Design Details

In the base experiment, all human participants saw the exact same instructions; there were no treatment variables. The study was conducted using a web-based experimental software developed by one of the authors.

There were two types of subjects participating in each market: in-lab participants, collected at the university, and online participants using Prolific. For each market, the in-lab participants had biometric data measured using Empatica E4 SCR devices and newer (acquired April 2024) Empatica EmbracePlus devices. We do not analyze any biometric data in this paper.

Before participating in the experiment:

1. Subjects sign a consent form agreeing to participate in the experiment.

Table 4.12: Descriptive statistics and forecast errors for each model and humans.

Source	Variable	Mean	SD	Min	Max	Q1	Q3
claude-3.5-sonnet	Price	14.65	1.00	13.00	20.00	14.00	15.00
	Forecast current	14.67	0.98	13.00	20.00	14.00	15.00
	Forecast t+2	14.78	1.12	14.00	21.00	14.00	15.00
	Forecast t+5	14.90	1.49	14.00	22.00	14.00	16.00
	Forecast t+10	15.13	2.04	14.00	25.00	14.00	16.00
	Forecast error current	0.09	0.49	-2.00	3.00	0.00	0.00
	Forecast error t+2	-0.01	0.68	-2.00	3.00	0.00	0.00
	Forecast error t+5	-0.12	1.10	-4.00	4.00	-1.00	1.00
	Forecast error t+10	-0.38	1.52	-7.00	3.00	-1.00	1.00
gemini-1.5-pro	Price	13.43	2.02	7.00	20.00	13.00	14.50
	Forecast current	13.68	1.91	6.00	21.00	13.00	15.00
	Forecast t+2	13.87	1.77	7.00	22.00	14.00	15.00
	Forecast t+5	14.26	1.48	8.00	23.00	14.00	15.00
	Forecast t+10	14.55	1.11	9.00	22.00	14.00	15.00
	Forecast error current	0.14	0.88	-5.00	6.00	0.00	0.50
	Forecast error t+2	-0.12	0.84	-4.00	4.00	-1.00	0.00
	Forecast error t+5	-0.64	1.36	-5.00	5.00	-1.00	0.00
	Forecast error t+10	-1.23	1.69	-6.00	6.00	-2.00	0.00
gpt-3.5	Price	16.48	0.92	14.00	20.00	16.00	17.00
	Forecast current	17.98	2.15	14.00	35.00	17.00	18.00
	Forecast t+2	19.92	2.06	14.00	38.00	19.00	21.00
	Forecast t+5	22.69	2.42	15.00	45.00	21.00	24.00
	Forecast t+10	25.28	2.76	17.00	55.00	24.00	27.00
	Forecast error current	-1.54	1.92	-16.00	2.00	-2.00	-0.50
	Forecast error t+2	-3.42	1.73	-18.00	0.50	-4.00	-2.50
	Forecast error t+5	-6.12	2.24	-26.00	1.50	-7.00	-5.00
	Forecast error t+10	-8.64	2.68	-39.00	0.00	-10.00	-7.00
gpt-4o	Price	14.94	0.87	14.00	20.00	15.00	15.00
	Forecast current	14.95	0.91	14.00	21.00	15.00	15.00
	Forecast t+2	15.20	1.10	13.00	22.00	15.00	15.00
	Forecast t+5	15.98	1.45	14.00	25.00	15.00	16.00
	Forecast t+10	17.06	2.10	14.00	28.00	16.00	18.00
	Forecast error current	0.04	0.41	-2.00	2.50	0.00	0.00
	Forecast error t+2	-0.19	0.46	-2.00	2.00	0.00	0.00
	Forecast error t+5	-0.94	0.94	-6.00	1.00	-1.00	0.00
	Forecast error t+10	-2.05	1.68	-11.50	2.00	-3.00	-1.00
grok-2	Price	18.17	2.13	14.00	22.00	16.00	20.00
	Forecast current	18.15	2.45	14.00	23.00	16.00	20.00
	Forecast t+2	18.39	2.64	14.00	24.00	16.00	20.00
	Forecast t+5	18.90	2.76	14.00	26.00	17.00	21.00
	Forecast t+10	19.07	3.02	14.00	28.00	17.00	21.75
	Forecast error current	-0.31	1.37	-3.00	6.00	-1.00	0.00
	Forecast error t+2	-0.31	1.49	-3.00	6.00	-1.50	0.00
	Forecast error t+5	-0.42	1.67	-5.00	5.00	-2.00	1.00
	Forecast error t+10	0.00	2.53	-7.50	7.00	-2.00	2.00
mistral-large	Price	15.17	2.09	12.00	20.00	14.00	16.00
	Forecast current	15.34	2.21	10.00	22.00	14.00	16.00
	Forecast t+2	15.77	2.45	9.00	23.00	14.00	17.00
	Forecast t+5	16.60	2.75	9.00	25.00	14.00	18.00
	Forecast t+10	17.76	3.23	14.00	27.00	15.00	19.00
	Forecast error current	0.00	1.04	-4.00	4.00	0.00	0.00
	Forecast error t+2	-0.36	0.95	-4.00	3.00	-1.00	0.00
	Forecast error t+5	-1.14	1.50	-9.00	4.00	-2.00	0.00
	Forecast error t+10	-2.38	2.96	-14.00	6.00	-3.00	0.00
Human	Price	31.31	16.42	12.00	102.00	19.00	40.00
	Forecast current	15.50	8.66	1.00	30.00	8.00	23.00
	Forecast t+2	16.50	8.08	3.00	30.00	9.75	23.25
	Forecast t+5	18.00	7.21	6.00	30.00	12.00	24.00
	Forecast t+10	20.50	5.77	11.00	30.00	15.75	25.25
	Forecast error current	1.67	8.70	-84.00	92.00	-1.00	2.00
	Forecast error t+2	1.69	11.12	-79.00	95.00	-3.00	6.00
	Forecast error t+5	-0.10	17.49	-149.00	96.00	-7.00	9.00
	Forecast error t+10	-0.53	25.44	-178.00	96.00	-12.00	15.00

2. Subjects fill in a short socio-demographic questionnaire where they provide information on their age, gender, study background, work experience, and level of English. The full survey is shown in the appendix.

3. Subjects watch a video with the experimental instructions on how to trade in the software. Participants were provided with written instructions and had to respond to 5 comprehension questions on the experimental software. If participants could not answer at least one question correctly, they were unable to participate. This was done to ensure that we only allowed participants who understood the instructions of the experiment. The full quiz is shown in the appendix.

4. Participants then logged in to the market to perform the actual trading experiment. In-lab participants will have their biometric data measured using either an Empatica E4 SCR device or a newer (acquired April 2024) Empatica EmbracePlus device.

5. Participants were then asked to respond to a post-experimental survey, where they were asked to qualitatively outline their strategy. The full survey is shown in the appendix.

Subjects received payments for their participation:

- A base payment of $20 and an average bonus of approx. $10
- Bonus payment is calculated based on: (i) participants' trading results at a rate of 200 experimental CASH units to $1, (ii) forecasting ability (i.e. each price forecast that is within 2.5 units of actual price is rewarded with 5 CASH), (iii) a randomly chosen lottery that participants select from the experiment and (iv) the quiz result (subject receive $0.25 for each of the 5 comprehension questions answered correctly.

Sample size: We collected data from 19 markets, with an average of 18 subjects/-market, resulting in a total of 342 participants. In each market session, we collected biometric data from three to five in-lab participants.

Instructions

Instructions were given by a video which can be viewed on Youtube. Below is the transcript of those instructions.

Welcome.

This experiment is about economic decision-making in an experimental stock market. The experiment will consist of a series of 30 trading periods in which you will have the opportunity to buy or sell shares of an asset that can yield payments in the future. Understanding these instructions may help you earn money and if you 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.

There are two assets in this experiment cash and stock you begin with 100 units of cash and four shares of stock. Stock is traded in a market each period among all the experimental subjects in units of cash. When you buy stock the price you agreed to pay is deducted from your amount of cash. When you sell stock the price you sold at is added to your amount of cash. The reward from holding stock is called dividend. Each period, every unit of stock earns a low or high dividend of either 0.4 cash per unit or 1.00 cash per unit with equal probability. These dividend payments are the same for everyone in all periods the dividend in each period does not depend on whether the previous dividend was low or high. The reward from holding cash is given by a fixed interest rate of 5% each period.

Price conversion at the end of the 30 periods of trading each unit of stock is automatically converted to 14 cash. For example if the market price for round 30 is 20 and you have three stocks, you'll receive 42 cash not 60 cash. Then your experimental cash units are converted to US dollars at a rate of 200 cash per 1 US dollar to determine how much you will be paid at the end of the experiment. If you buy shares for more than 14 as you get near Round 30 it is possible those shares will be terminated at a value of 14 if you cannot sell them.

Let's take an example. Suppose one period you have 120 units of cash and 5 units of stock. The dividend is 0.4 per unit of stock. Then your new cash amount is going to be the sum of the initial 120 cash, a 5% increase in cash, and total dividends of 5 times 0.4. Your total cash would therefore be 120 + 6 + 2 resulting in 128 cash. Notice that keeping cash will earn a return 5% per period and using cash to buy units of stock will also yield dividend earnings.

Each trading period will have four main screens that you will see one after the other. Order submission, forecast prices, round results, and choosing a lottery. This sequence is repeated 30 times. Let's look at each screen in a bit more detail.

On the order submission page in the upper right corner you will see the time
left for this page and the trading period. In this example the timer was set
on purpose for 5 minutes but in the actual trading task you'll only have 20 seconds. For this page also, the next button will not be displayed in the actual trading task so you cannot click it.

Below that you will see the order submission rectangle where you can buy or sell stocks. The price is displayed on the y-axis with the bold horizontal line being the previous market price. The quantity of shares you can buy is displayed on the x-axis. The vertical bold line represents the threshold between selling and buying. You can can select the price level at which you want to buy or sell using the mouse or cursor. For example, I can enter a buy order at 15 and submit two sell orders at 16. You can see the orders you entered in the lower right corner. You can cancel orders by clicking on the X next to them. Orders are not carried from one period to the next. If I want to buy at 16 and sell at 15 I will get an error and the sell order will not go through similarly. If I place a sell order at 15 and then I want
to submit a buy order at 16 the buy order will not go through. You can only buy stocks if you have enough cash, and you can only sell stocks you own. For example, if I want to sell five stocks I will get an error. Orders are limit orders. A limit order to buy one unit at a price of 15 means you would like to buy one unit at the price of 15 and at any lower price. Similarly a limit order to sell two units at 16 means you would like to sell
two units at the price of 16 and at any higher price.

A single market price is determined by matching the number of people who would like to buy at that price with the same number of people who would like to sell. The lower left part of the screen displays your personal stats
such as the number of shares you own, your current cash, and the total stock
value. For the market stats you can see the current market price, the interest rate, the dividends value, and the buyback value of the stock at the end of the 30 trading periods; which is 14. The upper left corner displays the graph of the stock market from the beginning until the current trading period. We can also see the traded volume for each period.

In the Forecast Prices screen the upper right part of the screen will change.
On this page, you will submit your best estimate of what you think the market
price will be for this period, two periods in advance, five periods in advance, and 10 periods in advance. As you can see the slider can be dragged
from left to right in this example I am setting the Price Forecast for this current trading period to 14. For two periods ahead I will set it to 14 as well. You can set the price forecast as you want. During the trading task, if the forecast is within 2.5 units from the actual realized price for each of the forecasted periods, then you will receive 2.5 units of cash as reward for each correct forecast. For example, if the actual price for period 1 is going to be 15 then you will be rewarded for that forecast. If the actual price for period 3 is 18 then you will not receive the reward for
that forecast.

On the round results screen you will see what the market price was for this period, how many shares were traded, as well as the number of shares you traded. You will also see updates on the new current cash and stock value and the dividend you earned. For period one, as you had four shares with a dividend of 0.4 for that period you won $1.60 from holding four stocks and $5 from holding $100 cash.

On the last page you will see two pie charts with different monetary rewards
and associated probabilities. You must select one of the pie charts. For example on the left side you have a lottery of winning either $7.25 with a probability of 74% or $9.00 with a probability of 26%. On the right side you
have a lottery of winning either $19.00 with 26% probability or $11.50 with 74% probability. After you select one of the lotteries a new set of lotteries
will appear. The next button will not appear in the actual trading task so you
cannot click on it. There will be four iterations of different lotteries with
different monetary rewards and probabilities. You will have 5 seconds for each
pair of lotteries, so you must be quick. At the end, one Lottery will be selected at random from those that you chose throughout the experiment, and you will receive the outcome of the lottery divided by 10. It’s in your

```
interest to actually choose accordingly.

Here are the key points. You will trade One stock for 30 trading periods
using cash. You start with 100 units of cash and four stocks. Each period
stock provides a dividend of either 0.4 or 1.00 while interest provides 5%
reward each trading period. The experiment consists of four pages, order
submission, forecasting prices, round results, and choosing a lottery.
    After
the last trading round all of your shares are converted to 14 cash units.
    If
you buy shares for more than 14, as you get near Round 30 it is possible
those shares will be terminated at a value of 14 if you cannot sell them.
You will also take a five question quiz. After this, pay attention to the
answers you give. You will earn a reward for each question you answer
correctly on the first try. If you fail to answer at least one question
correctly then you cannot continue.

The experiment will begin shortly.
```

4.7 Prompts

Additional Prompting Considerations

In adapting the task to LLM-agents, our primary goal was to imitate as best as possible the experiment as presented to human subjects. To this end, the description of the task setting mirrors the original instruction video, which details both the experiment premise as well as how to participate. Additional instructions are either specific to the text-based format of the market for LLM agents or general tips aimed at ensuring that agents understand the setup. They are told that their primary objective is to maximize total earnings. To reduce the necessary computation for the task, we make the following adjustments to the “screen" sequence for human subjects outlined in Appendix 4.A:

1. We combine the Order Submission and Forecast screens
2. We defer presenting round results until order entry of the following round

As no additional market data is revealed between the Order Submission and Forecast screens, modification (1) does not meaningfully change the experimental design. Similarly, modification (2) still provides updated market data before querying future market decisions, and so does not impact the setting. Thus, every round of the experiment, each agent simultaneously submits orders to the market as well as

expectations regarding current and future market prices. At each round, agents enter their market orders (if any) and price forecasts by filling in the provided JSON template.

We also include a “Practice Reflection" portion of the experiment, which invites agents to share specific insights and/or lessons from the practice rounds to be accessible for the remainder of the experiment. Such emulates the learning focus of the practice rounds for human subjects as well as the rhetorical questions posed after the practice rounds. Following the experiment, agents are also asked to complete a “Final Reflection" to distill down any key lessons learned through their participation.

Portfolio Data Format

```
   * Your Portfolio (Round XX):
      - Market price (Previous Round): XX
      - Buyback price: XX
      - # of shares owned: XX
      - Current cash: XX
      - Stock value: XX

%}
```

Standard Prompt Shell

```
You are a subject participating in a trading experiment. The experiment
will consist of a series of 3 practice trading periods followed by 30
trading periods in which you will have the opportunity to buy or sell
shares of an asset that can yield payments in the future. Understanding
instructions may help you earn money. If you make good decisions, you
might earn a considerable amount of money that will be paid at the end
of the experiment.

There are two assets in this experience: cash and stock. You begin with
```

100 units of cash and 4 shares of stock. Stock is traded in a market each period among all of the experimental subjects in units of cash. When you buy stock, the price you agreed to pay is deducted from your amount of cash. When you sell stock, the price you sold at is added to your amount of cash. The reward from holding stock is dividend. Each period, every unit of stock earns a low or high dividend of either 0.4 cash or 1.0 cash per unit with equal probability. These dividend payments are the same for everyone in all periods. The dividend in each period does not depend on whether the previous dividend was low or high. The reward from holding cash is given by a fixed interest rate of 5% each period. At the end of the 30 periods of trading, each unit of STOCK is automatically converted to 14 CASH. If the market price for round 30 is 20 and you have 3 stocks, youll receive 3x14=42 CASH, not 3x20=60 CASH. Then, your experimental CASH units are converted to US dollars at a rate of 200 CASH = $1 US, to determine how much the user will be paid at the end of the experiment. If you buy shares for more than 14 as you get near round 30, it is possible those shares will be terminated at a value of 14 if you cannot sell them.

Lets see an example. Suppose at the end of period 7, you have 120 units of CASH and 5 units of STOCK. The dividend that round is 0.40 per unit of stock. Your new cash amount for period 8 is going to be:

CASH = 120 + (120 x 5%) + (5 x 0.40)

= 120 + 6 + 2

= 128

Notice that keeping cash will earn a return of 5% per period and using cash to buy units of stock will also yield dividend earnings.

For each period, you will be provided with past market and portfolio history (prices, volumes, your filled orders) and you will simultaneously complete the following two tasks:

[ORDER SUBMISSION]:
In addition to past market and portfolio history, you will be provided with:

[# of Shares]: Number of shares of STOCK that you currently own. Each share that you own pays out a dividend at the end of each round. You CANNOT attempt to sell more shares than you own.

[Current Cash]: The amount of CASH that you currently have. Your CASH earns interest that is paid out at the end of each period. You CANNOT attempt to buy shares worth more than the cash you have.

[STOCK Value]: The value of your STOCK at the market price of the last round of play

[Market Price]: This is market clearing price from the last round of play

Using this information, you will submit orders to the market. All orders will be limit orders. For example, a limit order to BUY 1 STOCK @ 15 means that you would like to buy a STOCK at any price of 15 or less. Keep in mind the following points:

- Orders are NOT carried between periods
- SELL order prices must be greater than all BUY order prices + BUY order prices must be less than all SELL order prices
- You can only sell STOCK that you own and purchase STOCK with CASH you already have
- You are not required to submit orders every round and you may submit multiple orders each round
- Depending on market conditions, you may need to cross the spread to get fills on buy/sell orders

[PRICE FORECASTING]:

You will be asked to submit your predictions for the market price this period, two periods in advance, 5 periods in advance, and 10 periods in advance. In addition to past market and portfolio history, you will be provided with the range in which your prediction should fall. Your

prediction should be a non-negative, integer value. If your forecast is within 2.5 units of the actual price for each of the forecasted periods, then you will receive 5 units of cash at the end of the experiment as reward for each correct forecast.

For example, if you forecast the market price of period 1 to be 14 and the actual price is 15, then you will be rewarded for your forecast. However, if the actual price is 18, then you will not receive the reward. Additionally, during the experiment, you will complete PRACTICE REFLECTION and EXPERIMENT REFLECTION:

[PRACTICE REFLECTION]:

After completing the practice rounds, you will be asked to reflect on your practice experience. This reflection will be accessible to future versions of yourself during the main experiment. This can be helpful in passing along lessons learned to future versions of yourself.

[EXPERIMENT REFLECTION]:

At the end of the experiment, you will be asked to reflect on your experience, including any insight and/or strategies that you may have developed.

To summarize, here are the key points:

- You will trade one STOCK for 30 trading periods using CASH

- You start with 100 units of CASH and 4 STOCKS
- Each period, STOCK provides a dividend of either 0.4 or 1.0, while interest provides 5% reward
- You will complete each of the aforementioned tasks
- After the last trading round (30), all of your shares are converted to 14 CASH each. If you buy shares for more than 14 as you get near round 30, it is possible those shares will be terminated at 14 if you cannot sell them. You will keep any CASH you have at the end of the experiment.
- You are trading against other subjects in the experiment who may be susceptible to the same influences as you and may not always make optimal decisions. They, however, are also trying to maximize their earnings.
- Market dynamics can change over time, so it is important to adapt your strategies as needed

You will now complete the ORDER SUBMISSION + PRICE FORECASTING task.
Now let me tell you about the resources you have for this task. First, here are some files that you wrote the last time I came to you with a task. Here is a high-level description of what these files contain:

- PLANS.txt: File where you can write your plans for what strategies to test/use during the next few rounds.
- INSIGHTS.txt: File where you can write down any insights

you have regarding your strategies. Be detailed and precise

but keep things succinct and don't repeat yourself.

These files are passed between stages and rounds so try to focus on general strategies/insights as opposed to only something stage-specific.

Now, I will show you the current content of these files.

Filename: PLANS.txt

++++++++++++++++++++

<PLANS>

++++++++++++++++++++

Filename: INSIGHTS.txt

++++++++++++++++++++

<INSIGHTS>

++++++++++++++++++++

Here is the game history that you have access to:

Here is your practice round reflection:

Filename: PRACTICE REFLECTION (read-only)

++++++++++++++++++++

<PRACTICE REFLECTION>

++++++++++++++++++++

Filename: MARKET DATA (read-only)

++++++++++++++++++++

<MARKET DATA>

++++++++++++++++++++

Here is your current portfolio information:

Filename: CURRENT PORTFOLIO (read-only)

++++++++++++++++++++

<CURRENT PORTFOLIO>

++++++++++++++++++++

PRACTICE ROUND HISTORY/REFLECTION SHOULD ONLY BE USED TO LEARN THE EXPERIMENT SETTING AND MAY NOT REFLECT MARKET CONDITIONS IN THE MAIN EXPERIMENT.

Here is some key information to consider during your price forecasting:

- Make sure to submit a forecast within the specified range for each forecast input
- Use your previous history access to make informed decisions
- Remember that accurate (within 2.5 units) forecasts will earn you a reward at the end of the experiment

Here is some key information to consider during your order submission:

- You can only sell STOCK that you own and purchase STOCK with CASH you already have
- You are not required to submit orders every round and you may submit

```
  multiple orders each round for one or both sides
- Limit prices this round MUST be integer values between <MIN_LIMIT_PRICE>
  and <MAX_LIMIT_PRICE>. It is important that they are integer values
  within this range
- Make use of the provided history and your own strategies to make informed
  decisions
- Market dynamics can change over time, and so it might be necessary to
  adapt your strategies as needed
- Depending on market conditions, you may need to be aggressive or
  conservative in your trading strategies to maximize your earnings

Now you have all the necessary information to complete the task. Remember
YOUR TOP PRIORITY is to maximize your total earnings at the END of the 30
main experiment rounds. You have <# ROUNDS LEFT> rounds remaining.
First, carefully read through the information provided. Now, fill in the
below JSON template to respond. YOU MUST respond in this exact JSON format.

{
   "observations_and_thoughts": "<fill in here>",
   "new_content": {
      "PLANS.txt": "<fill in here>",
      "INSIGHTS.txt": "<fill in here>"
      "price_forecasts": [
```

```
{
    "round": <CURRENT ROUND>,
    "min_value": 0,
    "max_value": <FORECAST UPPER BOUND>,
    "forecasted_price": "<fill in here>"
},
{
    "round": <CURRENT ROUND + 2>,
    "min_value": 0,
    "max_value": <FORECAST UPPER BOUND>,
    "forecasted_price": "<fill in here>"
},
{
    "round": <CURRENT ROUND + 5>,
    "min_value": 0,
    "max_value": <FORECAST UPPER BOUND>,
    "forecasted_price": "<fill in here>"
},
{
    "round": <CURRENT ROUND + 10>,
    "min_value": 0,
```

```
            "max_value": <FORECAST UPPER BOUND>,
            "forecasted_price": "<fill in here>"
        }       ]
    },
    "submitted_orders": [
        {
            "order_type": "<BUY or SELL>",
            "quantity": <# of STOCK units>,
            "limit_price": <LIMIT_PRICE>
        },
        {
            "order_type": "<BUY or SELL>",
            "quantity": <# of STOCK units>,
            "limit_price": <LIMIT_PRICE>
        }
        // Add more or less orders as needed
    ]
}
```

Market Data Format

```
    Round XX:
        * Market + Portfolio Data:
            - Market price: XX
            - Market volume: XX
            - # of shares owned: XX
```

```
    - Current cash: XX
    - Stock value: XX
    - Dividend earned: XX
    - Interest earned: XX
    - Submitted orders:

        * <BUY/SELL> XX shares at XX per share

        * <BUY/SELL> XX shares at XX per share

    - Executed trades:

        -* No executed trades

* Forecasts:
    - Round XX price forecast: XX
    - Round (XX + 2) price forecast: XX
    - Round (XX + 5) price forecast: XX
    - Round (XX + 10) price forecast: XX
```

Fundamental Value Shock Treatment Prompt

For the two fundamental value shock treatments, we retain the same prompts, with the exception that from Round 15 onwards, the possible dividend outcomes as well as the redemption value are adjusted. Models are not informed of this change prior to Round 15 and thus, this operates as a "shock" to all participants.

Mechanically we inject the following prompt after the round recap screen is shown before Round 15.

```
[News Alert]: The company has recently announced it will now
    [halved/doubled]
all dividends to [0.2/0.5 OR 0.8/2.0]. The asset redemption value has now
[halved/doubled] to [$7.0 OR $14.0].
```

References

Abdelsamie, M. and H. Wang (Dec. 2024). "Comparative analysis of LLM-based market prediction and human expertise with sentiment analysis and machine learning integration". In: *2024 7th International Conference on Data Science and Information Technology (DSIT)*. IEEE, pp. 1–6.

AI, Mistral (2024). *Mistral Large (24.11)*. https://docs.mistral.ai/getting-started/models/models_overview/. Accessed: 2025-05-11.

Akata, Elif et al. (2023). "Playing Repeated Games with Large Language Models". In: *arXiv preprint arXiv:2305.16867*. URL: https://arxiv.org/abs/2305.16867.

Anlló, Hernán et al. (2024). "Do Large Language Models Reason Like Us?" In: *Communications Psychology*. URL: https://communities.springernature.com/posts/do-large-language-models-reason-the-way-we-do.

Anthropic (2024). *Introducing Claude 3.5 Sonnet*. https://www.anthropic.com/news/claude-3-5-sonnet. Accessed: 2025-05-11.

Bhat, R. and B. Jain (June 2024). "Stock Price Trend Prediction Using Emotion Analysis of Financial Headlines with Distilled LLM Model". In: *Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments*, pp. 67–73.

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

Bond, Shaun A. (2023). "Large Language Models and Financial Market Sentiment". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.4584928. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4584928.

Bostian, AJ, Jacob Goeree, and Charles A Holt (2005). "Price bubbles in asset market experiments with a flat fundamental value". In: *Draft for the Experimental Finance Conference, Federal Reserve Bank of Atlanta September*. Vol. 23, p. 2005.

Bostian, Andrew A. and Charles A. Holt (2009). "Price bubbles with discounting: A web-based classroom experiment". In: *Journal of Economic Education* 40.1, pp. 27–37. DOI: 10.3200/JECE.40.1.27-37.

Brookins, Philip and Jason DeBacker (2024). "Playing Games with GPT: What Can We Learn About a Large Language Model from Canonical Strategic Games?" In: *Economics Bulletin* 44.1, pp. 25–37. URL: http://www.accessecon.com/Pubs/EB/2024/Volume44/EB-24-V44-I1-P3.pdf.

Chen, Daniel L., Martin Schonger, and Chris Wickens (2016). "oTree—An open-source platform for laboratory, online, and field experiments". In: *Journal of Behavioral and Experimental Finance* 9, pp. 88–97. ISSN: 2214-6350. DOI: https://doi.org/10.1016/j.jbef.2015.12.001. URL: https://www.sciencedirect.com/science/article/pii/S2214635016000101.

Chen, Q. (2025). "Stock Price Prediction with LLM-Guided Market Movement Signals and Transformer Model". In: *FinTech and Sustainable Innovation*, pp. 1–7.

Cheng, J. and P. Chin (2024). "Empirical Asset Pricing with Large Language Model Agents". In: *arXiv preprint arXiv:2409.17266*.

Cobbe, Karl et al. (2021). "Training Verifiers to Solve Math Word Problems". In: *arXiv preprint arXiv:2110.14168*. URL: https://arxiv.org/abs/2110.14168.

Cui, Ziyan, Ning Li, and Huaikang Zhou (Aug. 25, 2024). *Can AI Replace Human Subjects? A Large-Scale Replication of Psychological Experiments with LLMs*. Rochester, NY. DOI: 10.2139/ssrn.4940173. URL: https://papers.ssrn.com/abstract=4940173 (visited on 09/25/2025).

DeepMind, Google (2024). "Gemini 1.5: Unlocking multimodal understanding across millions of tokens". In: *arXiv preprint arXiv:2403.05530*. URL: https://arxiv.org/abs/2403.05530.

Ding, Q. et al. (2024). "TradExpert: Revolutionizing Trading with Mixture of Expert LLMs". In: *arXiv preprint arXiv:2411.00782*.

Fatouros, George et al. (Dec. 2024). "Can Large Language Models beat wall street? Evaluating GPT-4's impact on financial decision-making with MarketSenseAI". In: *Neural Computing and Applications*. ISSN: 1433-3058. DOI: 10.1007/s00521-024-10613-4. URL: http://dx.doi.org/10.1007/s00521-024-10613-4.

Fish, Sara, Yannai A. Gonczarowski, and Ran I. Shorrer (2024). *Algorithmic Collusion by Large Language Models*. arXiv: 2404.00806 [econ.GN]. URL: https://arxiv.org/abs/2404.00806.

Hendrycks, Dan et al. (2021). "Measuring Massive Multitask Language Understanding". In: *arXiv preprint arXiv:2009.03300*. URL: https://arxiv.org/abs/2009.03300.

Holt, Charles A., Megan Porzio, and Michelle Y. Song (2017). "Price Bubbles, Gender, and Expectations in Experimental Asset Markets". In: *European Economic Review* 100, pp. 72–94. DOI: 10.1016/j.euroecorev.2017.05.005. URL: https://doi.org/10.1016/j.euroecorev.2017.05.005.

Horton, John J. (2023). *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?* Working Paper. DOI: 10.3386/w31122. URL: https://www.nber.org/papers/w31122.

Jia, J. J. et al. (2024). "Decision-making behavior evaluation framework for LLMs under uncertain context". In: *Advances in Neural Information Processing Systems*. Vol. 37, pp. 113360–113382.

Kahneman, Daniel and Amos Tversky (2013). "Prospect theory: An analysis of decision under risk". In: *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, pp. 99–127.

Kang, Y. et al. (2024). "Can Large Language Models Effectively Process and Execute Financial Trading Instructions?" In: *arXiv preprint arXiv:2412.04856*.

Kim, B. (2025). "Large Language Model Driven Technical Analysis: Enhancing Predictive Accuracy and Interpretability in Stock Trading". In: *International Journal of Internet, Broadcasting and Communication* 17.1, pp. 205–215.

Kirtac, Kemal and Guido Germano (2024). "Sentiment Trading with Large Language Models". In: *arXiv preprint arXiv:2412.19245*. URL: https://arxiv.org/abs/2412.19245.

Kuruvilla, J. S. and M. Mythily (2025). "Financial LLM for Stock Price Analysis and Investment Recommendation". In: *Journal of Telematics and Informatics* 13.1.

Leng, Yan (2024). "Can LLMs Mimic Human-Like Mental Accounting and Behavioral Biases?" In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.4705130. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4705130.

Li, X. et al. (May 2025). "HedgeAgents: A Balanced-Aware Multi-Agent Financial Trading System". In: *Companion Proceedings of the ACM on Web Conference 2025*, pp. 296–305.

Li, Y. et al. (2024). "A Reflective LLM–Based Agent to Guide Zero–Shot Cryptocurrency Trading". In: *arXiv preprint arXiv:2407.09546*.

Liu, K. M. and M. C. Lo (2025). "LLM-Based Routing in Mixture of Experts: A Novel Framework for Trading". In: *arXiv preprint arXiv:2501.09636*.

OpenAI (2025). *OpenAI API*. Accessed: 2025-05-11. URL: https://openai.com/api/.

Oprea, S. V. and A. Bâra (2024). "A Multi-Agent System Based on LLM for Trading Financial Assets". In: *Ovidius University Annals, Series Economic Sciences* 24.2.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan (2002). "Thumbs up? Sentiment classification using machine learning techniques". In: *arXiv cs/0205070*.

Řehůřek, Radim and Petr Sojka (May 2010). "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, pp. 45–50.

Saqur, R. (2024). "What Teaches Robots to Walk, Teaches Them to Trade too–Regime Adaptive Execution using Informed Data and LLMs". In: *arXiv preprint arXiv:2406.15508*.

Shi, J. and B. Hollifield (2024). “Predictive Power of LLMs in Financial Markets”. In: *arXiv preprint arXiv:2411.16569*.

Siddique, M. T. et al. (2025). “Enhancing Automated Trading with Sentiment Analysis: Leveraging Large Language Models for Stock Market Predictions”. In: *The American Journal of Engineering and Technology* 7.03, pp. 185–195.

Smith, Alec et al. (2014). “Irrational exuberance and neural crash warning signals during endogenous experimental market bubbles”. In: *Proceedings of the National Academy of Sciences* 111.29, pp. 10503–10508. DOI: 10.1073/pnas.1318416111. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1318416111. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1318416111.

Smith, Vernon L., Gerry L. Suchanek, and Arlington W. Williams (1988). “Bubbles, crashes, and endogenous expectations in experimental spot asset markets”. In: *Econometrica* 56.5, pp. 1119–1151. DOI: 10.2307/1911361.

Sun, Youfa et al. (Jan. 2010). “Algorithm of Call Auction Price”. In: *Applied Mechanics and Materials* 20-23. DOI: 10.4028/www.scientific.net/AMM.20-23.981.

Thaler, Richard (1985). “Mental accounting and consumer choice”. In: *Marketing science* 4.3, pp. 199–214.

Wei, Jason, Xuezhi Wang, et al. (2022). “Chain of thought prompting elicits reasoning in large language models”. In: *arXiv preprint arXiv:2201.11903*.

xAI (2024). *Grok-2 Beta Release*. https://x.ai/blog/grok-2. Accessed: 2025-05-11.

Xiao, Y. et al. (2024). “TradingAgents: Multi-agents LLM Financial Trading Framework”. In: *arXiv preprint arXiv:2412.20138*.

Yu, X. et al. (2023). “Temporal Data Meets LLM–Explainable Financial Time Series Forecasting”. In: *arXiv preprint arXiv:2306.11025*.

Yu, Y., H. Li, et al. (May 2024). “FinMem: A Performance-Enhanced LLM Trading Agent with Layered Memory and Character Design”. In: *Proceedings of the AAAI Symposium Series*. Vol. 3. 1, pp. 595–597.

Yu, Y., Z. Yao, et al. (2024). “FinCon: A Synthesized LLM Multi-Agent System with Conceptual Verbal Reinforcement for Enhanced Financial Decision Making”. In: *Advances in Neural Information Processing Systems* 37, pp. 137010–137045.

Yun, S. (2024). “Pretrained LLM Adapted with LoRA as a Decision Transformer for Offline RL in Quantitative Trading”. In: *arXiv preprint arXiv:2411.17900*.

Zhang, C. et al. (2024). “When AI Meets Finance (StockAgent): Large Language Model-Based Stock Trading in Simulated Real-World Environments”. In: *arXiv preprint arXiv:2407.18957*.

Zhang, W. et al. (Aug. 2024). "A Multimodal Foundation Agent for Financial Trading: Tool-Augmented, Diversified, and Generalist". In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4314–4325.

Zhang, Xiao-Yang et al. (2023). "Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of Large Language Models". In: *arXiv preprint arXiv:2306.12659*. URL: https://arxiv.org/abs/2306.12659.

Zhao, H. et al. (2024). "Revolutionizing Finance with LLMs: An Overview of Applications and Insights". In: *arXiv preprint arXiv:2401.11641*.

Zheng, Lianmin et al. (2023). "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena". In: *arXiv preprint arXiv:2306.05685*. URL: https://arxiv.org/abs/2306.05685.

Zhou, L. et al. (2025). "LLM-Augmented Linear Transformer–CNN for Enhanced Stock Price Prediction". In: *Mathematics* 13.3, p. 487.

Chapter 5

NEURAL RESPONSES DURING DETECTION OF REAL VS AI-GENERATED FACES

Thomas Henning, Colin Camerer, and Antonio Rangel

5.1 Introduction

A major social question emerging as AI-generated material becomes increasingly lifelike, whether in image generation models or other sensory modalities, is how well people can distinguish AI-generated content from reality. In the Image generation domain, algorithms have significantly improved their ability to generate realistic images and have also democratized their use. Now, many humans find AI images to be indistinguishable from reality, and image generation tools are being weaponized, leading to rising concerns for an upcoming crisis in social disinformation, including the phenomenon of "fake news" (U.S. Department of Homeland Security, 2021). Simultaneously, the ability of "AI" algorithms to detect these images has been mixed at best, and "deep-fake detection" remains an open question in computer science (Martin-Rodriguez, Garcia-Mojon, and Fernandez-Barciela, 2023). In response, there is a need to design technologies that can help distinguish AI-generated and real images. In this study, we investigate the role of neural responses in the development of such technologies.

The fundamental breakthrough that led to the AI image explosion was the latent Diffusion Model and its combination with commercially available text-to-image tools. Photo-realistic AI image generation tools are not new; GAN models (Goodfellow et al., 2014) have been known to create realistic faces that are nearly indistinguishable from real photographs of humans (Nightingale and Farid, 2022). However, GANs, while capable image generators, are more difficult to train and struggle to generate a wide variety of images due to a concept known as mode collapse (Bhagyashree, Kushwaha, and Nandi, 2020). In contrast, Latent Diffusion models are significantly easier to train and provide a greater variety of outputs, enabling numerous companies to create large models that can generate almost any type of image imaginable, which are often difficult for humans and other AI models to identify.

In response to these risks, there is a need to design technologies that can help distinguish AI-generated and real images. And a need to design technologies that can help identify specific interpretable features that can be used to make these distinctions. In this study, we investigate the role of neural responses in the development of such technologies.

Previous neuroforecasting studies have shown that neuronal data can outperform human behavior in predicting real-world outcomes in certain domains (Knutson and Genevsky, 2018). Here, we build on this methodology to test whether fMRI BOLD responses to face images can help distinguish AI-generated images from real

images above and beyond an individual's behavioral judgments. Additionally, we also identify neural correlates with underlying features that both distinguish AI faces from real faces and are interpretable to humans. This presents an opportunity to train or teach individuals on which features are relevant in distinguishing between AI-generated images and real images.

Neuroforecasting combines neural biological measurements and other physiological measurements taken from a small group of subjects to predict the aggregate or collective outcomes for another, larger group. Neuroforecasting has shown initial success in predicting numerous different aggregate demand markets from movie box office sales (Boksem, Smidts, and Scholte, 2015), newspaper article views (Falk, Scholz, O'Donnell, et al., 2017), to smoking reduction from health campaigns (Schmälzle et al., 2020)).

There is evidence from perception literature that the brain can process information without conscious awareness, indicating that in some circumstances, the brain takes in information that the subject cannot react to behaviorally. The literature often refers to this concept as "processing without awareness or not consciously perceived (NCP) stimuli" (Mudrik and Deouell, 2022). Papers such as (Fahrenfort et al., 2017; Gaal, Lamme, and Ridderinkhof, 2014; Pitts, Martínez, and Hillyard, 2012) suggest that neural data may be more sensitive than behavioral tasks in some perception based tasks such as; processing of face identity without awareness (Quiroga, 2012), detection of man-made/natural real-life pictures of objects in scenes (Fang et al., 2016), differences in trustworthiness between faces that are manipulated to appear close or far way (Bryan, Perona, and Adolphs, 2012).

While relatively scarce (at least compared to the AI image generation literature), there is a substantial body of literature on detecting AI-generated images. Most of the literature involves training deep-learning and/or machine-learning models to automatically detect images (Rana et al., 2022). The belief being that more traditional statistical and ML models are unable to capture more complex patterns that are poorly understood or easily represented using simple features (Rafique et al., 2023). In their paper, they utilize a multi-level approach. They first use an Error Level Analysis to determine if the image has been modified, they then feed this image to a Convolutional Neural Network for deep feature extraction. The resultant feature vectors are then classified using Support Vector Machines and K-Nearest Neighbors models.

While there are numerous deep learning papers where the underlying features en-

abling detection are often uninterpretable, the literature exploring human-interpretable features that will allow detection is sparse. Because of this, there is little work on attempting to augment individual human detection capabilities, and also little research on understanding how and why humans are occasionally capable of distinguishing these images and why they are sometimes unable to do so.

5.2 Materials and Methods

Our study has two separate components: (1) an online behavioral experiment designed to get measures of subjects' difficulty distinguishing between real and AI-generated faces, and (2) an fMRI experiment designed to identify neuronal differences between AI real and AI-generated faces.

The purpose of the online behavioral experiment was to identify 200 images (100 AI-generated and 100 Real) for which subjects had difficulty distinguishing between real and AI-generated faces. We used this experiment to select two sets of images, each balanced with 50 real and 50 AI images: (1) an easy set, images that were classified correctly by roughly 70% (+/- 10%) of the subjects, (2) a hard set, images that were classified correctly by roughly 50% (+/- 10%) of the subjects).

Image Selection We collected 300 real face photographs from publicly available image websites (e.g., Pexels.com). We initially used broad search terms (e.g., "man smiling," "woman smiling," "woman frowning") and then iteratively refined queries to obtain better demographic and contextual balance with the AI-generated images (e.g., "Black middle-aged woman smiling"). We excluded images containing multiple faces, images with heavy post-processing (e.g., airbrushing or obvious Photoshop manipulation), and images in which the face was partially or fully obscured. We further excluded images that were highly atypical or obfuscated (e.g., face covered by food, masks, props), as well as images that were excessively dark or blurry. We aimed to assemble a diverse set of stimuli varying in facial expression, age, race, gender, background, and lighting conditions to capture a wide range of image domains.

We generated 300 AI face images using Midjourney v5.1 (https://www.midjourney.com). We first generated an initial pool of 1,000 synthetic human face images using text-only prompts, and then selected 300 images based on manual inspection to prioritize photorealistic images that would plausibly be mistaken for real photographs. Prompts were generated in two ways: (i) hand-written prompts, and (ii) prompts obtained by providing Midjourney with a real image and requesting a corresponding

text prompt, which was then reused as a text-only input to generate a distinct synthetic image. This approach helped preserve key characteristics of the real images (e.g., approximate age, background, race) and reduce imbalance between AI and real stimulus sets. We relied on this method because Midjourney v5.1 did not always produce consistently photorealistic human face images from naïve text prompts alone. Example prompts and corresponding images are provided in Figure 5.1. The full stimulus set is publicly available at https://github.com/TaoTeTom/AI_Detection.



Figure 5.1: A) Prompt → Image Prompt: "Frontal view of a real young dominican woman with long brown hair, wearing a formal suit, sitting inside an interviewing office" B) Real Image → Prompt → Image Prompt: smiling asian woman sitting at desk and writing on notebook, in the style of light teal and black, uniformly staged images, representational, contact printing, nonrepresentational. C) Real Image D) Real Image → Prompt → Image Prompt: a young man with long hair is standing by a fence, in the style of afro-caribbean influence, light indigo and dark amber, strong facial expression, organic formations, dinocore, quadratura, confessional.

All images were resized to 512×512 pixels and cropped such that the primary subject's face was centered. We did not apply global corrections for brightness, saturation, contrast, or hue, as such transformations introduced visible perceptual artifacts in some stimuli (e.g., unnatural skin tones and lip coloration), which could itself serve as a cue for classification. Instead, we quantified these image statistics for both sets and verified that their distributions substantially overlapped. We additionally compared saliency maps across image sets using the Saliency Attentive Model (Cornia et al., 2018), and assessed high-level visual similarity by generating image embeddings using DINOv2 (Oquab et al., 2023) and comparing embedding

distributions across real and AI-generated image sets.

Online Behavioral Experiment. Eighty English-speaking adults (18–65 years) were recruited via Prolific and gave informed consent under Caltech IRB approval (IR19-0200). Participants received a $10 base payment plus an accuracy-contingent performance bonus of up to $10 (bonus = $10 × accuracy rate). Eleven participants were excluded based on pre-specified criteria[1], leaving a final sample of $N = 69$ (40 female; mean ± SD age = 23.4 ± 4.8 years).

The online behavioral experiment was designed to identify face images that participants found difficult to distinguish as AI-generated or real. Using mean classification accuracy across online participants as our difficulty metric, we selected a final stimulus set of 200 images (100 AI-generated, 100 real) split into two subsets: an easy set (50 real, 50 AI; ~70% correct) and a hard set (50 real, 50 AI; ~50% correct, near chance).

After reading task instructions, participants completed a brief practice block (using faces generated by a different AI model) before beginning the main task. The main task consisted of a single run in which participants viewed all 600 images (300 AI-generated; 300 real), presented in random order. On each trial, participants viewed one face stimulus and were tasked with determining whether the face image was real or AI-generated using keyboard arrow keys (A sample task trial can be seen in Figure 5.2). Participants had up to 3 seconds to respond. If no response was made within 3 seconds, participants received a prompt instructing them to respond more quickly. No trial-by-trial accuracy feedback was provided. Participants were informed that the task included both real and AI-generated images, but were not informed that the stimulus set was balanced 50/50. Figure 5.3 shows the accuracy and reaction time distributions for the all the images shown in the online experiment separately for real and AI-generated stimuli.

[1]Participants were excluded from analysis if their response patterns suggested inattention or automated responding. Specifically, participants were excluded if (i) their median response time (RT) across trials was below 300 ms, (ii) their RT showed a sustained collapse indicative of disengagement (defined as a ≥50% decrease in median RT from the first third to the final third of the task), or (iii) they exhibited highly stereotyped responding, defined as selecting the same response ("AI" or "Real") for ≥50 consecutive trials. All exclusion thresholds were chosen to remove implausibly fast responding and highly non-informative response patterns while retaining typical human variability in response speed.

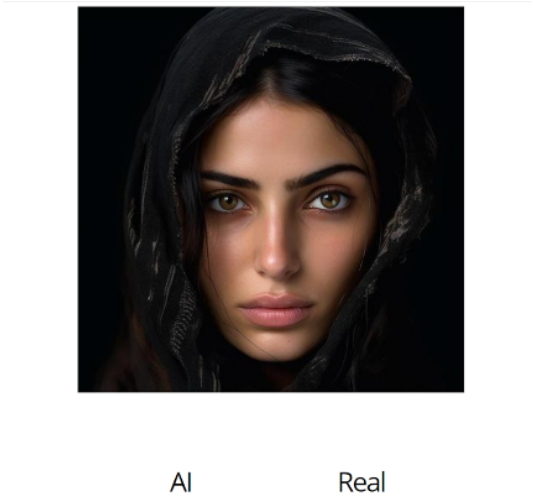


Figure 5.2: Sample task screen for the online behavioral task.

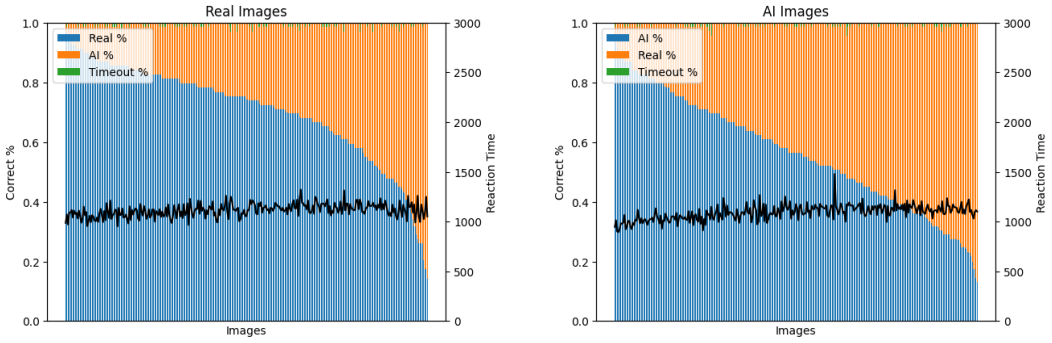


Figure 5.3: Accuracy and reaction time distributions for the images shown in the online task, shown separately for real and AI-generated stimuli.

To construct the stimulus set for the fMRI experiment, we used mean response accuracy from the online study as an index of stimulus difficulty. We selected two subsets of images. The *hard* set comprised 100 images (50 AI-generated and 50 real) that were correctly classified between 40% and 60% of the time (mean = 51.14%), corresponding to stimuli near chance performance. The *easy* set comprised 100 images (50 AI-generated and 50 real) that were correctly classified between 60% and 80% of the time (mean = 70.57%). Within each subset, we additionally matched the distributions of race, sex, and age across real and AI categories to reduce potential demographic confounds. Example stimuli are shown in Figure 5.4. The full stimulus set is publicly available at:
https://github.com/TaoTeTom/AI_Detection/tree/main/Images/Image_Bins.



Figure 5.4: Examples from the fMRI stimulus set. (A) Hard AI image, (B) Hard real image, (C) Easy AI image, and (D) Easy real image.

fMRI Task Fifty-three healthy subjects (31 F, 18–46 yr) were recruited and scanned for our data sample. Individuals were recruited from the Caltech and Pasadena City College communities. Subjects were excluded based on the following exclusion criteria: professional/recreational AI-ML or extensive photography experience, neurological/psychiatric disorders, psychoactive medication, and non-native English speaking.

Thirteen subjects were excluded from further analysis due to reasons such as experimenter error, incorrect task sequence given, and/or data corruption that caused fMRI-prep to fail to pre-process the data, leaving a final usable count of forty subjects for analysis. Caltech's Institutional Review Board approved all procedures. Participants were paid $75 + a performance bonus for participating.

The task consisted of two components: (i) an image classification task and (ii) a confidence survey administered between runs while participants remained in the scanner. The classification task was designed to elicit both behavioral responses and neural activity, allowing us to identify brain regions associated with discrimination of AI-generated versus real face images and to test whether neural signals predict trial-level behavioral performance. The confidence survey assessed participants' subjective confidence in their overall performance for each run. No fMRI volumes were acquired during the confidence survey.

Scanner Image Classification Task During scanning, participants completed two runs (easy and hard; order counterbalanced across participants). Each run consisted

of 100 face images (50 AI-generated and 50 real). Before each run, participants completed a brief 10-trial practice block using a different set of images to familiarize them with the button-response procedure; no feedback was provided. Trials were separated by a jittered intertrial interval (ITI) ranging from 4 to 9 s (mean = 6.5 s, SD = 1.33 s). Participants had up to 3 s to respond on each trial, the task structure is visualized in Figure 5.5.

Responses were recorded using two buttons on a four-button response box. The button response mapping (left = real; right = AI) was reversed between runs to counterbalance motor responses. If a participant responded before the 3-second response window elapsed, the stimulus was removed and the remaining response-window time was added to the subsequent fixation interval (ITI).

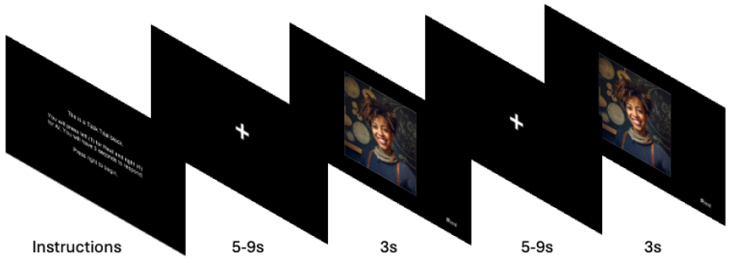


Figure 5.5: fMRI Task.

MRI acquisition and Pre-processing Images were acquired on a Siemens Prisma 3T with a 32-channel head coil: 72 oblique axial slices (30° tilt), TR = 1.12 s, TE = 30 ms, flip angle = 54°, multi-band 4, 2 mm isotropic voxels, IPAT 2, echo spacing = 0.56 ms. A high-resolution T1-weighted MPRAGE (1 mm isotropic) was also collected.

fMRI data were preprocessed using fMRIPrep 23.1.3 (Esteban et al., 2019). Pre-processing included slice-timing correction, motion correction, co-registration of functional images to each participant's anatomical T1-weighted image, and spatial normalization to MNI space. Post-fMRIPrep analyses were performed in Python using NiBabel and Nilearn. Temporal filtering and additional nuisance regression were applied during first-level modeling as described below.

fMRI Analysis We estimated several different first-order general linear models (GLMs) of the BOLD responses. Each GLM was estimated in three steps. First, we estimated the model for each individual separately. Second, we calculated contrast statistics at the individual level. Third, we computed second-level group statistics by

performing one-sample t tests on the single-subject contrast coefficients. A spatial smoothing isotropic Gaussian kernel of 6mm full width at half maximum (FWHM) was applied to all functional data before model estimation.

GLM 1: Real v AI Encoding. We used a Least Squares All (LSA) GLM framework to identify regions that contrastly encode neural responses to AI-generated versus Real images during face classification. The model included two regressors of interest:

- (R1): An indicator function for the duration of AI-image presentation;
- (R2): An indicator function for the duration of Real-image presentation;

Each regressor of interest was modeled from stimulus to response time. The model included three motor nuisance stick regressors (left-button press, right-button press, no-button press) of no interest, as well as six motion regressors of no interest (translation and rotation). The regressors of interest and the motor regressors were convolved with a canonical hemodynamic response function (HRF), the motion regressors were not convolved.

Several first-level contrasts were defined and submitted to group-level random effects analyses:

- (C1): Real - AI, identifying regions with stronger responses to Real images compared to AI-generated images
- (C2): AI - Real identifying regions with stronger responses to AI-generated images compared to Real images

We performed a whole-brain voxel-wise analysis contrasting BOLD responses to Real versus AI-generated faces. To identify clusters of significant activation, we applied a threshold of $p < 0.005$ and a cluster-extent threshold of 50 voxels. Cluster-wise inference was carried out using the resulting statistical maps from the Real > AI contrast.

GLM 2: Perceived AI vs. Perceived Real Classification Model. The model included three regressors of interest:

- (R1): An indicator function for the duration of trials in which the participant classified the image as AI-generated;

- (R2): An indicator function for the duration of trials in which the participant classified the image as Real;
- (R3): An indicator function for the duration of missed (timeout) trials in which no response was made;

Each regressor of interest was modeled from stimulus to response time. The model included three motor nuisance stick regressors (left-button press, right-button press, no-button press) of no interest, as well as six motion regressors of no interest (translation and rotation). The regressors of interest and the motor regressors were convolved with a canonical hemodynamic response function (HRF), the motion regressors were not convolved.

Several first-level contrasts were defined and submitted to group-level random effects analyses:

- (C1): Perceived AI – Perceived Real, identifying regions more active when participants believed an image was AI-generated;
- (C2): Perceived Real – Perceived AI, identifying regions more active when participants believed an image was Real;
- (C3): Missed – (Perceived AI + Perceived Real), isolating activation during nonresponse trials;

We performed a whole-brain voxel-wise analysis contrasting BOLD responses to Perceived Real versus Perceived AI-generated faces. To identify clusters of significant activation, we applied a threshold of $p < 0.005$ and a cluster-extent threshold of 50 voxels. Cluster-wise inference was carried out using the resulting statistical maps from the Perceived Real > Perceived AI contrast.

GLM 3: Correct vs. Incorrect Classification Model The model included two regressors of interest:

- (R1): An indicator function for the duration of correctly classified trials;
- (R2): An indicator function for the duration of incorrectly classified trials;
- (R3): An indicator function for the duration of unanswered trials;

Each regressor of interest was modeled from stimulus onset to response time. The model included three motor nuisance stick regressors (left-button press, right-button press, no-button press) of no interest, as well as six motion regressors of no interest (translation and rotation). The regressors of interest and the motor regressors were convolved with a canonical hemodynamic response function (HRF), the motion regressors were not convolved. A spatial smoothing isotropic Gaussian kernel of 6mm full width at half maximum (FWHM) was applied to all functional data before model estimation.

Several first-level contrasts were defined and submitted to group-level random effects analyses:

- (C1): Correct – Incorrect, identifying regions with stronger responses during correct trials
- (C2): Incorrect – Correct, identifying regions with stronger responses during incorrect trials
- (C3): Missed – (Correct + Incorrect), identifying regions associated with trials where no behavioral response was made

We performed a whole-brain voxel-wise analysis contrasting BOLD responses to Correct versus Incorrect trials. To identify clusters of significant activation, we applied a threshold of $p < 0.005$ and a cluster-extent threshold of 50 voxels. Cluster-wise inference was carried out using the resulting statistical maps from the Correct > Incorrect contrast.

5.3 Results

Figure 5.8 shows the accuracy and reaction time distributions for the all the images shown in the fmri experiment separately for real and AI-generated stimuli.

Accuracy. Table 5.1 and Figure 5.6 examine the accuracy of responses in the online task and show that we successfully identified an easy stimulus set whose difficulty replicated in the scanner. In the easy set, accuracy was nearly identical in the online cohort (70.54%) and the in-scanner cohort (70.40%), with no significant difference between cohorts ($p = 0.803$). All p-values in this section are Welch's t-test. Within both cohorts, accuracy did not differ between real and AI images (scanner: 70.20% vs. 70.60%, $p = 0.891$; online: 71.56% vs. 69.53%, $p = 0.086$), indicating no systematic bias in discriminability across image types.

The online experiment also yielded a hard stimulus set with near-chance performance, though accuracy was modestly higher in-scanner. In the hard set, in-scanner participants achieved 54.55% accuracy, while the independent online cohort achieved 51.18%. Accuracy differed significantly between the two cohorts ($p = 0.005$), though accuracy did not differ between real and AI images (scanner: 55.70% vs. 53.40%, $p = 0.468$; online: 51.21% vs. 51.14%, $p = 0.957$) in either cohorts. We felt comfortable with this difference since there was no systematic discriminability differences between image classes.

When collapsing across both stimulus sets, overall accuracy did not differ between in-scanner (62.47%) and online participants (60.86%; $p = 0.118$). Taken together, these results validate our stimulus selection procedure. The online pretest successfully identified two stimulus sets that produced the intended separation in difficulty in the scanner (easy: ∼70% correct; hard: near chance). Additionally accuracy did not differ between real and AI images within either set, indicating no systematic differences in discriminability across image types. Although as stated above, in-scanner participants performed modestly better than the online cohort for the hard set, overall performance did not differ between cohorts when collapsing across both sets.

Table 5.1: Behavioral Accuracy for Real and AI conditions.

Condition	Total (%)	Real (%)	AI (%)	Diff (pp)	p
Scanner Easy (70)	70.40 ± 2.88	70.20 ± 3.93	70.60 ± 4.35	-0.40	0.891
Scanner Hard (50)	54.55 ± 3.12	55.70 ± 4.34	53.40 ± 4.62	2.30	0.468
Scanner Total	62.47 ± 2.38	62.95 ± 3.22	62.00 ± 3.56	0.95	0.695
Online Easy (70)	70.54 ± 1.17	71.56 ± 1.63	69.53 ± 1.68	2.02	0.086
Online Hard (50)	51.18 ± 1.15	51.21 ± 1.60	51.14 ± 1.69	0.06	0.957
Online Total	60.86 ± 1.58	61.38 ± 2.32	60.34 ± 2.18	1.04	0.516

Values are mean ± 95% CI, p from Welch's t-test. p values represent difference between AI and Real values

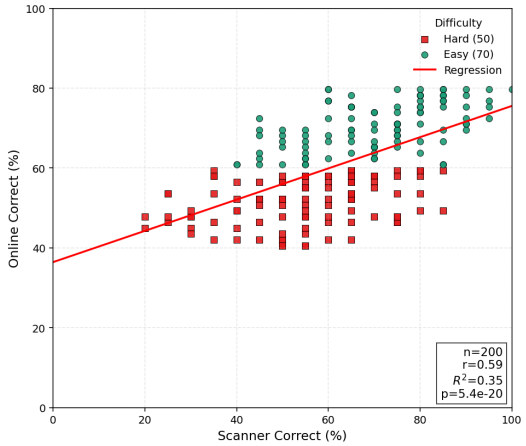


Figure 5.6: Online vs Scanner Behavioral Accuracy image comparison, online vs scanner. The variance is wider across scanner subjects because the online images were directly selected to fall within a set accuracy range.

Reaction Times. We also captured reaction times (RTs) in the online and fMRI tasks as shown in Table 5.2 and Figure 5.7. Responses were slower in the scanner than online (scanner: easy = 1524.43 ms, hard = 1576.25 ms; online: easy = 1106.93 ms, hard = 1133.06 ms; cohort effect: $p < 0.001$). In both cohorts, RTs were slower for hard than easy images (scanner: $p < 0.001$; online: $p < 0.001$). Finally, responses were faster for AI than real images in both cohorts (scanner: AI = 1504.90 ms vs. real = 1595.71 ms, $p < 0.001$; online: AI = 1101.62 ms vs. real = 1138.37 ms, $p < 0.001$). Overall, reaction time patterns corroborated the difficulty manipulation (hard > easy) and suggested a consistent response asymmetry; participants responded faster for AI than real images across cohorts. Figure 5.3 shows the accuracy and reaction time distributions for the all the images shown in the online experiment separately for real and AI-generated stimuli.

Table 5.2: Reaction times (ms) by condition.

Condition	Total	Real	AI	Diff	p
Scanner Easy (70)	1524.43 ± 25.92	1575.39 ± 31.86	1473.47 ± 36.62	101.92	< 0.001
Scanner Hard (50)	1576.25 ± 23.69	1616.03 ± 29.95	1536.48 ± 34.17	79.55	< 0.001
Scanner Total	1550.34 ± 17.78	1595.71 ± 21.86	1504.98 ± 25.39	90.73	< 0.001
Online Easy (70)	1106.93 ± 14.76	1140.86 ± 20.00	1073.00 ± 17.68	67.86	< 0.001
Online Hard (50)	1133.06 ± 13.16	1135.87 ± 16.21	1130.25 ± 21.30	5.62	0.674
Online Total	1120.00 ± 9.97	1138.37 ± 12.66	1101.62 ± 14.75	36.74	< 0.001

Values are mean ± 95% CI, p from Welch's t-test. p values represent the difference between AI and Real values.

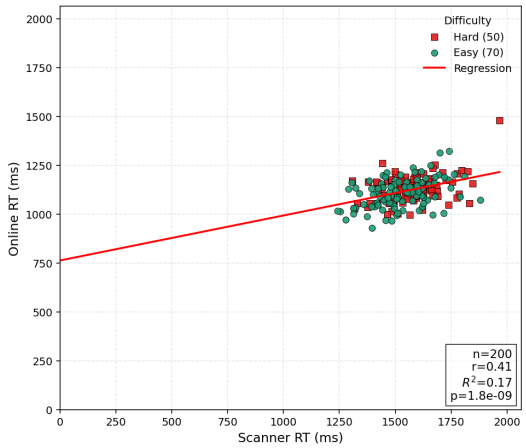


Figure 5.7: Online vs Scanner reaction time image comparison, online vs scanner. The variance is wider across scanner subjects because the online images were directly selected to fall within a set accuracy range.

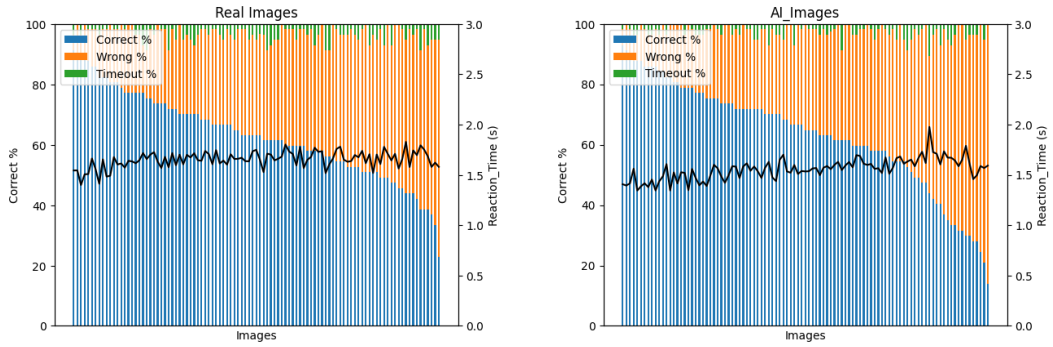


Figure 5.8: Accuracy and reaction time distributions for the images shown in the fMRI task, shown separately for real and AI-generated stimuli.

BOLD responses to Real vs AI-generated faces. GLM 1 allowed us to identify regions that exhibited different responses to real vs. AI-generated faces. Results are shown in Table 5.3 and Figure 5.9.

The contrast produced six surviving clusters with effects in both directions. Real > AI effects appeared in the intracalcarine cortex and an adjacent posterior occipital cluster, atlased as Left Cerebral Cortex. AI > Real effects appeared in lingual gyrus, precuneus, cingulate gyrus, and frontal pole. The pattern places the Real > AI effects in early visual cortex and the AI > Real effects in a more anatomically distributed set of medial parietal, cingulate, and anterior prefrontal regions.

The Real > AI effects in intracalcarine cortex and adjacent posterior occipital cortex fall within early retinotopic visual cortex, a region commonly associated with low-

level visual analysis including the encoding of contrast, edges, spatial frequency, and local image structure (Kravitz et al., 2013). A relevant consideration when interpreting this direction of effect is that diffusion-model image outputs have been shown in the computer-vision and image-forensics literature to differ systematically from natural photographs in low-level statistical properties. Diffusion-generated images have been reported to exhibit characteristic anomalies in their power spectra and autocorrelation structure, with smoother local texture and reduced high-frequency content relative to real photographs (Corvi, Cozzolino, Zingarini, et al., 2023; Corvi, Cozzolino, Poggi, et al., 2023). These differences are sufficiently consistent that they form the basis for a substantial line of work on detecting diffusion-generated images using frequency-domain features, color statistics, and patch-level low-level statistics (Wang et al., 2020; Ojha, Li, and Lee, 2024; Bammey, 2024; Lorenz, Durall, and Keuper, 2023; Sundar et al., 2024). Independent psychophysical work has also documented systematic differences in artifact structure and low-level image properties between diffusion outputs and real photographs that can be exploited for human discrimination (Lu, Groh, et al., 2025). Against this background, a stronger drive for early retinotopic cortex by real photographs than by diffusion-generated faces would be expected on the basis of low-level image statistics alone, independently of any face-specific processing or evaluative interpretation. The Real > AI direction in early visual cortex in the present contrast is therefore consistent with this stimulus-property account.

A lingual gyrus cluster appeared in the AI > Real direction. The lingual gyrus sits within ventral occipital and occipito-temporal cortex, where it has been implicated in visual recognition, object and scene processing, and the transformation of early visual features into higher-level perceptual representations (Kravitz et al., 2013). In the broader ventral-stream literature, regions in this territory are associated with category-level and mid-level visual representations and have been shown to respond differentially to image-statistical and category manipulations (Kravitz et al., 2013; Heydt, 2023). The lingual cluster here is also relatively medial and inferior, falling outside the canonical face-selective regions identified in the face-perception literature, namely the fusiform face area, occipital face area, and posterior superior temporal sulcus face region (Kanwisher, McDermott, and Chun, 1997; Haxby, Hoffman, and Gobbini, 2000; Kanwisher and Yovel, 2006). We do not attempt to attribute the present lingual effect to a specific perceptual mechanism but note that ventral occipito-temporal involvement is recurrent in literatures on differential responses to image categories that vary in visual properties.

We also saw increased activity in the cluster around the precuneus region for AI images > Real images. The precuneus has been implicated in visuospatial imagery, episodic memory retrieval, first-person perspective taking, self-referential processing, and internally directed cognition (Cavanna and Trimble, 2006). It is a major node of the default-mode network (Buckner, Andrews-Hanna, and Schacter, 2008; Andrews-Hanna, Smallwood, and Spreng, 2014; Utevsky, Smith, and Huettel, 2014) and has been characterized as a functional core of that network, which evidence suggests has significant involvement during externally directed tasks involving evaluation of complex stimuli, including aesthetic and self-relevant judgments and visuospatial planning (Vessel, Starr, and Rubin, 2012; Vatansever, D. K. Menon, and Stamatakis, 2017; Smallwood et al., 2021). Precuneus and adjacent medial parietal cortex have also been reported to modulate during processing of unfamiliar, ambiguous, or affectively relevant images and during reappraisal of emotional stimuli (Sheline et al., 2009). The appearance of a precuneus effect in the AI > Real direction is broadly consistent with this pattern of medial parietal involvement in evaluative and internally directed processing of complex visual stimuli.

We also see an AI > Real cluster in the Anterior and mid-cingulate regions which have been associated with attentional control, salience monitoring, conflict detection, error processing, and the integration of default-mode and control-network processing (Ridderinkhof et al., 2004; Botvinick, 2007; Shenhav, Botvinick, and Cohen, 2013; Leech and Sharp, 2014). In perceptual tasks involving uncertainty, ambiguity, or atypical stimuli, anterior cingulate involvement has frequently been reported in connection with monitoring and evaluative processes rather than with perceptual encoding per se (Ridderinkhof et al., 2004; Ullsperger, Danielmeier, and Jocham, 2014). The present contrast does not adjudicate among these interpretations, but the anatomical location of the cluster is consistent with regions repeatedly implicated in monitoring and evaluation across a range of perceptual decision tasks.

A frontal pole cluster appeared in the AI > Real direction. Frontal pole, also referred to as anterior prefrontal cortex or Brodmann area 10, has been associated with high-level cognitive control, abstract evaluation, prospective reasoning, multitasking, and the coordination of multiple goals or task sets (Ramnani and Owen, 2004; Koechlin and Hyafil, 2007; Burgess, Dumontheil, and Gilbert, 2007). Frontal-pole engagement has been reported during evaluation of unfamiliar or atypical stimuli, during integration of internally generated and externally presented information, and during tasks requiring metacognitive evaluation of one's own perceptual or

cognitive states (Burgess, Gilbert, and Dumontheil, 2007; Fleming and Dolan, 2012). As with the cingulate effect, the present design does not isolate evaluation, novelty processing, metacognition, or cognitive control as separable mechanisms, but the regional involvement is recurrent in the literature on higher-order evaluation of complex or unfamiliar stimuli.

It is also worth noting where these effects do not fall. The face-perception literature has identified a relatively consistent set of face-selective regions, with the fusiform face area in mid-fusiform cortex, the occipital face area in lateral occipital cortex, and a face-selective region in the posterior superior temporal sulcus, sometimes considered together with surrounding regions as a distributed face-processing network (Kanwisher, McDermott, and Chun, 1997; Haxby, Hoffman, and Gobbini, 2000; Kanwisher and Yovel, 2006; Grill-Spector et al., 2017). The principal AI > Real effects in the present contrast did not fall within these canonical face-selective regions. The pattern is therefore better characterized as a distinction between posterior visual responses in the Real > AI direction and a more distributed medial parietal, cingulate, and anterior prefrontal pattern in the AI > Real direction, rather than as differential recruitment of classical face-selective cortex. This is consistent with the possibility that the contrast indexes broad differences in low-level image statistics and higher-order evaluative processing rather than face-specific perceptual computation, although the present design cannot conclusively distinguish these accounts.

Taken together, the regions implicated by the Real vs. AI contrast span systems that have, in prior work, been associated with low-level visual encoding, ventral-stream visual recognition, default-mode and medial parietal evaluative processing, anterior cingulate monitoring and salience, and anterior prefrontal evaluation and control. Two observations from the prior literature are worth highlighting. First, generative-image and image-statistics work provides a parsimonious account of why early retinotopic cortex would respond more strongly to real photographs than to diffusion-generated faces, independent of any face-specific or evaluative interpretation. Second, the AI > Real effects fell outside canonical face-selective cortex and instead implicated regions more often associated with evaluative, monitoring, and internally directed processing in prior fMRI work.

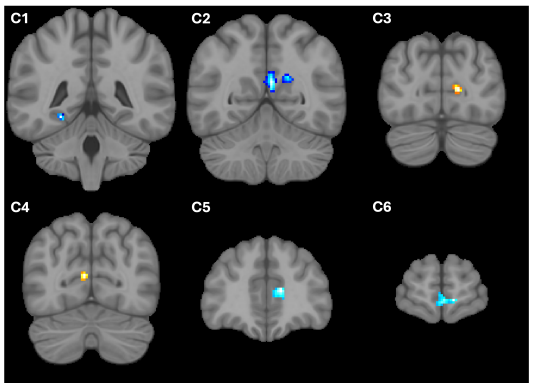


Figure 5.9: Cluster masks from GLM 1, thresholded at voxelwise $p < 0.005$ with cluster extent $k > 50$ voxels. Panels labeled C1–C6 correspond to cluster IDs in Table 5.3.

Table 5.3: Significant clusters from the Real vs. AI contrast (GLM 1). Threshold: p-value < 0.005, $k > 50$ voxels.

Cluster ID	X	Y	Z	Peak t	Cluster Size (mm^3)	ROI Region
1	-26	-44	-10	-4.68	432	Lingual Gyrus
2	4	-56	20	-4.60	1952	Precuneus Cortex
3	14	-80	8	4.07	528	Intracalcarine Cortex
4	-6	-72	14	3.78	448	Left Cerebral Cortex
5	12	42	10	-3.69	464	Cingulate Gyrus
6	10	64	-6	-3.68	464	Frontal Pole

Real > AI encoding contrast (sorted by absolute t value at the cluster level, with subpeaks kept attached to their parent cluster). ROIs were determined using the Harvard–Oxford atlas within a 4 mm radius of each peak coordinate; the majority label was chosen. Cluster size is shown only for the highest-$|t|$ peak within each cluster.

BOLD responses to Perceived Real vs Perceived AI-generated faces. GLM 2 allowed us to identify regions that exhibited different responses to faces that subjects perceived to be real vs. images subjects perceived to be AI-generated. Results are shown in Table 5.4 and Figure 5.10.

The contrast produced eight surviving clusters, with the majority in the Perceived Real > Perceived AI direction and a single cluster in the Perceived AI > Perceived Real direction. The Perceived Real > Perceived AI effects fell predominantly in occipital, ventral occipito-temporal, sensorimotor, and medial motor-related regions, with an additional cerebellar cluster. The Perceived AI > Perceived Real effect was localized to anterior prefrontal cortex.

The intracalcarine and occipital pole clusters are situated within early retinotopic visual cortex and adjacent striate and peristriate regions, where activity is commonly

associated with low-level visual encoding including contrast, spatial frequency, edges, and retinotopic spatial structure (Kravitz et al., 2013). Their appearance in the Perceived Real > Perceived AI direction parallels the early visual effects observed in the objective Real vs. AI contrast. As discussed for GLM 1, a parsimonious account of stronger early visual cortex responses to images judged as real is provided by the image-statistics literature, which has documented systematic differences in low-level statistical properties between diffusion-generated and natural photographs, including differences in power spectra, autocorrelation structure, smoother local texture, and reduced high-frequency content (Corvi, Cozzolino, Zingarini, et al., 2023; Corvi, Cozzolino, Poggi, et al., 2023; Lu, Groh, et al., 2025). These low-level differences form the basis for a substantial computer-vision literature on detecting diffusion outputs using frequency-domain, color, and patch-level statistics (Wang et al., 2020; Ojha, Li, and Lee, 2024; Bammey, 2024; Lorenz, Durall, and Keuper, 2023; Sundar et al., 2024). To the extent that these statistical regularities also influence subjective judgments of realness, an early visual cortex effect in the perceived-realness contrast would be expected even in the absence of any explicit AI-detection process.

Effects in inferior temporal and fusiform cortex fall within the ventral occipito-temporal stream, where activity has been associated with higher-level object and face processing (Kravitz et al., 2013; Heydt, 2023). Inferior temporal cortex has long been characterized as a high-level node of the ventral stream involved in the representation of objects, places, faces, and other categories of visual stimuli (Kravitz et al., 2013; Martin and Barense, 2023). The fusiform gyrus, and in particular the mid-fusiform region, contains the fusiform face area and adjacent face-selective patches that have been extensively characterized in the face-perception literature (Kanwisher, McDermott, and Chun, 1997; Kanwisher and Yovel, 2006; Grill-Spector et al., 2017). Face-selective regions extending from inferior occipital gyrus to anterior temporal lobe have been proposed to constitute a hierarchically organized face-processing network (Haxby, Hoffman, and Gobbini, 2000; Collins and Olson, 2014; Iidaka, 2014). Although the present design does not include a face localizer and we therefore cannot establish that these clusters fall specifically within face-selective subregions, their anatomical location is consistent with ventral-stream territory commonly implicated in face processing. This is a point of divergence from the objective Real vs. AI contrast, which did not yield significant effects in inferior temporal or fusiform regions. One possibility is that subjective realness judgments tracked features more closely related to ventral-stream face representations than the objective image category alone, although the present design cannot distinguish

this from alternative accounts such as differences in attention, decision criterion, or stimulus-level confounds.

Effects in postcentral gyrus and juxtapositional lobule cortex fall within sensorimotor and medial motor-related regions. Postcentral gyrus is most commonly associated with primary somatosensory cortex (Gomez-Ramirez, Hysaj, and Niebur, 2016). The juxtapositional lobule cortex, in Harvard–Oxford terminology, encompasses the supplementary motor area and pre-supplementary motor area on the medial wall, regions broadly implicated in motor planning, action preparation, response selection, and the timing of voluntary action (Picard and Strick, 1996; Nachev, Kennard, and Husain, 2008; Ohbayashi, 2021). Supplementary motor regions have also been reported in fMRI studies of perceptual decision-making and choice reaction-time tasks, where they have been associated with response preparation and selection processes that interact with broader cortical decision systems (Cunnington et al., 2003; Nachev, Kennard, and Husain, 2008). Because the present design used a counterbalanced two-alternative forced-choice judgment, these effects are unlikely to reflect a simple confound between condition and a fixed motor response. However, they may still reflect response-related processes associated with subjective decision-making, such as action preparation, response selection, or decision difficulty, rather than perceptual processing of the face stimuli alone. The exploratory contrast does not separate these perceptual and response-related contributions. These sensorimotor regions did not appear in the objective Real vs. AI contrast, which is consistent with the possibility that they are more closely tied to the behavioral judgment component of the task.

The frontal pole cluster, the only Perceived AI > Perceived Real effect, falls within anterior prefrontal cortex. This region has been associated with high-level cognitive control, abstract evaluation, prospective reasoning, multitasking, and the coordination of multiple goals or task sets (Ramnani and Owen, 2004; Koechlin and Hyafil, 2007; Burgess, Dumontheil, and Gilbert, 2007). Frontal pole has additionally been implicated in metacognitive evaluation of one's own perceptual or cognitive states (Burgess, Gilbert, and Dumontheil, 2007; Fleming and Dolan, 2012). The appearance of a frontal pole effect in the Perceived AI > Perceived Real direction parallels the frontal pole effect observed in the objective AI > Real contrast in GLM 1, although the location and extent are not identical. One descriptive interpretation is that images judged as AI-generated, like images that are objectively AI-generated, were associated with relatively greater anterior evaluative processing. The present

design does not isolate evaluation, novelty processing, metacognition, or cognitive control as separable mechanisms. The cerebellar cluster appeared in the Perceived Real > Perceived AI direction. The cerebellum has increasingly been implicated in non-motor cognitive functions, including timing, prediction, attentional processing, and contributions to perceptual and cognitive task performance (Strick, Dum, and Fiez, 2009; Schmahmann, 2019). Without a more focused analysis we do not interpret the cerebellar cluster further, but note that cerebellar involvement is increasingly common in fMRI studies of perceptual and cognitive tasks.

The relationship between the perceived-realness contrast and the objective Real vs. AI contrast is informative even under exploratory framing. Both contrasts implicated early visual cortex in the same direction: regions of intracalcarine cortex and adjacent occipital territory responded more strongly to images that were either objectively real or perceived as real than to images that were either objectively AI-generated or perceived as AI-generated. Both contrasts also implicated frontal pole in the same direction: anterior prefrontal cortex responded more strongly to images that were either objectively AI-generated or perceived as AI-generated. This convergence is notable given that participants classified images correctly above chance but imperfectly (63% accuracy), implying that subjective judgments of realness tracked some features of the objective real-versus-AI distinction while also containing substantial decision noise.

The convergence is also restricted to early visual cortex and frontal pole, both of which are large and reliably activated regions in many fMRI tasks, increasing the chance that overlap occurs on geometric grounds. The two contrasts also differed in important respects. The objective Real vs. AI contrast yielded AI > Real effects in lingual gyrus, precuneus, and cingulate gyrus that did not appear in the perceived-realness contrast, while the perceived-realness contrast yielded Perceived Real > Perceived AI effects in inferior temporal gyrus, fusiform gyrus, postcentral gyrus, juxtapositional lobule cortex, and cerebellum that did not appear in the objective contrast. These divergences indicate that subjective realness and objective image category, while related, recruited at least partially distinct sets of regions in this dataset.

A possible descriptive interpretation is that early visual cortex effects in both contrasts may be driven by low-level image statistics, which would be expected to influence both the objective image category and the subjective realness judgment without requiring a specific cognitive mechanism. The frontal pole convergence is

consistent with the possibility that AI-related categories, both objective and subjective, engage broadly evaluative anterior prefrontal processing. The divergences are consistent with the possibility that subjective realness judgments additionally engage ventral-stream face-relevant processing and response-related sensorimotor processing, while objective image category additionally engages medial parietal and cingulate evaluative processing. The present design does not isolate perceptual fluency, uncertainty, naturalness, salience, response-related processing, or explicit AI detection as separable psychological mechanisms, and these interpretations should be treated as hypotheses suggested by the data rather than as confirmed effects, particularly given the exploratory framing of these analyses.

Our results find that the regions implicated by the Perceived Real vs. Perceived AI contrast span systems that have, in prior work, been associated with low-level visual encoding, ventral-stream object and face processing, sensorimotor response processing, supplementary motor planning, cerebellar contributions to perceptual and cognitive task performance, and anterior prefrontal evaluation. The pattern partially parallels and partially diverges from the objective Real vs. AI contrast, with the most reliable convergence at the level of regional involvement in early visual cortex and frontal pole. The convergence with prior literature, and with the objective contrast, is at the level of regional involvement only; the present design does not establish that the same computations are engaged here as in the cited studies or as in the objective contrast.

Table 5.4: Significant clusters from the Perceived Real vs. Perceived AI contrast (GLM 2). Threshold: p-value < 0.005, $k > 50$ voxels.

Cluster ID	X	Y	Z	Peak t	Cluster Size (mm^3)	ROI Region
1	-44	-70	-28	5.10	1864	Cerebellar
2	-30	-86	-8	4.72	6896	Intracalcarine
3	-8	-90	8	4.32	1248	Occipital Pole
4	4	8	68	4.30	896	Juxtapositional Lobule Cortex
5	-52	40	-4	-4.26	1056	Frontal Pole
6	54	-40	-22	4.19	584	Inferior Temporal Gyrus
7	-36	-80	-44	3.80	416	Cerebellar
8	4	-34	78	3.75	472	Postcentral Gyrus
9	36	-80	-8	3.55	552	Fusiform Gyrus

Perceived Real > Perceived AI encoding contrast (sorted by t value, grouped by cluster). ROIs determined by Harvard–Oxford atlas within a 4mm radius of coordinates; the majority label was chosen.

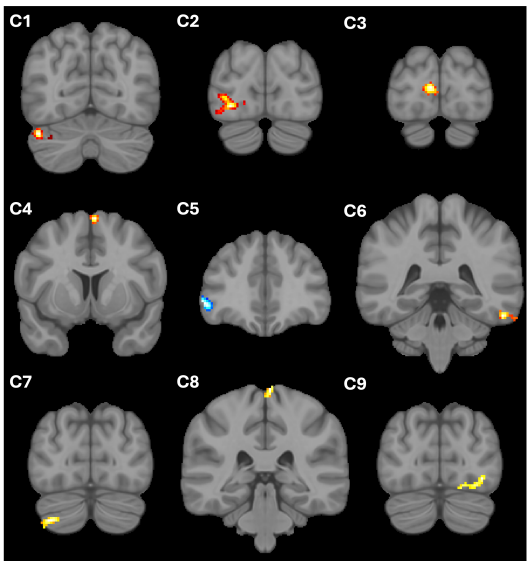


Figure 5.10: Cluster masks from GLM 2, thresholded at voxelwise $p < 0.005$ with cluster extent $k > 50$ voxels. Panels labeled C1–C9 correspond to cluster IDs in Table 5.4.

BOLD responses to Correct vs Incorrect responses. GLM 3 allowed us to identify regions that exhibited different responses to correct responses vs incorrect responses. Results are shown in Table 5.5 and Figure 5.11.

The most striking feature of this contrast is the sheer extent of the surviving pattern. At the same voxelwise and cluster-extent thresholds used in GLMs 1 and 2, the accuracy contrast produced 22 surviving clusters, all in the same direction, indicating greater activation for incorrect than correct trials throughout the surviving voxels. Cluster sizes for the largest clusters in this contrast are an order of magnitude larger than those observed in either of the earlier contrasts. The clusters span posterior visual and ventral occipito-temporal regions, sensorimotor and opercular regions, medial and lateral parietal regions, frontal and medial wall regions, subcortical regions including thalamus, and cerebellum. There is essentially no major cortical or subcortical system that does not contain a surviving cluster.

Because this contrast produced many supra-threshold clusters, region-specific interpretations should be treated cautiously. The broad spatial extent of the effect makes it difficult to attribute the pattern to a small number of focal cognitive processes. Thus, the cluster labels should be interpreted mainly as anatomical summaries of supra-threshold effects, rather than as strong evidence for distinct psychological mechanisms.

Within that strong caveat, the regions implicated by the contrast are nonetheless familiar from existing literatures, and we describe their typical involvement below. We do this primarily to characterize what kinds of systems differed between trial types, not to claim that any specific cognitive process was selectively engaged on incorrect trials.

The posterior visual and ventral occipito-temporal clusters fall within regions commonly situated in visual and ventral-stream processing literature, where they are associated with the encoding of low-level visual structure and the transformation of visual features into higher-order object and face-relevant perceptual representations (Kravitz et al., 2013; Heydt, 2023; Martin and Barense, 2023). Ventral-stream regions including fusiform and inferior temporal cortex have been reported to show stronger responses on difficult trials in studies of perceptual decision-making with degraded or ambiguous visual stimuli (Heekeren, Marrett, and Ungerleider, 2008; Philiastides, Heekeren, and Sajda, 2014). Their involvement on incorrect trials is consistent with the possibility that incorrect classifications were drawn disproportionately from harder, more visually ambiguous trials, although the breadth of the broader pattern makes it difficult to definitively attribute these particular clusters to perceptual processing.

The sensorimotor, opercular, parietal, and medial wall clusters fall within systems commonly associated with response preparation and execution, somatosensory processing, salience and interoceptive monitoring, attentional control, and performance monitoring (Cunnington et al., 2003; Cavanna and Trimble, 2006; Nachev, Kennard, and Husain, 2008; Andrews-Hanna, Smallwood, and Spreng, 2014; V. Menon and Uddin, 2010; Ullsperger, Danielmeier, and Jocham, 2014; Gomez-Ramirez, Hysaj, and Niebur, 2016; Ohbayashi, 2021). Within this set, paracingulate and adjacent medial prefrontal regions in particular have been reported in fMRI studies of error monitoring, conflict detection, and performance evaluation (Ullsperger and Cramon, 2001; Ridderinkhof et al., 2004; Botvinick, 2007; Shenhav, Botvinick, and Cohen, 2013; Ullsperger, Danielmeier, and Jocham, 2014), and central opercular and adjacent insular territory has been associated with arousal, interoceptive monitoring, and salience processing under uncertainty (V. Menon and Uddin, 2010; Ullsperger, Danielmeier, and Jocham, 2014). The combined recruitment of these systems on incorrect trials is consistent with the broader literature on performance monitoring under demanding task conditions, but the present contrast cannot adjudicate among conflict, monitoring, salience, response-conflict, and effort-related interpretations.

The frontal clusters, including middle frontal gyrus, superior frontal gyrus, and frontal pole, fall within regions commonly associated with cognitive control, evaluative processing, prospective reasoning, multitasking, and metacognitive evaluation of one's own perceptual or cognitive states (Ramnani and Owen, 2004; Koechlin and Hyafil, 2007; Burgess, Dumontheil, and Gilbert, 2007; Fleming and Dolan, 2012). The presence of similar regions on incorrect trials is consistent with this broader literature, although as with the medial wall clusters, the spatial extent of the overall contrast cautions against attributing these effects to any specific control or evaluation process.

Thalamic and cerebellar clusters indicate that the incorrect-trial effect was not restricted to cortical visual, sensorimotor, or frontal systems. Thalamic involvement has been reported in attention, arousal, and cortico-subcortical loops supporting task performance, and the cerebellum has increasingly been implicated in non-motor cognitive functions including timing, prediction, and contributions to perceptual and cognitive task performance (Strick, Dum, and Fiez, 2009; Schmahmann, 2019). Their presence here is more consistent with general task-engagement and arousal differences between trial types than with a specific perceptual or evaluative process.

Taken together, the regions implicated by the contrast span essentially all major cortical systems plus subcortical and cerebellar regions, with all surviving clusters showing the same direction of effect. Several non-mutually-exclusive mechanisms could plausibly contribute to such a pattern, including increased visual ambiguity, decision uncertainty, response conflict, error monitoring, longer effective processing time on harder trials, and correlations with reaction time, arousal, or effort (Heekeren, Marrett, and Ungerleider, 2008; Shenhav, Botvinick, and Cohen, 2013; Ullsperger, Danielmeier, and Jocham, 2014). The result is more consistent with a broad task-difficulty or effort-related difference between correct and incorrect trials than with a focal cognitive contrast, and the present design does not separately isolate perceptual ambiguity, confidence, response conflict, attention, or error monitoring.

The relationship between this contrast and the earlier Real vs. AI and Perceived Real vs. Perceived AI contrasts is informative even under exploratory framing. The accuracy contrast included posterior visual and ventral occipito-temporal regions, frontal pole, and precuneus, all of which appeared in the earlier objective or perceived contrasts. At a descriptive level, this is consistent with the possibility that trial accuracy is related to some of the same broad systems that distinguish real from AI-generated faces or that distinguish images perceived as real from images

perceived as AI-generated. We are nonetheless cautious in reading this as substantive convergence. The accuracy contrast is spatially extensive enough that overlap with more focal contrasts is expected on geometric grounds alone, and anatomical overlap does not by itself establish shared underlying computation. Given that the earlier contrasts also did not replicate consistently across splits, the apparent convergence may partly reflect shared exploratory noise structure rather than a stable neural commonality.

The accuracy contrast was also clearly broader than the earlier contrasts. Whereas the Real vs. AI and Perceived Real vs. Perceived AI models produced a roughly posterior-visual versus anterior-evaluative organization, the accuracy model additionally implicated opercular, sensorimotor, parietal, temporal, thalamic, and cerebellar regions. This suggests that, in this dataset, trial accuracy was not simply a weaker version of the real-versus-AI distinction but instead indexed a broader task-performance or effort dimension. Such a dimension might include perceptual difficulty, uncertainty, attentional control, effort, and response selection, but the present analyses cannot separate these contributions.

The Correct vs. Incorrect results are best interpreted as exploratory evidence that incorrect classifications in this task were associated with broader recruitment of perceptual, sensorimotor, parietal, and control-related systems, likely reflecting general task-difficulty or effort differences between trial types rather than a specific psychological mechanism or a localized neural correlate of accurate real-versus-AI judgment. The convergence with prior literature, and with the earlier contrasts, is at the level of regional involvement only; the present design does not establish that the same computations are engaged here as in the cited studies or as in the earlier contrasts.

Table 5.5: Significant clusters from the Correct vs. Incorrect contrast. Threshold: p-value < 0.005, $k > 50$ voxels.

Cluster ID	X	Y	Z	Peak t	Cluster Size (mm^3)	ROI Region
1	54	-48	-24	-5.85	2912	Temporal Occipital Fusiform Cortex
2	-4	4	-8	-5.49	4240	Left Cerebral
3	-28	12	62	-5.49	12816	Middle Frontal Gyrus
4	12	-86	-30	-5.23	1472	Lingual Gyrus
5	-10	-36	40	-5.19	11568	Precuneus Cortex
6	40	-74	-28	-5.11	3088	Cerebellar
7	22	-26	70	-5.11	9360	Precentral Gyrus
8	-16	40	0	-5.07	11896	Paracingulate Gyrus
9	56	-8	10	-5.04	14304	Superior Temporal Gyrus
10	30	38	48	-5.00	3344	Superior Frontal Gyrus
11	38	-56	30	-4.57	3688	Angular Gyrus
12	-58	-2	6	-4.56	6480	Central Opercular Cortex
13	-14	-24	66	-4.53	3848	Postcentral Gyrus
14	-26	-82	42	-4.53	6648	Lateral Occipital Cortex
15	-2	-6	10	-4.43	1304	Left Thalamus
16	-62	-40	-18	-4.34	1288	Inferior Temporal Gyrus
17	-6	-88	24	-4.32	3656	Cuneal Cortex
18	-18	-76	-12	-4.32	2568	Occipital Fusiform Gyrus
19	6	30	10	-4.15	576	Right Cerebral
20	-10	24	-2	-4.05	504	Left Cerebral
21	64	10	22	-4.04	424	Precentral Gyrus
22	-20	58	0	-3.96	648	Frontal Pole

Correct vs. Incorrect contrast (sorted by absolute t value at the cluster level, with subpeaks kept attached to their parent cluster). ROIs were determined using the Harvard–Oxford atlas within a 4 mm radius of each peak coordinate; the majority label was chosen. Cluster size is shown only for the highest-$|t|$ peak within each cluster.

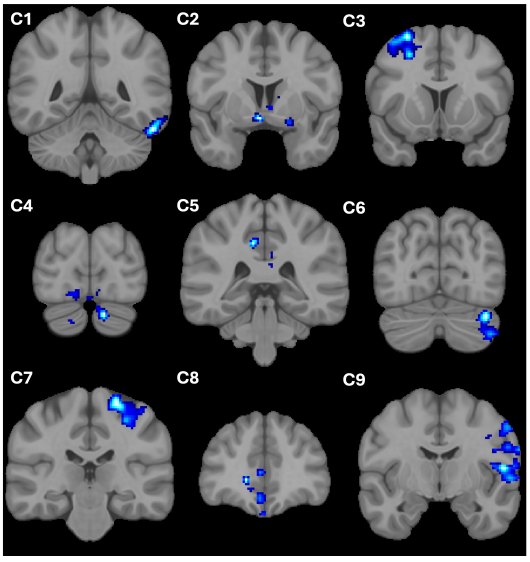
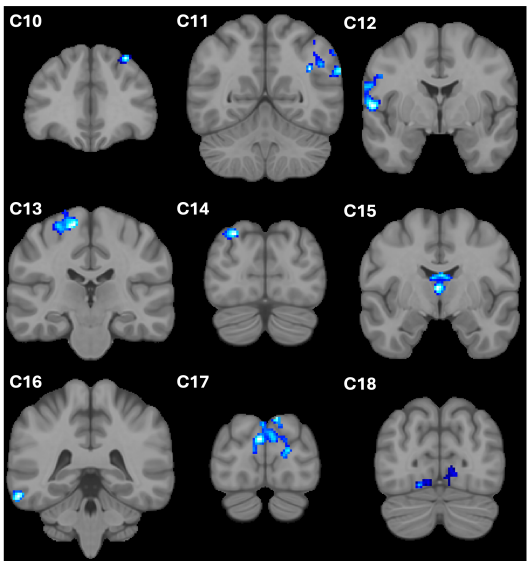
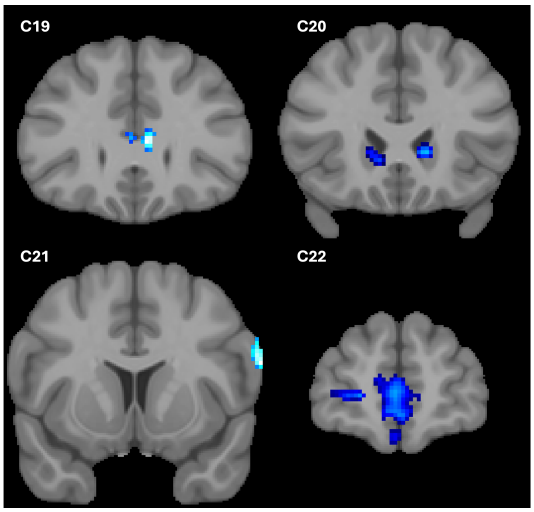


Figure 5.11: Cluster masks from GLM 3, thresholded at voxelwise $p < 0.005$ with cluster extent $k > 50$ voxels. Panels labeled C1–C9 correspond to cluster IDs in Table 5.5.



Cluster masks from GLM 3, thresholded at voxelwise $p < 0.005$ with cluster extent $k > 50$ voxels. Panels labeled C10–C18 correspond to cluster IDs in Table 5.5.



Cluster masks from GLM 3, thresholded at voxelwise $p < 0.005$ with cluster extent $k > 50$ voxels. Panels labeled C19–C22 correspond to cluster IDs in Table 5.5.

5.4 Conclusion

In this study, we tested whether neural responses to face images contain information that helps distinguish AI-generated faces from real photographs, and whether these signals provide information beyond overt behavioral judgments. Across both the behavioral pretest and the fMRI task, participants showed that some AI-generated faces are now genuinely difficult to identify, with performance approaching chance for the hard stimulus set, while easier images remained only modestly discrim-

inability. This confirms the central premise of the paper: modern synthetic faces are often hard for humans to classify, making improved detection methods an important scientific and societal problem.

Our neural results suggest that the brain does carry information related to image origin. Whole-brain encoding analyses revealed reliable differences in activity between real and AI-generated faces, with effects spanning early visual cortex as well as parietal and frontal regions. In particular, AI-generated faces tended to elicit relatively greater activation in regions such as the precuneus, cingulate, and frontal pole, whereas real faces showed stronger responses in more limited occipital regions. A related pattern emerged when images were analyzed according to participants' subjective judgments, suggesting that both stimulus-driven properties and perceptual interpretation contribute to the neural representation of synthetic versus real faces.

At the same time, the broader message of the paper should remain measured. The presence of distinguishable neural patterns does not yet imply a robust or practically deployable neuro-based deepfake detector. The evidence is better interpreted as proof of concept: fMRI signals appear to contain modest but meaningful information about whether a face is AI-generated, but those effects are distributed, relatively small, and likely sensitive to analytic choices and stimulus composition. The paper therefore supports the scientific claim that neural data can reveal aspects of AI-image detection that are not fully captured by overt behavior, while stopping short of claiming that neural decoding currently solves the deepfake detection problem.

An important contribution of this study is that it begins to connect detection performance to human-interpretable features. The behavioral feature analysis suggests that a compact set of latent perceptual attributes is informative for distinguishing AI-generated from real faces, raising the possibility that some of the relevant signal is not purely black-box but can be linked to image properties that humans may eventually be trained to recognize more effectively. This provides a useful bridge between machine classification, human judgment, and neural representation, and points toward a more interpretable framework for studying synthetic media perception.

Several limitations should be acknowledged. The study uses modest sample sizes, focuses on a restricted set of face images, and relies on one primary image-generation pipeline for the AI stimuli. As a result, it remains unclear how well the observed neural signatures generalize across generators, across broader image classes, or across other forms of manipulated media. Future work should test larger and more

diverse samples, compare multiple image-generation systems directly, and more systematically manipulate candidate perceptual features to determine which cues are stably represented in the brain and most useful for classification.

Overall, this paper provides initial evidence that neural activity differentiates real from AI-generated faces and may capture aspects of synthetic-image detection beyond conscious report alone. Rather than offering a definitive neural solution to deepfake detection, the results establish a foundation for future work at the intersection of neuroforecasting, perception, and AI safety. As synthetic media become more realistic and widespread, understanding how the brain encodes "realness" may prove increasingly important both for basic science and for the development of more effective, interpretable detection tools.

5.A Decoding Methods

In this section, we explore the results of the exploratory decoding analysis conducted on half (20 subjects) of the total dataset. We find no robust ability to decode real from AI-generated faces using our methods. We do see some weak signal that approaches a trend but nothing that is significant. **GLM 4: Trial-Level Estimation Using Least Squares–Separate (LSS)**

To generate subject-specific, single-trial beta maps for multivariate decoding analyses, we employed a Least Squares–Separate (LSS) approach Mumford et al., 2012. The LSS method involves estimating a separate general linear model (GLM) for each trial in the experiment, allowing us to obtain an unbiased voxel-wise activation estimate for every individual stimulus presentation.

Specifically, for each of the 200 trials (100 AI and 100 Real images), we fit a separate GLM for each subject. Each model contained:

- R1: An indicator function for the duration of image presentation modeling the BOLD response to the single target trial;
- R2: A single indicator function for the duration of image presentation modeling 'other stimulus' regressor for the combined response to all other trials in the session;

The regressors of interest were modeled from stimulus onset with a duration equal to the subject's reaction time on that trial. The 'other stimulus' regressor was also modeled from stimulus onset with a duration equal to a subject's reaction time, but was concatenated with all trials other than the stimulus of interest. The model also included three motor (left-button press, right-button press, no-button press) nuisance regressors as stick regressors (with a duration of 0) of no interest, as well as six motion regressors of no interest (translation and rotation). The regressors of interest and the motor regressors were convolved with a canonical hemodynamic response function (HRF). A spatial smoothing isotropic Gaussian kernel of 4mm full width at half maximum (FWHM) was applied to all functional data before model estimation.

This procedure yielded 200 beta maps per subject, one for each trial, forming the basis of our decoding analysis.

Building Decoding Input Vectors

Single-trial beta maps were first estimated with the LSS GLM (see 5.A) and then averaged across subjects to improve the signal-to-noise ratio. This produced 200 mean beta images (100 AI, 100 Real).

Using the exploratory dataset only, we re-estimated GLM 1 and thresholded the *Real* > *AI* contrast at $p < 0.005$ (cluster extent $k > 50$ voxels). The resulting binary mask was then applied unchanged to the exploratory datasets after which each mean beta image was vectorized[2] to yield one row vector per stimulus that contained all voxels present in our mask. We acknowledge the circularity of this approach and view this as a exploratory analysis designed to measure the upper limit of neuronal data's ability to classify these images.

Neuro-Only Model Design Generalization was quantified with stratified 20-fold cross-validation, such that the classes balance between AI and Real images was equal throughout every fold. The evaluation metric was the area under the ROC curve (ROC-AUC), averaged across folds (though we also report accuracy).

Within each cross-validation fold, the voxel input vectors were decomposed with PCA; the top ten components (explaining roughly 22% of training variance on average; range 21.7%-22.3% across 20 folds) were retained. The same fold-specific loadings were then applied to the test partition. Refitting PCA in every fold guarantees that feature extraction remains blind to the held-out data.

Reduced vectors were then entered into an unregularized sparse linear support vector classifier (SVC), with an L1 norm. We chose L1 because it encourages voxel-sparse solutions that have been shown to generalize well in fMRI decoding (Mohr et al., 2015). We did not perform any parameter tuning and instead chose to use PCA rather than regularization to combat overfitting. A supplementary analysis with an unregularized L2-penalized SVC yielded indistinguishable performance.

Statistical significance was evaluated with a non-parametric permutation procedure. Condition labels were randomly shuffled 1000 times within each outer fold, and the entire model pipeline was recomputed for every shuffle, yielding a null distribution of decoding accuracies against which the empirical score was compared.[3]

[2] Vectorization employed `NiftiMasker` with 1 mm FWHM Gaussian smoothing, and per-sample z-scaling.

[3] Permutation testing was done utilizing the `permutation_test_score` function from sklearn. All analyses were conducted in Python 3.11 using `nilearn` 0.10, `numpy` 1.26, and `scikit-learn` 1.4. Code, preprocessing scripts, and pre-computed beta images are openly available at github.com/TaoTeTom/AI_Detection.

Neuro+Behavior Model Design In this modification we include subject's behavioral response as an additional regressor to the ten principal components from the earlier design. Generalization was also quantified using stratified 20-fold cross-validation, ensuring an equal class balance between AI and Real images throughout every fold. The evaluation metric was the area under the receiver operating characteristic (ROC) curve (ROC-AUC), averaged across folds (though we also report accuracy).

Within each cross-validation fold, the voxel input vectors were decomposed with Principal Component Analysis PCA; the top ten components (explaining roughly 22% of training variance on average; range 21.7%-22.3% across 20 folds) were retained. The same fold-specific loadings were then applied to the test partition. Refitting PCA in every fold guarantees that feature extraction remains blind to the held-out data. We then added the average group behavioral accuracy for each image, creating an input matrix of 11 features: our 10 neuro-PCs and the average behavioral classification score across scanner subjects.

PCA reduced vectors were then fed into an unregularized sparse linear SVC, with an L1 norm. We chose L1 because it encourages voxel-sparse solutions that have been shown to generalize well in fMRI decoding (Mohr et al., 2015). We did not perform any parameter tuning and instead chose to use PCA rather than regularization to combat overfitting. A supplementary analysis with an unregularized L2-penalized SVC yielded indistinguishable performance; the numbers were exactly the same.

We report the results of the classifier in both the confirmatory sample and the combined results. Statistical significance was evaluated with a non-parametric permutation procedure. Condition labels were randomly shuffled 1000 times within each outer fold, and the entire model pipeline was recomputed for every shuffle, yielding a null distribution of decoding accuracies against which the empirical score was compared.[4]

Cross Mask Easy vs Hard Mask Detection We also tested whether whether our neural data and decoding upper bound pipeline performed better for easy trials or for hard trials, and also whether masks formed from easy image trials or hard image trials would be more informative in predicting the other difficulty task.

To do so, we first split our exploratory dataset into two sections based on difficulty.

[4]Permutation testing was done utilizing the `permutation_test_score` function from sklearn. All analyses were conducted in Python 3.11 using `nilearn` 0.10, `numpy` 1.26, and `scikit-learn` 1.4. Code, preprocessing scripts, and pre-computed beta images are openly available at github.com/TaoTeTom/AI_Detection.

For each difficulty section, we generated a mask by re-estimating our initial encoding with the same cluster correction[5]., Real-AI GLM. This left us with an "easy-trial" mask and a "hard-trial" mask.

For each difficulty section, we generated two sets of input vectors. One for each of the two masks above. This left us with four different models: Easy-Block/Easy-Mask (EBEM), Easy-Block/Hard-Mask (EVBHM), Hard-Block/Easy-Mask (HBEM), Hard-Block/Hard-Mask (HBHM).

Binned Responses We also aimed to investigate the relationship between neural decoding performance and behavioral difficulty, implementing a leave-one-bin-out cross-validation procedure. Single-trial beta maps (from the LSS GLM; see Section 5.A) were vectorized within the binary ROI mask and concatenated across subjects to form a trial-by-voxel matrix. We used the same binary ROI mask generated from the standard decoding analysis (GLM). Images, along with their vectorized single-trial beta maps, were assigned to bins based on their behavioral accuracy in the exploratory sample ("Correct Bin"), with Bin 1 containing the 10 images with the lowest accuracy and Bin 20 containing the highest.

For each iteration, all trials from one bin were held out as the test set, and the remaining 19 bins were used for training. The image vectors were then reduced to the top ten principal components using PCA, fit only on the training set, and then projected onto the held-out bin. Classification was performed using a sparse, unregularized linear SVC with an L1 penalty. Model performance for each bin was quantified using the area under the ROC curve (AUC), with accuracy also computed for descriptive purposes. This procedure produced a per-bin profile of decoding performance that could be directly compared to mean behavioral accuracy for the same stimuli.

Within Group Decoding We then tested the ability of our decoding model to decode within groups. We first split our stimulus set in half into a training and testing set, ensuring that for each iteration, all four trial types (AI/Real × Easy/Hard) were evenly represented across the training and testing sets.

We then re-estimated our initial encoding GLM (GLM 1) on the training set only. We then built a mask from the beta maps in this training set, but applied a cluster correction (p-value = 0.001 and cluster-size > 20). The cluster correction was performed using the hyperparameter settings previously optimized on the exploratory

[5] p-value < 0.005 and cluster-size > 50)

dataset. This mask was then applied to our held-out test set beta images, generated by GLM 5.A.

We then reduced the masked input vectors to their first principal component. This choice was determined through a grid search on the exploratory dataset,[6] and this best-performing parameter set indicated retaining only the first principal component.

The resulting PC-reduced vectors were then decoded using a leave-one-per-class-out cross-validation procedure (one AI image and one Real image held out in each fold), yielding a 50-fold model.

To evaluate whether decoding performance was limited by the amount of data available for decoder training (as opposed to mask construction), we also implemented a 30 Mask/70 Train split of the stimulus set. In this variant, 30% of the stimuli were used solely to estimate the encoding model (GLM 1) and to construct the voxelwise mask using the previously optimized hyperparameters ($t = 2.3$, $p = 0.052$, cluster size > 40). The remaining 70% of the stimuli—after masking—were used to train and evaluate the decoder using leave-one-per-class-out cross-validation (one AI and one Real stimulus per fold, yielding a 70-fold model).

This approach allows us to test whether providing substantially more data to the decoding stage (relative to the 50/50 split) improves classification performance. If decoding accuracy increases under the 30 Mask/70 Train allocation, it would indicate that decoder performance is sample-limited; if accuracy remains comparable, it suggests that mask quality, rather than decoder training size, is the primary constraint. To assess stability, we repeated this procedure across ten random 30 Mask/70 Train partitions.

Across Group Robustness We also tested whether our pipeline generalized to unseen participants, we separated our subjects into random splits (5 "mask creator subjects" and 15 "mean-beta map creator subjects"). We repeated this random split five times. We then ran our GLM 1 pipeline on the "mask creator subjects" and calculated the mean beta map across these subjects to generate our mask. We then used this mask on our beta maps to create mean-beta maps composed of the 15 "mean-beta map creator subjects."

[6]Grid parameters: cluster-forming t-values $t \in (2, 2.3, 2.5, 2.75, 3.0, 3.25, 3.5, 3.75, 4.0)$; cluster-extent thresholds $\in (10, 15, 20, 30, 40, 50)$; number of retained principal components $\in (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 20, 25, 30)$. For each parameter combination, we computed voxelwise masks and selected the combination that produced the largest number of masks reaching statistical significance in the exploratory dataset.

We then ran our test set through the same decoding pipeline as described earlier, ensuring that we used the mask generated from our"mask creator subjects" and our beta-vectors from our "mean-beta map creator subjects" for each robustness check iteration.

All behavioral accuracy and reaction time comparisons between conditions (e.g., Easy vs. Hard, AI vs. Real, Online vs. fMRI) were evaluated using two-tailed Welch's t-tests, which do not assume equal variances between groups. Reported p-values correspond to these tests unless otherwise noted. Permutation testing was used exclusively for evaluating classifier performance.

Searchlight Analysis Pipeline

Neuro-Only Searchlight Decoding To complement our ROI-based and GLM-masked decoding analyses, we performed a whole-brain searchlight analysis to localize predictive neural information throughout the cortex without a priori regional constraints.

For each subject, we conducted a voxel-wise searchlight decoding analysis using a 5.6 mm spherical radius. Within each sphere, we extracted the mean beta-activation patterns from our single-trial LSS maps and concatenated them with trial-level behavioral responses (e.g., classification decisions). This combined neural-behavioral input was then used to train a linear SVC without regularization. Model performance was evaluated using a 10-fold cross-validation strategy applied within each sphere.

The decoding accuracy maps from each subject were transformed into standard MNI space. To produce a group-level map, we applied a threshold of 58% decoding accuracy and identified voxels that survived this threshold in at least four participants. Cluster correction was applied at a minimum cluster size of 40 voxels.

Whole-Brain Searchlight Decoding: Neuro + Behavior

To assess whether behavioral responses enhance neural decoding of image identity, we performed a second whole-brain searchlight analysis that incorporated both BOLD activation and trial-level behavioral inputs. For each subject, we ran a 5.6 mm radius searchlight using the same single-trial beta maps as in previous decoding models. In each searchlight sphere, voxel activation was concatenated with behavioral covariates (e.g., image judgment or confidence ratings), forming a combined feature vector for classification.

An unregularized support vector classifier (SVC) was trained and evaluated using a 10-fold cross-validation scheme, independently applied within each sphere. This process yielded a whole-brain decoding accuracy map for each subject based on Neuro + Behavior inputs.

To identify consistent predictive regions across participants, we binarized individual decoding maps using a 58% accuracy threshold and selected voxels present in at least four subjects. A cluster-size threshold of 40 voxels was applied to minimize false positives.

5.B Decoding Results

Multivariable Pattern Analysis (MVPA) Decoding Real vs AI The upper-bound MVPA decoder, using 10 principal components, achieved a mean ROC-AUC of 69.8%. For comparison, participants' behavioral classification accuracy was 62.5%.[7] Because this analysis used circular feature selection, this result should not be interpreted as an unbiased estimate of generalizable decoding performance. Rather, it suggests that, under favorable feature selection conditions, group-average neural responses contained information about image authenticity that was at least comparable to, and potentially more separable than, participants' overt behavioral classifications. Permutation testing was conducted with 1000 label shuffles to establish a null distribution. The mean accuracy under the null distribution was 50.2%, and the empirical subject-wise p-value was $p < 0.001$, indicating that decoding performance was significantly above chance. We outline the performance curves in section 5.B.

To ensure that our decision to choose a PC count of 10 was not arbitrary, we also reran our model for each PC up until 100% variance explained (192 principal components) and generated a curve of AUC scores for every principal component amount. Below in Figure 5.12, the model's performance as a function of the number of principal components is shown. We also show the cumulative variance explained curve as a function of the total number of principal components in the inset.

[7] A parallel analysis using accuracy as the decoding metric produced a mean model score of 63.5%.

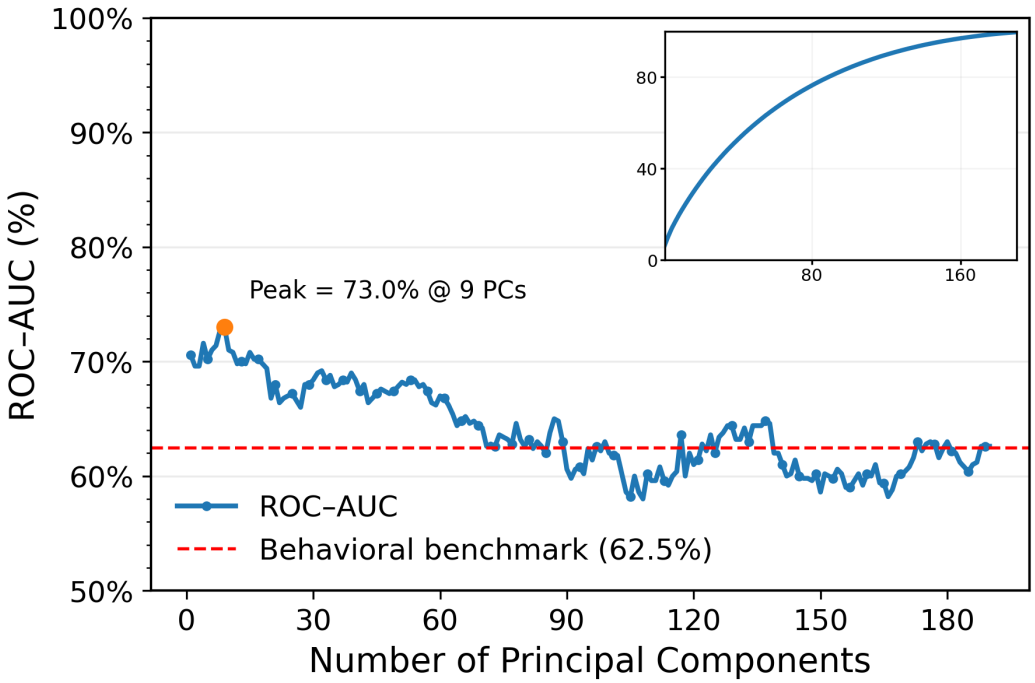


Figure 5.12: Neuro-only Data Classification AUC scores by number of Principal Components. **Inset:** Explained variance by Principal Component.

Our MVPA decoder when combined with subjects behavioral responses, with a principal component count of 10, achieved a mean ROC-AUC of 86.2%, outperforming our subjects' behavioral performance of 62.5% AUC, and the neuro-only model performance of 68.2% AUC.[8]

Permutation testing was conducted with 1000 label shuffles to establish a null distribution. The mean accuracy under the null distribution had an AUC of 50.2%, and the empirical subject-wise p-value was $p = 0.001$, indicating that decoding performance was significantly above chance.

We also performed a permutation test in which the neural data were randomly shuffled across trials while retaining the correct average behavioral score for each image. This procedure generated a null distribution of decoding performance with a mean AUC of 84.4%. The observed decoding accuracy exceeded this null distribution, yielding a subject-wise empirical p-value of $p = 0.047$, indicating that neuronal data adds significant classification power over behavioral signal alone.

[8] A parallel analysis with accuracy as the metric returned a model score of 80.0%.

To ensure that our decision to choose a PC count of 10 was not arbitrary, we reran our model for each PC up until 100% variance explained (192 principal components) and generated a curve of AUC scores for every principal component amount. Our model remained the same. Below in Figure 5.13 the model's performance as a function of the number of principal components is shown. We also show the cumulative variance explained curve as a function of the total number of principal components in the inset.

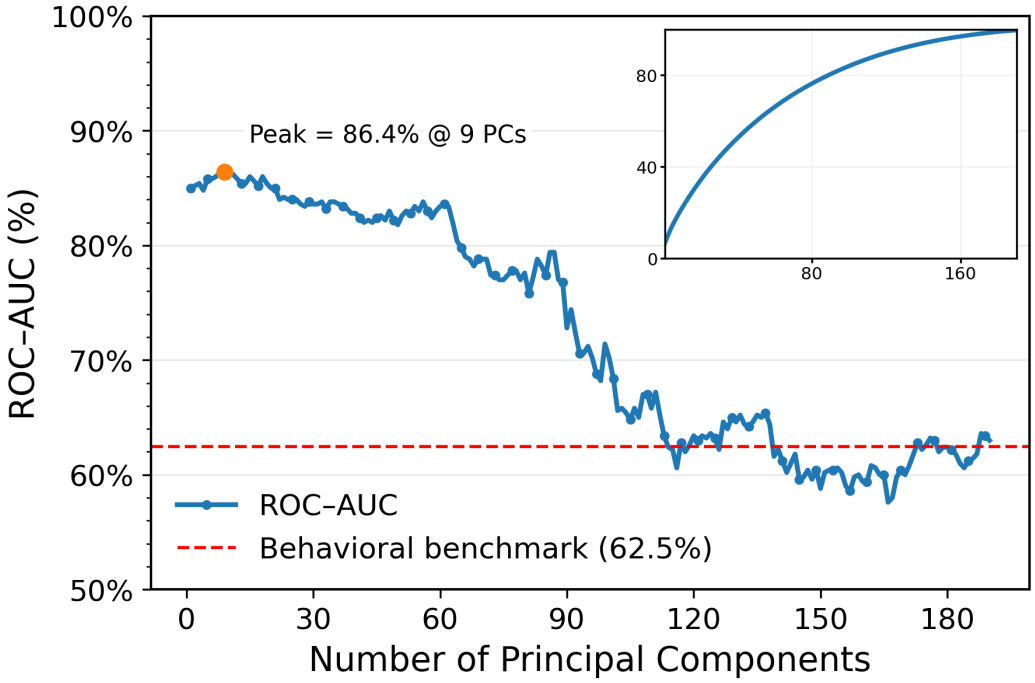


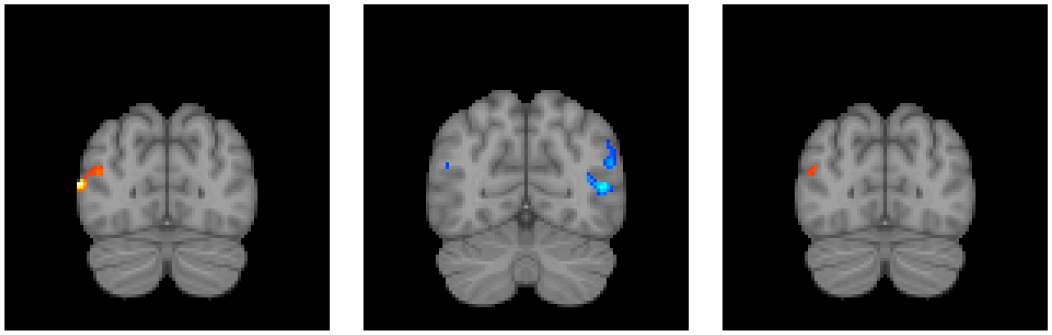
Figure 5.13: Neuro+Behavior Data Classification AUC scores by number of Principal Components. **Inset:** Explained variance by Principal Component.

Easy vs Hard Classification Performance Although humans performed better on easy than hard trials, neuro-decoding performance was similar (p-value = 0.757) across conditions: ROC–AUC of 65.8% (p-value from chance = 0.020) for easy trials and 68.3% (p-value from chance = 0.011) for hard trials. These results suggest that images that are easier for humans to identify consciously may not be easier to decode in the brain.

Our results show that when the difficulty of the mask and the input vectors was matched (Easy–Easy and Hard–Hard), decoding performance was robust: both models achieved ROC-AUC values of 65.0% and 61.0%, respectively, with permutation tests indicating significance from chance ($p = 0.028$, $p = 0.002$, respectively).

In contrast, cross-difficulty transfer was weak. Voxels identified from easy trials provided virtually no predictive value for hard trials (ROC-AUC = 45.0%, p = 0.802), while the reciprocal transfer from hard masks to easy vectors was also not above chance (ROC-AUC = 58.0%, p = 0.089 for ROC-AUC and non-significant for accuracy). These results indicate that informative neural patterns are largely difficulty-specific: successful decoding hinges on selecting a feature space (mask) tailored to the current trial type, rather than on the intrinsic ease or difficulty of the stimuli themselves.

Below in Figure 5.14 and Table 5.6, you can see the masks generated for the easy images and hard images as well as the overlaps between the two masks.



(a) Easy Mask y = -78 (b) Hard Mask y = -68 (c) Overlap Mask y = -78

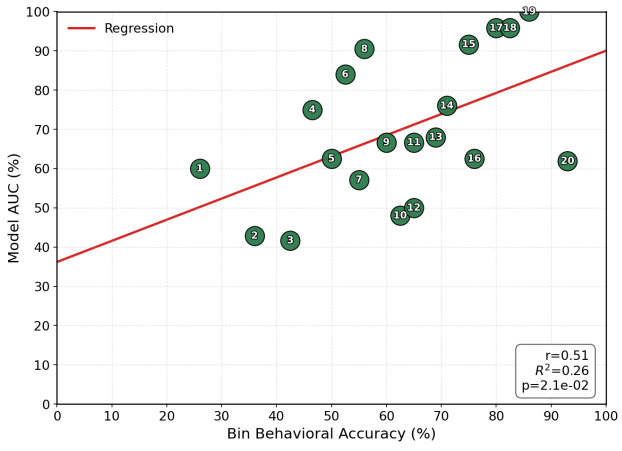
Figure 5.14: Coronal view of masks generated from GLM 1 on a) Easy Trials, b) Hard trials, c) Overlap of Easy Mask and Hard Mask.

Table 5.6: Cluster tables for Easy, Hard, and Overlap. Easy/Hard entries show peak coordinates grouped by cluster (thresholded at $p < 0.005$, $k > 50$ voxels). Overlap shows the intersection computed voxelwise ($|t| > 3$) with component size.

Cluster ID	X	Y	Z	Peak t-value	Cluster Size (mm^3)	Region
Easy Contrast						
1	-54	-78	12	6.79	632	Lateral Occipital Cortex
1a	-48	-82	18	3.76		Lateral Occipital Cortex
1b	-42	-78	20	3.52		Lateral Occipital Cortex
2	6	-32	32	-5.55	440	Cingulate Gyrus
2a	4	-34	36	-5.21		Cingulate Gyrus
Hard Contrast						
1	46	-68	10	-5.23	992	Lateral Occipital Cortex
1a	38	-70	16	-4.63		Lateral Occipital Cortex
1b	42	-64	12	-4.45		Lateral Occipital Cortex
1c	54	-74	6	-4.42		Lateral Occipital Cortex
2	-44	-74	28	-4.94	1600	Lateral Occipital Cortex
Overlap (Easy ∩ Hard)						
1	-44	-78	20	—	56	Lateral Occipital Cortex

Most of the overlap occurs in the visual cortex. However, we do see some overlap in the Right Cingulate Gyrus/Precuneus region, consistent with Cluster 1 from our standard encoding analysis shown in Table 5.3.

Behavioral Binning



Bin	Model AUC (%)	Bin Behavioral Accuracy (%)
1	**60.0**	26.0
2	**42.9**	36.0
3	41.7	42.5
4	**75.0**	46.5
5	**62.5**	50.0
6	**84.0**	52.5
7	**57.1**	55.0
8	**90.5**	56.0
9	**66.7**	60.0
10	48.0	62.5
11	**66.7**	65.0
12	50.0	65.0
13	68.0	69.0
14	**76.0**	71.0
15	**91.7**	75.0
16	62.5	76.0
17	**95.8**	80.0
18	**95.8**	82.5
19	**100.0**	86.0
20	61.9	93.0

Figure 5.15: Per-bin decoding performance (AUC) compared to behavioral accuracy. Bold indicates Model AUC higher than behavioral accuracy.

Model AUC exceeds behavioral accuracy in most bins, particularly in those with

lower behavioral performance. This suggests that the decoding model captures discriminative information even when human participants perform poorly, indicating that the neural signal contains task-relevant features not fully reflected in behavior. However, in bins where behavioral accuracy is already high, the model does not consistently surpass human performance; this indicates that neuro-data provides additional predictive information for images that are more uncertain compared to those that are more easily classified.

Within Group Decoding As stated earlier in the methods section, to validate the reliability of our decoding pipeline, we implemented a train-test split procedure with region masking derived solely from the training data. We report the results of this below

Table 5.7: SVC ROC across 50 Mask/50 Train, Neuro-Data

Mask	**AUC**	***p*-Value**
Mask 1	0.60	0.117
Mask 2	0.64	0.029
Mask 3	0.60	0.107
Mask 4	0.64	0.026
Mask 5	0.60	0.109
Mask 6	0.64	0.039
Mask 7	0.64	0.027
Mask 8	0.62	0.06
Mask 9	N/A	N/A
Mask 10	0.58	0.183
Mean	**0.617**	**N/A**

No clusters survived the threshold for Mask 9.

Across train–test splits, the most consistent mask overlap occurred in a cluster located in the precuneus/retrosplenial cortex. Figure 5.16 and Table 5.8 summarize the spatial overlap across within-group decoder masks. The peak consistency was centered in the left precuneus cortex, with additional overlap in the cingulate cortex, occipital fusiform cortex, and right precuneus. These regions were also identified in the initial GLM 1 cluster map (Figure 5.9), suggesting that the same broad anatomical regions contributed across multiple analytic variants.

Overall, the non-circular within-group decoding analysis provides evidence for weak but above-chance neural signal in some train–test splits. Several training-derived

masks produced ROC-AUC values above chance at conventional or marginal significance thresholds, with AUC values ranging from 0.58 to 0.64 and a mean AUC of 0.617 across usable masks. However, the effect was not consistent across all splits and was sensitive to the precise mask definition. These results therefore support a cautious interpretation: neural responses contain some information relevant to distinguishing real from AI-generated faces, but this information is modest and does not yet provide a robust or reliably generalizable decoding signal.

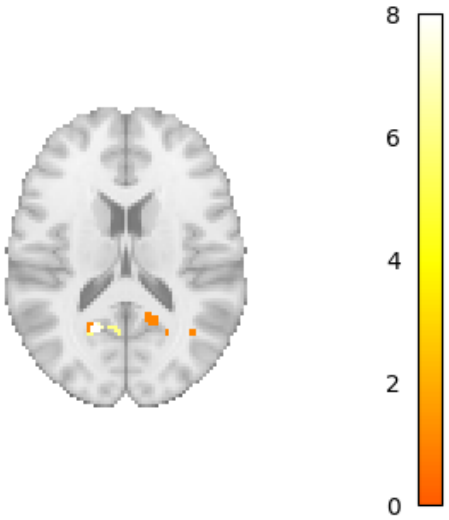


Figure 5.16: Overlap heatmap of regions predictive in multiple decoding robustness checks. The cluster in the Precuneus/Retrosplenial area was consistently implicated across permutations (8/10) | (8/9) of the robustness checks that didn't fail.

Table 5.8: Overlap clusters with peak MNI coordinates, peak overlap counts, and Harvard–Oxford ROI labels.

Cluster	X	Y	Z	Peak overlap (n)	Size (mm^3)	Region (HO)
1	-16	-60	16	8	2712	Left Precuneous Cortex
2	-12	-20	36	6	680	Cingulate Cortex
3	28	-44	-10	2	304	Temporal Occipital Fusiform Cortex
4	20	-54	20	2	136	Right Precuneous Cortex

Cross-Group Robustness As shown in Table 5.9, 0/5 of our cross group splits showed that neural data modeling was significantly higher than chance. This may be due to the low ability of 15 mean-beta map creator subjects" to create accurate

mean-beta images and the inability of five "mask creator subjects" to create accurate masks.

We ran this analysis with 10 "mask creator subjects" and 10 "mean-beta map creator subjects" as well as 15 "mask creator subjects" and 5 "mean-beta map creator subjects"; neither condition produced results.

Table 5.9: SVC classification performance across iterations. Accuracy remains at chance level with no significant decoding.

Iteration	**SVC Accuracy**	**Permutation Score**	p**-value**
1	0.488	0.499	0.63
2	0.488	0.504	0.64
3	0.550	0.504	0.127
4	0.542	0.500	0.16
5	0.482	0.501	0.66

Figure 5.17 shows the overlap map of peak decoding accuracy across subjects, highlighting consistent regions of high information content. The regions that survived thresholding included early visual and associative areas, such as the *Left Lingual Gyrus*, *Primary Visual Cortex*, and *Left Inferior Occipital Gyrus*, as well as higher-order integrative areas, including the *Medial Prefrontal Cortex* and *Anterior Cingulate Cortex*. These findings suggest that classification-relevant signals are distributed across both early perceptual and higher-order evaluative networks.

Neuro-Only Searchlight Results

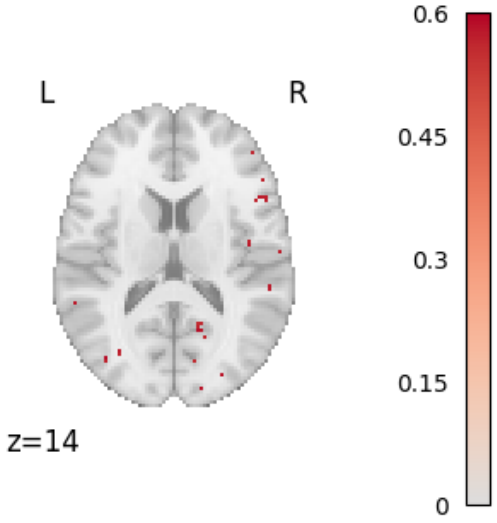


Figure 5.17: Group-level average peak decoding accuracy map from whole-brain searchlight analysis. Only voxels exceeding 58% decoding accuracy in at least 4 subjects and belonging to clusters > 40 voxels are shown.

Table 5.10: Significant clusters from whole-brain searchlight decoding (Threshold: 58%, Cluster size > 40 voxels)

Cluster ID	X (mm)	Y (mm)	Z (mm)	Peak Accuracy	Cluster Size (mm^3)	ROI Region
1	-14	-84	-4	65.0%	264	Left Lingual Gyrus
2	-10	-80	2	59.5%	—	Primary Visual Cortex
3	-8	46	22	64.5%	344	Medial Prefrontal Cortex
4	-10	36	14	59.5%	—	Anterior Cingulate Cortex
5	-14	-88	-12	61.0%	56	Left Inferior Occipital Gyrus
6	-12	-76	2	59.5%	24	Left Lingual Gyrus

Neuro + Behavior Searchlight Results Compared to the neural-only decoding (Figure 5.17), the addition of behavioral information caused the overlapping regions to scatter. We found that multiple participants had consistent decoding contributions only from bilateral *Temporal Poles* and midline structures as opposed to the numerous areas found in the neuro-only searchlight analysis.

Table 5.11: Significant clusters from Neuro + Behavior searchlight decoding (Threshold: 58%, Cluster size > 40 voxels)

Cluster ID	X (mm)	Y (mm)	Z (mm)	Peak Accuracy	Cluster Size (mm^3)	ROI Region
1	26	20	-36	68.9%	880	Right Temporal Pole
2	0	-46	-70	68.9%	984	Midline Occipital
3	-22	16	-38	68.9%	632	Left Temporal Pole

Latent-Variable Encoding Models After finding that there were some aspects of the image that allowed the brain to outperform conscious behavioral performance the next step was to understand what features enabled the brain to do so. What was the potential mechanism that enabled the brain to outperform behavior, but also prevented the brain from implementing behavior? While a full understanding of this mechanism is outside the scope of this paper, we did gain an understanding of some of the potential latent image features that may be relevant in this process. In this section, we describe how we identified 14 latent image features that enable the above behavioral classification. This process was conducted to determine if there were any clues as to what signal the brain was using to make the above behavior classification.

To quantify the perceptual and conceptual properties that could influence face classification, we conducted an online rating study on Prolific.[9] The full stimulus set comprised 200 images. Each image was presented three times, and on each presentation, participants rated a single attribute randomly selected from the 14-item list proposed by Miller et al., 2023. Consequently, every participant rated each image three times, once per attribute, striking a balance between task brevity and adequate sampling of ratings across images. All attributes were assessed on a continuous 0–100 scale, with a 10s response window (mean response time $\approx$ 2s). Data quality was monitored through 11 attention-check trials; only participants who answered all checks correctly received full compensation.

Behavioral Predictive Performance After averaging attribute ratings across raters, every face was encoded as a 14-dimensional feature vector. A logistic-regression model was then fit using *LogisticRegression* from *scikit-learn* 1.4 with the library's default hyperparameters: $C = 1.0$ (L2 regularization), the *liblinear* solver, and unbalanced class weights. The dataset was stratified and randomly split into training (two-thirds) and test (one-third) partitions. On the held-out test set, the classifier achieved an accuracy of 74 %, substantially exceeding the average human face-classification rate of 62.5 %. This result indicates that a compact set of latent perceptual attributes is highly diagnostic of AI generation, reproducing the result found in Miller et al., 2023.

Upper-Bound Decoding Model Performance Curves

[9] Participants were not informed that the stimulus set contained both real and AI-generated faces.

Table 5.12: Latent perceptual attributes adopted from Miller et al., 2023

#	**Attribute**	**Brief description**
1	Age	Apparent facial age.
2	Alive in the eyes / uncanny valley	Eyes look “alive” vs. artificial.
3	Attractive	Overall attractiveness.
4	Distinctive / Average	Uniqueness of the face.
5	Expressive	Intensity of facial expression.
6	Eye Contact	Degree of direct gaze.
7	Familiar	Perceived familiarity.
8	Genuinely Happy	Authenticity of a happy expression.
9	Image Quality	Technical sharpness and clarity.
10	Congruent Lighting	Realistic, coherent lighting.
11	Memorable	Subjective memorability.
12	Proportional (holistic fit)	Harmony of facial proportions.
13	Smooth-Skinned	Smoothness / absence of blemishes.
14	Symmetrical	Left–right facial symmetry.

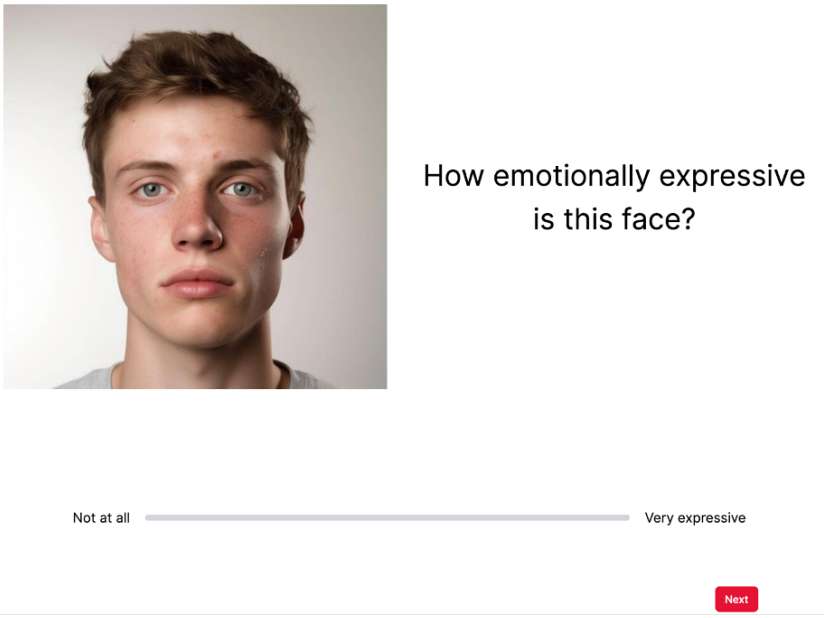


Figure 5.18: Schematic of the latent-variable rating task. Each participant rated three randomly assigned attributes (on a 0–100 scale) for all 200 images. Mean attribute scores served as inputs to subsequent behavioral and fMRI analyses.

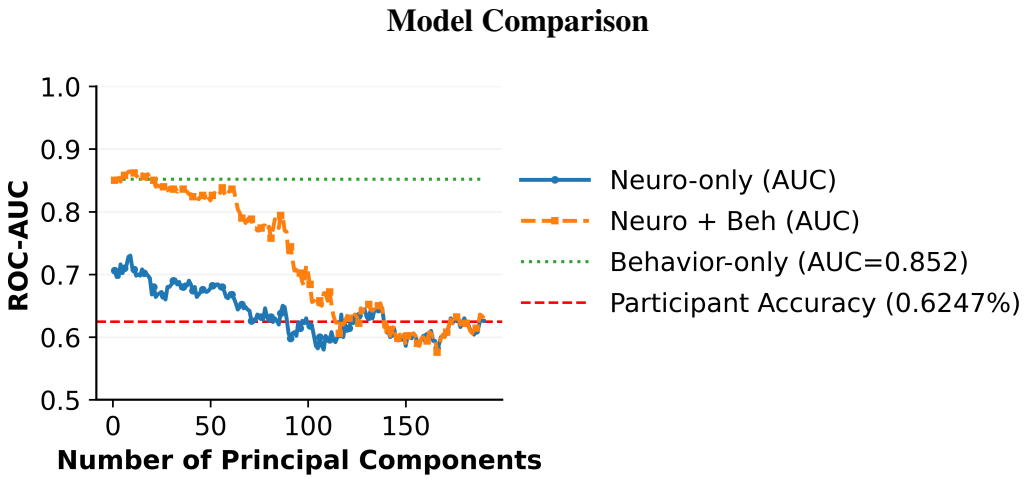


Figure 5.19: AUC scores by Principal Components, comparison between Neuro-Only, Neuro+Behavior, and Behavioral Only Models.

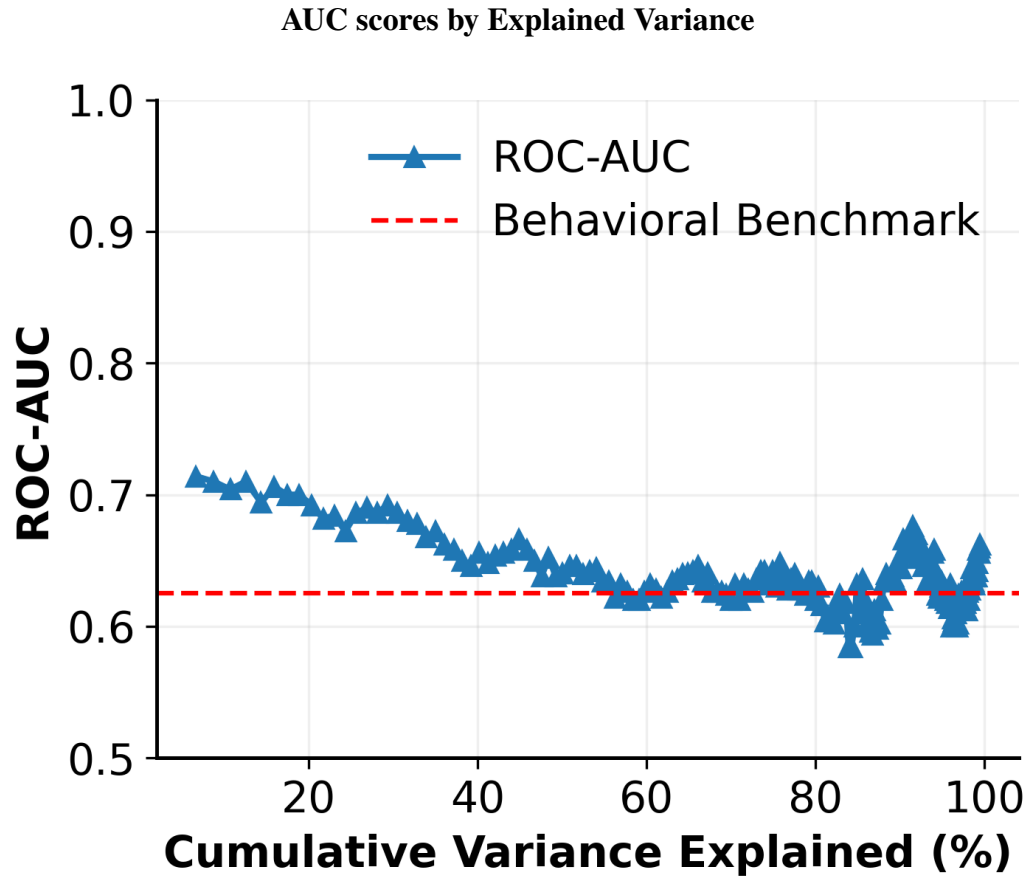


Figure 5.20: AUC scores by Explained Variance, horizontal reference line indicates behavioral score.

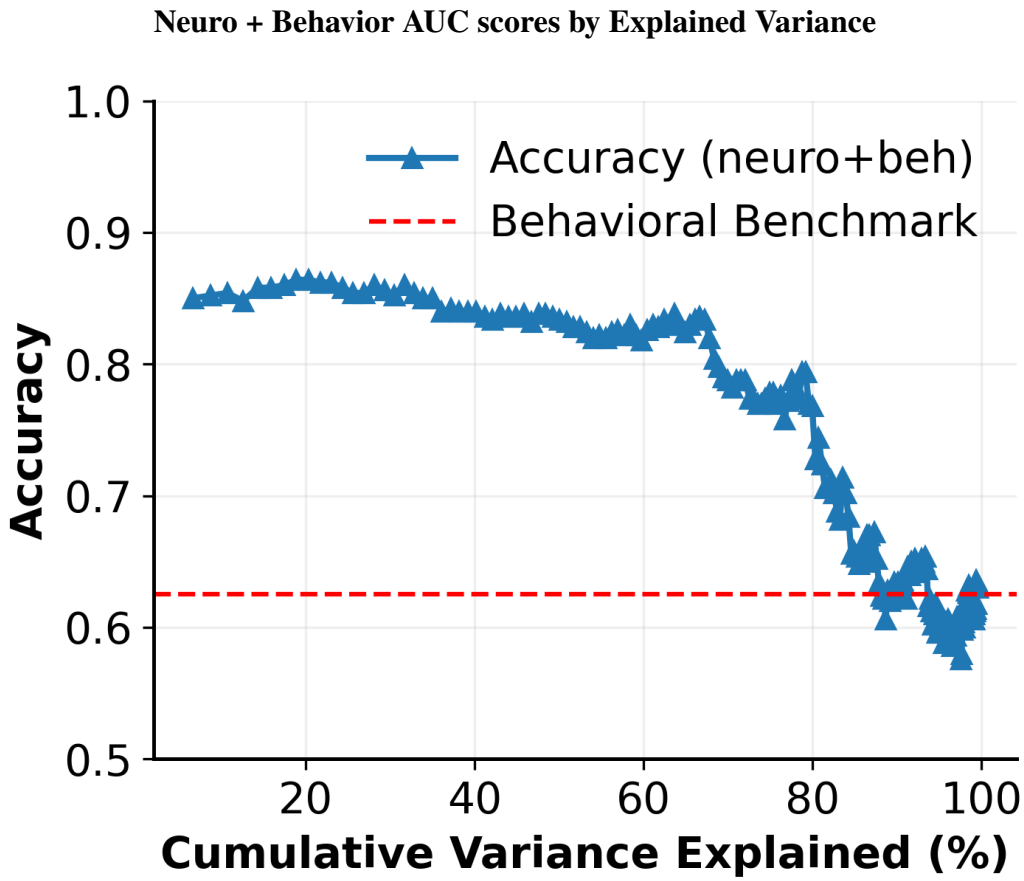


Figure 5.21: Neuro + Behavior AUC scores by Variance Explained, horizontal reference line indicates behavioral score.

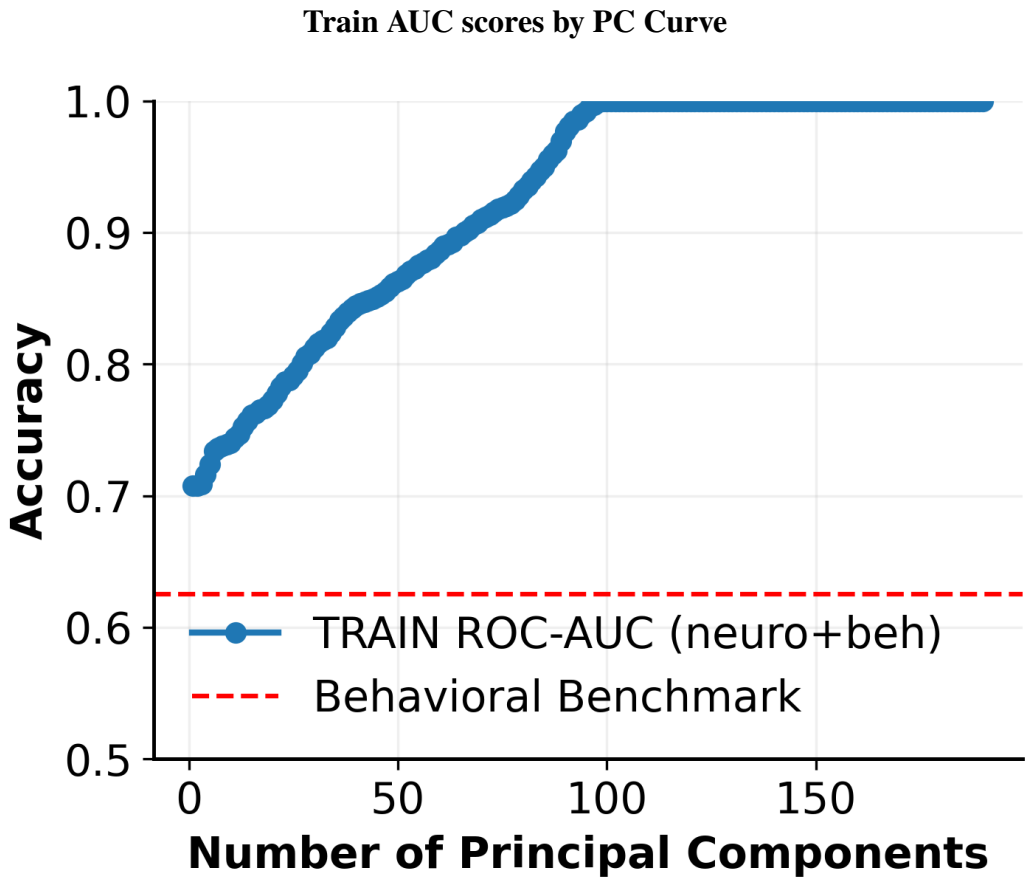


Figure 5.22: Training AUC scores by Principal Components, horizontal reference line indicates behavioral score; represents the AUC in the training folds.

Model Bin

Table with Model Accuracy

Table 5.13: Per-bin decoding performance compared to behavioral accuracy.

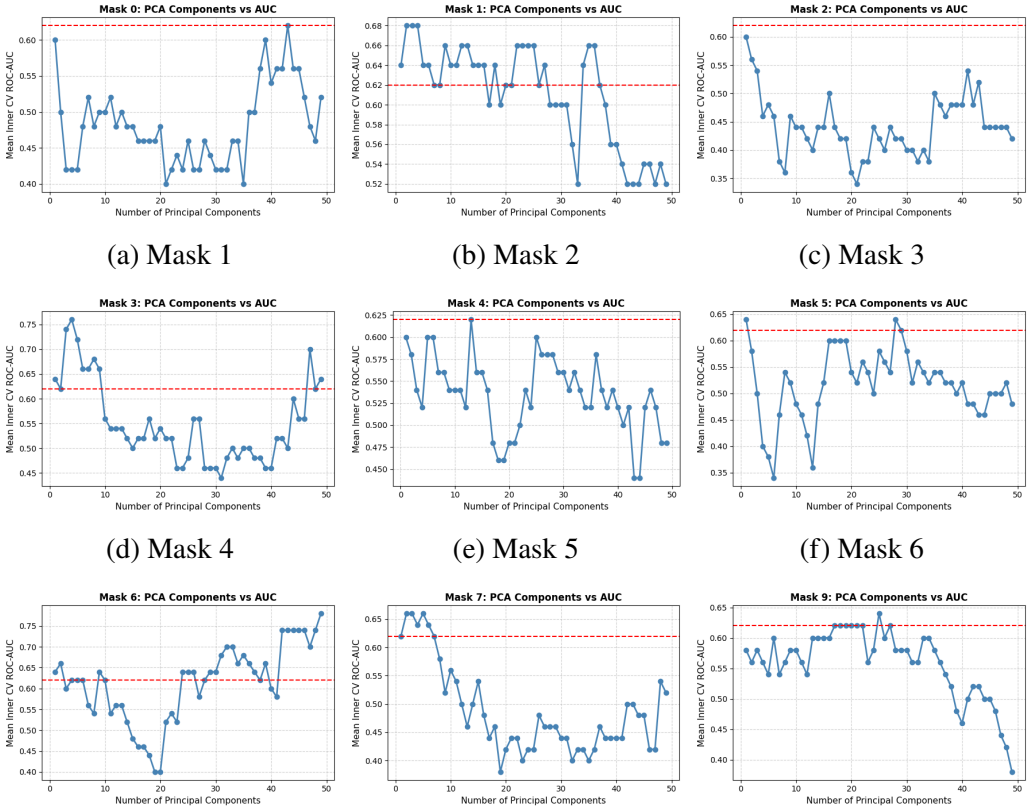
Bin	Model AUC	Model Accuracy	Bin Behavioral Accuracy
1	0.600	0.600	26.0
2	0.429	0.300	36.0
3	0.417	0.300	42.5
4	0.750	0.700	46.5
5	0.625	0.500	50.0
6	0.840	0.800	52.5
7	0.571	0.400	55.0
8	0.905	0.700	56.0
9	0.667	0.400	60.0
10	0.480	0.600	62.5
11	0.667	0.600	65.0
12	0.500	0.600	65.0
13	0.680	0.700	69.0
14	0.760	0.600	71.0
15	0.917	0.800	75.0
16	0.625	0.600	76.0
17	0.958	0.900	80.0
18	0.958	0.900	82.5
19	1.000	0.900	86.0
20	0.619	0.700	93.0

5.C PC Curve Plots, Within Subject Decoding

For each of these plots we used the same configuration described in section 5.A of the main text. The behavior was added as well to the same configuration.

Neuro-Only

50 Mask/ 50 Train Split



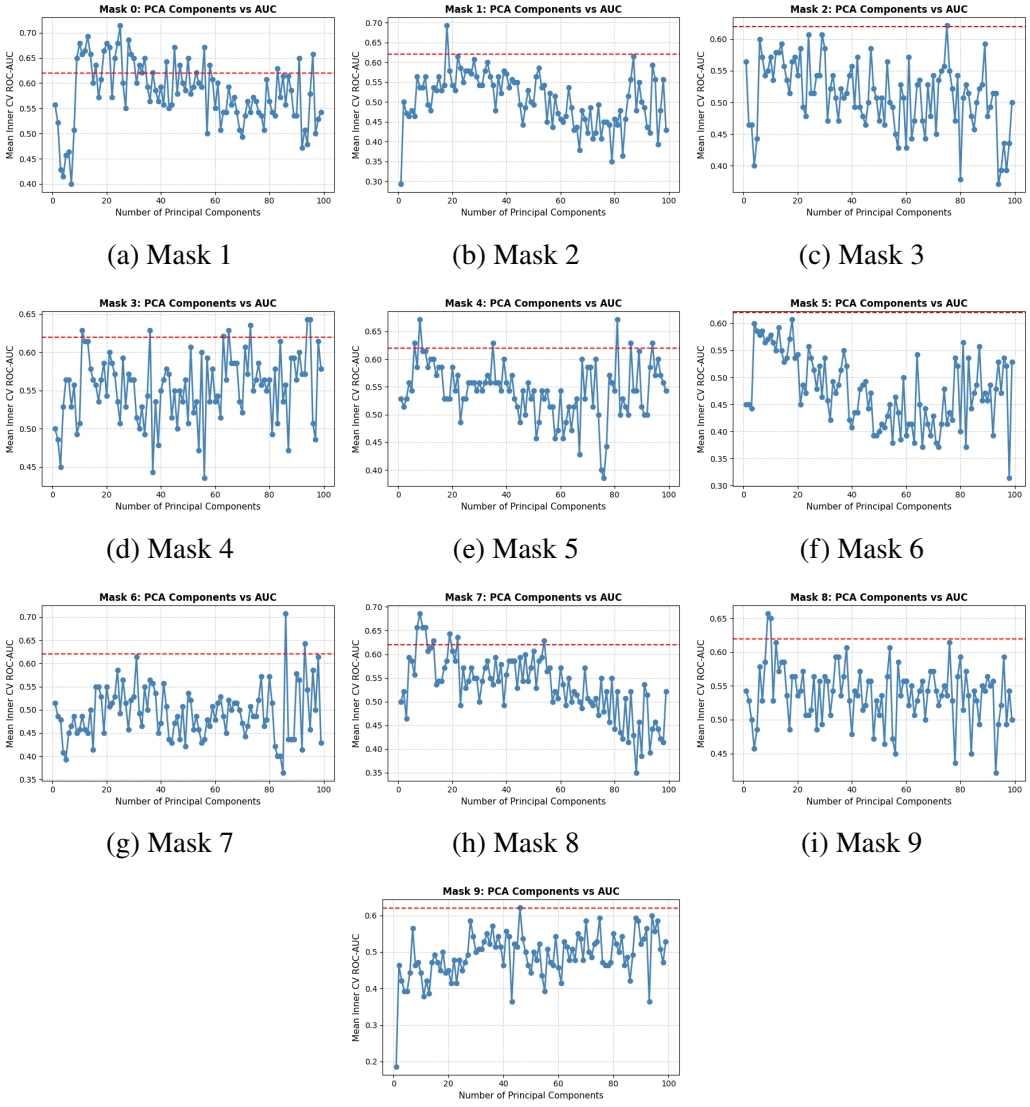
(a) Mask 1 (b) Mask 2 (c) Mask 3

(d) Mask 4 (e) Mask 5 (f) Mask 6

(g) Mask 7 (h) Mask 8 (i) Mask 10

Figure 5.23: PC decoding curves for the 50 Mask / 50 Train within-subject split. Each panel shows the decoding performance for each mask as principal component increases. Horizontal line represents average behavior accuracy across subjects.

30 Mask/ 70 Train Split



(a) Mask 1 (b) Mask 2 (c) Mask 3

(d) Mask 4 (e) Mask 5 (f) Mask 6

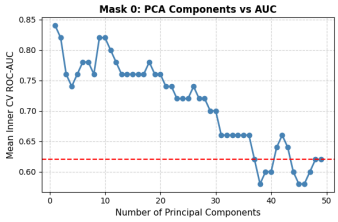
(g) Mask 7 (h) Mask 8 (i) Mask 9

(j) Mask 10

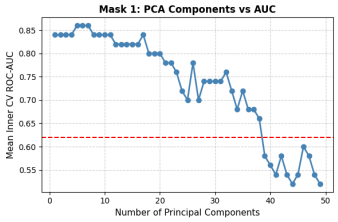
Figure 5.24: PC decoding curves for the 30 Mask / 70 train within-subject split. Each panel shows the decoding performance for each mask as principal component increases. Horizontal line represents average behavior accuracy across subjects.

Neuro and Behavior

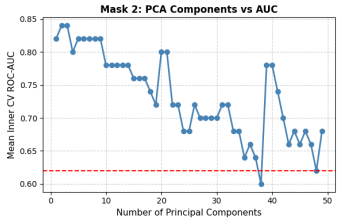
50 Mask/ 50 Train Split



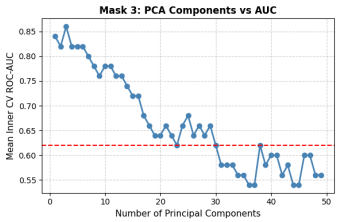
(a) Mask 1



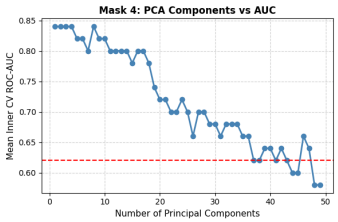
(b) Mask 2



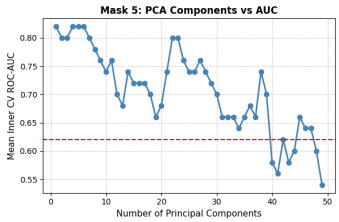
(c) Mask 3



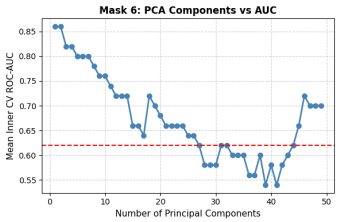
(d) Mask 4



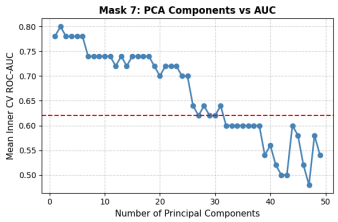
(e) Mask 5



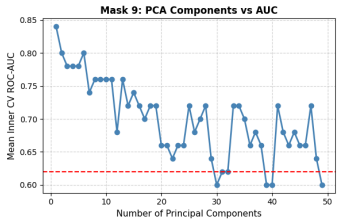
(f) Mask 6



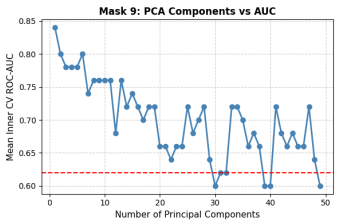
(g) Mask 7



(h) Mask 8



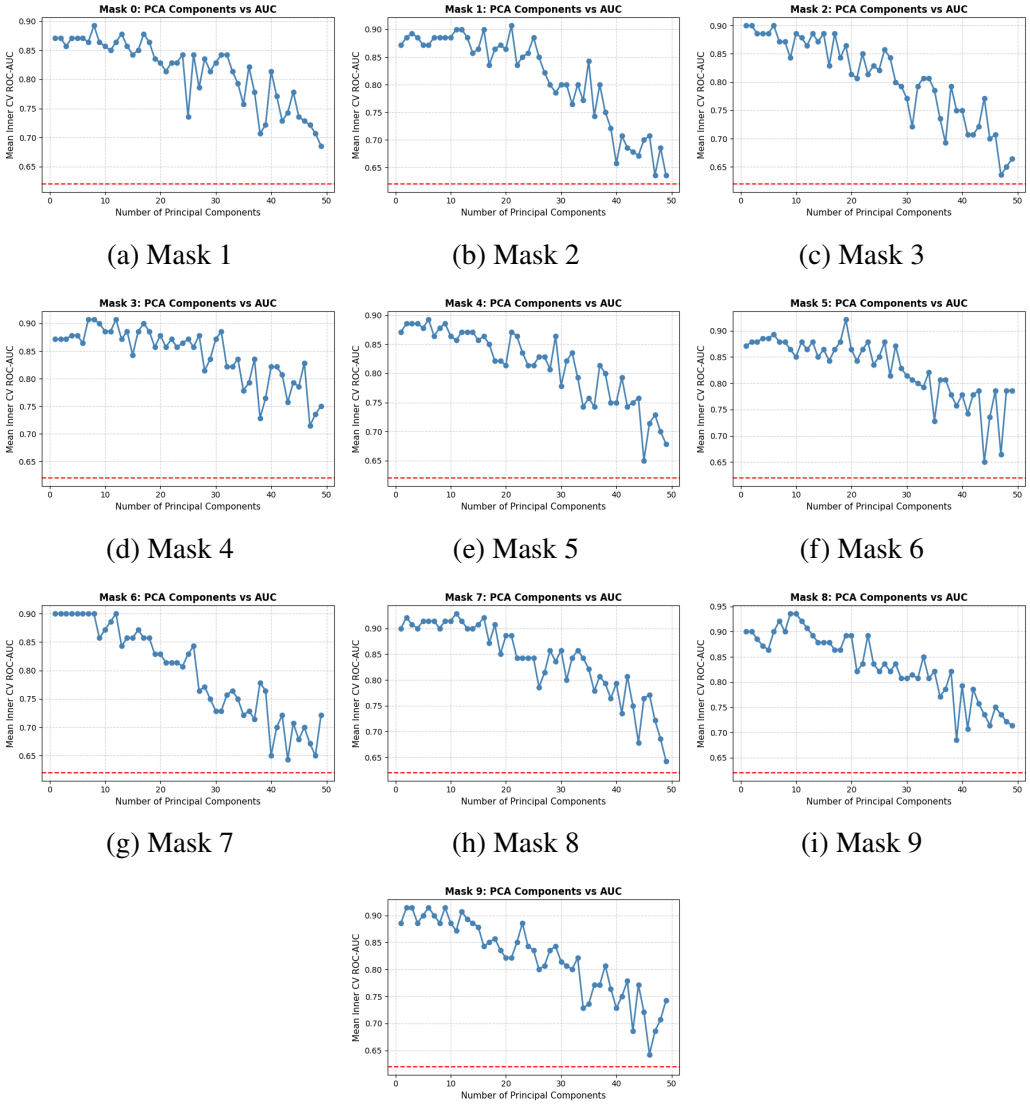
(i) Mask 9



(j) Mask 10

Figure 5.25: PC decoding curves for the 50 Mask / 50 Train within-subject split. Each panel shows the decoding performance for each mask as principal component increases. Each mask iteration is combined with behavioral data to show the combined predictive power of Neuro and Behavior. Horizontal line represents average behavior accuracy across subjects.

30 Mask/ 70 Train Split



(a) Mask 1 (b) Mask 2 (c) Mask 3

(d) Mask 4 (e) Mask 5 (f) Mask 6

(g) Mask 7 (h) Mask 8 (i) Mask 9

(j) Mask 10

Figure 5.26: PC decoding curves for the 30 Mask / 70 Train within-subject split. Each panel shows decoding performance across principal components for each mask using combined neural and behavioral features. Each mask iteration is combined with behavioral data to show the combined predictive power of Neuro and Behavior. Horizontal line represents average behavior accuracy across subjects.

References

Andrews-Hanna, Jessica R., Jonathan Smallwood, and R. Nathan Spreng (2014). "The Default Network and Self-Generated Thought: Component Processes, Dynamic Control, and Clinical Relevance". In: *Annals of the New York Academy of Sciences* 1316, pp. 29–52. DOI: 10.1111/nyas.12360.

Bammey, Quentin (2024). "Synthbuster: Towards Detection of Diffusion Model Generated Images". In: *IEEE Open Journal of Signal Processing* 5, pp. 1–9.

Bhagyashree, V. Kushwaha, and G. C. Nandi (2020). "Study of Prevention of Mode Collapse in Generative Adversarial Network (GAN)". In: Chennai, India: IEEE, pp. 1–6. DOI: 10.1109/CICT51604.2020.9312049. URL: https://ieeexplore.ieee.org/document/9312049.

Boksem, Maarten AS, Ale Smidts, and H Steven Scholte (2015). "Brain Responses to Movie Trailers Predict Individual Preferences for Movies and Their Population-Wide Commercial Success". In: *Journal of Marketing Research* 52.4, pp. 482–492.

Botvinick, Matthew M. (2007). "Conflict Monitoring and Decision Making: Reconciling Two Perspectives on Anterior Cingulate Function". In: *Cognitive, Affective, & Behavioral Neuroscience* 7.4, pp. 356–366. DOI: 10.3758/CABN.7.4.356.

Bryan, R., P. Perona, and R. Adolphs (2012). "Perspective distortion from interpersonal distance is an implicit visual cue for social judgments of faces". In: *PLoS ONE* 7.9, e45301. DOI: 10.1371/journal.pone.0045301.

Buckner, Randy L., Jessica R. Andrews-Hanna, and Daniel L. Schacter (2008). "The Brain's Default Network: Anatomy, Function, and Relevance to Disease". In: *Annals of the New York Academy of Sciences* 1124, pp. 1–38. DOI: 10.1196/annals.1440.011.

Burgess, Paul W., Iroise Dumontheil, and Sam J. Gilbert (2007). "The Gateway Hypothesis of Rostral Prefrontal Cortex (Area 10) Function". In: *Trends in Cognitive Sciences* 11.7, pp. 290–298. DOI: 10.1016/j.tics.2007.05.004.

Burgess, Paul W., Sam J. Gilbert, and Iroise Dumontheil (2007). "Function and Localization within Rostral Prefrontal Cortex (Area 10)". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 362.1481, pp. 887–899. DOI: 10.1098/rstb.2007.2095.

Cavanna, Andrea E. and Michael R. Trimble (2006). "The precuneus: A review of its functional anatomy and behavioural correlates". In: *Brain* 129.3, pp. 564–583.

Collins, Jessica A. and Ingrid R. Olson (2014). "Beyond the FFA: The Role of the Ventral Anterior Temporal Lobes in Face Processing". In: *Neuropsychologia* 61, pp. 65–79. DOI: 10.1016/j.neuropsychologia.2014.06.005.

Cornia, Marcella et al. (2018). "Predicting Human Eye Fixations via an LSTM-Based Saliency Attentive Model". In: *IEEE Transactions on Image Processing* 27.10, pp. 5142–5154. DOI: 10.1109/TIP.2018.2851672.

Corvi, Riccardo, Davide Cozzolino, Giovanni Poggi, et al. (2023). "Intriguing Properties of Synthetic Images: From Generative Adversarial Networks to Diffusion Models". In: *arXiv preprint arXiv:2304.06408*. URL: https://arxiv.org/abs/2304.06408.

Corvi, Riccardo, Davide Cozzolino, Giada Zingarini, et al. (2023). "On the Detection of Synthetic Images Generated by Diffusion Models". In: pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095167.

Cunnington, Ross et al. (2003). "The Preparation and Readiness for Voluntary Movement: A High-Field Event-Related fMRI Study of the Bereitschafts-BOLD Response". In: *NeuroImage* 20.1, pp. 404–412. DOI: 10.1016/S1053-8119(03)00291-X.

Esteban, Oscar et al. (2019). "fMRIPrep: a robust preprocessing pipeline for functional MRI". In: *Nature Methods* 16.1, pp. 111–116. DOI: 10.1038/s41592-018-0235-4. URL: https://www.nature.com/articles/s41592-018-0235-4.

Fahrenfort, Johannes J. et al. (2017). "Perceptual Integration Without Conscious Access". In: *Proceedings of the National Academy of Sciences* 114.14, pp. 3744–3749. DOI: 10.1073/pnas.1617268114. URL: https://www.pnas.org/doi/10.1073/pnas.1617268114.

Falk, Emily B, Carolyn Scholz, Matthew B O'Donnell, et al. (2017). "Brain signals can predict how often a news article is shared online". In: *Proceedings of the National Academy of Sciences* 114.31, pp. 8260–8265.

Fang, Zhiyong et al. (2016). "Unconscious Processing of Negative Animals and Objects: Role of the Amygdala Revealed by fMRI". In: *Frontiers in Human Neuroscience* 10, p. 146. DOI: 10.3389/fnhum.2016.00146. URL: https://www.frontiersin.org/articles/10.3389/fnhum.2016.00146/full.

Fleming, Stephen M. and Raymond J. Dolan (2012). "The Neural Basis of Metacognitive Ability". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1594, pp. 1338–1349. DOI: 10.1098/rstb.2011.0417.

Gaal, Simon van, Victor A. F. Lamme, and K. Richard Ridderinkhof (2014). "Unconscious High-Level Information Processing: Implication for Neurobiological Theories of Consciousness". In: *Neuroscientist* 20.5, pp. 435–448. DOI: 10.1177/1073858413495536. URL: https://journals.sagepub.com/doi/10.1177/1073858413495536.

Gomez-Ramirez, Manuel, Kristjana Hysaj, and Ernst Niebur (2016). "Neural Mechanisms of Selective Attention in the Somatosensory System". In: *Journal of Neurophysiology* 116.3, pp. 1218–1231. DOI: 10.1152/jn.00637.2015.

Goodfellow, Ian et al. (2014). “Generative Adversarial Nets”. In: vol. 27.

Grill-Spector, Kalanit et al. (2017). “The Functional Neuroanatomy of Human Face Perception”. In: *Annual Review of Vision Science* 3, pp. 167–196. DOI: 10.1146/annurev-vision-102016-061214.

Haxby, James V., Elizabeth A. Hoffman, and M. Ida Gobbini (2000). “The Distributed Human Neural System for Face Perception”. In: *Trends in Cognitive Sciences* 4.6, pp. 223–233. DOI: 10.1016/S1364-6613(00)01482-0.

Heekeren, Hauke R., Sean Marrett, and Leslie G. Ungerleider (2008). “The Neural Systems That Mediate Human Perceptual Decision Making”. In: *Nature Reviews Neuroscience* 9.6, pp. 467–479. DOI: 10.1038/nrn2374.

Heydt, Rüdiger von der (2023). “Visual cortical processing—From image to object representation”. In: *Frontiers in Computer Science* 5, p. 1136987.

Iidaka, Tetsuya (2014). “Role of the Fusiform Gyrus and Superior Temporal Sulcus in Face Perception and Recognition: An Empirical Review”. In: *Japanese Psychological Research* 56.1, pp. 33–45. DOI: 10.1111/jpr.12018.

Kanwisher, Nancy, Josh McDermott, and Marvin M. Chun (1997). “The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception”. In: *Journal of Neuroscience* 17.11, pp. 4302–4311. DOI: 10.1523/JNEUROSCI.17-11-04302.1997.

Kanwisher, Nancy and Galit Yovel (2006). “The Fusiform Face Area: A Cortical Region Specialized for the Perception of Faces”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 361.1476, pp. 2109–2128. DOI: 10.1098/rstb.2006.1934.

Knutson, Brian and Alexander Genevsky (2018). “Neuroforecasting aggregate choice”. In: *Current Directions in Psychological Science* 27.2, pp. 110–115.

Koechlin, Etienne and Alexandre Hyafil (2007). “Anterior prefrontal function and the limits of human decision-making”. In: *Science* 318.5850, pp. 594–598.

Kravitz, Dwight J. et al. (2013). “The ventral visual pathway: an expanded neural framework for the processing of object quality”. In: *Trends in Cognitive Sciences* 17.1, pp. 26–49. DOI: 10.1016/j.tics.2012.10.011. URL: https://doi.org/10.1016/j.tics.2012.10.011.

Leech, Robert and David J. Sharp (2014). “The role of the posterior cingulate cortex in cognition and disease”. In: *Brain* 137.1, pp. 12–32.

Lorenz, Peter, Ricard Durall, and Janis Keuper (2023). “Detecting Images Generated by Deep Diffusion Models Using Their Local Intrinsic Dimensionality”. In: URL: https://openaccess.thecvf.com/content/ICCV2023W/DFAD/papers/Lorenz_Detecting_Images_Generated_by_Deep_Diffusion_Models_Using_Their_Local_ICCVW_2023_paper.pdf.

Lu, Negar, Matthew Groh, et al. (2025). "Characterizing Photorealism and Artifacts in Diffusion Model-Generated Images". In: DOI: 10.1145/3706598.3713962.

Martin, Chris B. and Morgan D. Barense (2023). "Perception and Memory in the Ventral Visual Stream and Medial Temporal Lobe". In: *Annual Review of Vision Science* 9, pp. 409–434. DOI: 10.1146/annurev-vision-120222-014200.

Martin-Rodriguez, Francisco, Ramon Garcia-Mojon, and Monica Fernandez-Barciela (2023). "Detection of AI-Created Images Using Pixel-Wise Feature Extraction and Convolutional Neural Networks". In: *Sensors* 23.22, p. 9037. DOI: 10.3390/s23229037. URL: https://www.mdpi.com/1424-8220/23/22/9037.

Menon, Vinod and Lucina Q. Uddin (2010). "Saliency, Switching, Attention and Control: A Network Model of Insula Function". In: *Brain Structure and Function* 214.5–6, pp. 655–667. DOI: 10.1007/s00429-010-0262-0.

Miller, Emily J. et al. (2023). "AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones". In: *Psychological Science* 34.12, pp. 1390–1403. DOI: 10.1177/09567976231207095.

Mohr, Holger et al. (2015). "Sparse regularization techniques provide novel insights into outcome integration processes". In: *NeuroImage* 104, pp. 163–176. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2014.10.025. URL: https://www.sciencedirect.com/science/article/pii/S1053811914008490.

Mudrik, Liad and Leon Y. Deouell (2022). "Neuroscientific Evidence for Processing Without Awareness". In: *Annual Review of Neuroscience* 45, pp. 403–423. DOI: 10.1146/annurev-neuro-110920-033151. URL: https://www.annualreviews.org/doi/10.1146/annurev-neuro-110920-033151.

Mumford, Jeanette A. et al. (2012). "Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses". In: *NeuroImage* 59.3, pp. 2636–2643. DOI: 10.1016/j.neuroimage.2011.08.076.

Nachev, Parashkev, Christopher Kennard, and Masud Husain (2008). "Functional Role of the Supplementary and Pre-Supplementary Motor Areas". In: *Nature Reviews Neuroscience* 9.11, pp. 856–869. DOI: 10.1038/nrn2478.

Nightingale, Sophie J. and Hany Farid (2022). "AI-synthesized faces are indistinguishable from real faces and more trustworthy". In: *Proceedings of the National Academy of Sciences* 119.8, e2120481119. DOI: 10.1073/pnas.2120481119. URL: https://www.pnas.org/doi/10.1073/pnas.2120481119.

Ohbayashi, Makoto (2021). "The roles of the cortical motor areas in sequential movements". In: *Frontiers in Behavioral Neuroscience* 15, p. 640659.

Ojha, Utkarsh, Yuheng Li, and Yong Jae Lee (2024). "Towards Universal Fake Image Detectors That Generalize Across Generative Models". In.

Oquab, Maxime et al. (2023). "DINOv2: Learning Robust Visual Features without Supervision". In: *arXiv preprint arXiv:2304.07193*.

Philiastides, Marios G., Hauke R. Heekeren, and Paul Sajda (2014). "Human Scalp Potentials Reflect a Mixture of Decision-Related Signals during Perceptual Choices". In: *Journal of Neuroscience* 34.50, pp. 16877–16889. DOI: 10.1523/JNEUROSCI.3012-14.2014.

Picard, Nathalie and Peter L. Strick (1996). "Motor Areas of the Medial Wall: A Review of Their Location and Functional Activation". In: *Cerebral Cortex* 6.3, pp. 342–353. DOI: 10.1093/cercor/6.3.342.

Pitts, Michael A., Antígona Martínez, and Steven A. Hillyard (2012). "Isolating Neural Correlates of Conscious Perception from Neural Correlates of Reporting One's Perception". In: *Frontiers in Psychology* 3, p. 108. DOI: 10.3389/fpsyg.2012.00108. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00108/full.

Quiroga, Rodrigo Quian (2012). "Concept cells: the building blocks of declarative memory functions". In: *Nature Reviews Neuroscience* 13.8, pp. 587–597. DOI: 10.1038/nrn3251.

Rafique, Riffat et al. (2023). "Deep fake detection and classification using error-level analysis and deep learning". In: *Scientific Reports* 13, p. 7422. DOI: 10.1038/s41598-023-34629-3. URL: https://www.nature.com/articles/s41598-023-34629-3.

Ramnani, Narender and Adrian M. Owen (2004). "Anterior Prefrontal Cortex: Insights into Function from Anatomy and Neuroimaging". In: *Nature Reviews Neuroscience* 5.3, pp. 184–194. DOI: 10.1038/nrn1343.

Rana, M. S. et al. (2022). "Deepfake Detection: A Systematic Literature Review". In: *IEEE Access* 10, pp. 25494–25513. DOI: 10.1109/ACCESS.2022.3154404. URL: https://ieeexplore.ieee.org/document/9716516.

Ridderinkhof, K. Richard et al. (2004). "The Role of the Medial Frontal Cortex in Cognitive Control". In: *Science* 306.5695, pp. 443–447. DOI: 10.1126/science.1100301.

Schmahmann, Jeremy D. (2019). "The Cerebellum and Cognition". In: *Neuroscience Letters* 688, pp. 62–75. DOI: 10.1016/j.neulet.2018.07.005.

Schmälzle, Ralf et al. (2020). "The Effectiveness of Online Messages for Promoting Smoking Cessation Resources: Predicting Nationwide Campaign Effects From Neural Responses in the EX Campaign". In: *Frontiers in Human Neuroscience* 14, p. 565772.

Sheline, Yvette I. et al. (2009). "The Default Mode Network and Self-Referential Processes in Depression". In: *Proceedings of the National Academy of Sciences* 106.6, pp. 1942–1947. DOI: 10.1073/pnas.0812686106.

Shenhav, Amitai, Matthew M. Botvinick, and Jonathan D. Cohen (2013). "The Expected Value of Control: An Integrative Theory of Anterior Cingulate Cortex Function". In: *Neuron* 79.2, pp. 217–240. DOI: 10.1016/j.neuron.2013.07.007.

Smallwood, Jonathan et al. (2021). "The Default Mode Network in Cognition: A Topographical Perspective". In: *Nature Reviews Neuroscience* 22, pp. 503–513. DOI: 10.1038/s41583-021-00474-4.

Strick, Peter L., Richard P. Dum, and Julie A. Fiez (2009). "Cerebellum and Non-Motor Function". In: *Annual Review of Neuroscience* 32, pp. 413–434. DOI: 10.1146/annurev.neuro.31.060407.125606.

Sundar, Anirudh et al. (2024). "Improved Detection of Latent Diffusion-Generated Images Through Aligned Datasets". In: *arXiv preprint arXiv:2410.11835*. URL: https://arxiv.org/abs/2410.11835.

U.S. Department of Homeland Security (2021). *Increasing Threats of Deepfake Identities*. Tech. rep. Accessed May 18, 2025. U.S. Department of Homeland Security. URL: https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf.

Ullsperger, Markus and D. Yves von Cramon (2001). "Subprocesses of Performance Monitoring: A Dissociation of Error Processing and Response Competition Revealed by Event-Related fMRI and ERPs". In: *NeuroImage* 14.6, pp. 1387–1401. DOI: 10.1006/nimg.2001.0935.

Ullsperger, Markus, Claudia Danielmeier, and Gerhard Jocham (2014). "Neurophysiology of Performance Monitoring and Adaptive Behavior". In: *Physiological Reviews* 94.1, pp. 35–79. DOI: 10.1152/physrev.00041.2012.

Utevsky, Amanda V., David V. Smith, and Scott A. Huettel (2014). "Precuneus Is a Functional Core of the Default-Mode Network". In: *Journal of Neuroscience* 34.3, pp. 932–940. DOI: 10.1523/JNEUROSCI.4227-13.2014.

Vatansever, Deniz, David K. Menon, and Emmanuel A. Stamatakis (2017). "Default Mode Network Engagement Beyond Self-Referential Internal Mentation". In: *Brain Connectivity* 7.9, pp. 603–613. DOI: 10.1089/brain.2017.0489.

Vessel, Edward A., G. Gabrielle Starr, and Nava Rubin (2012). "The Brain on Art: Intense Aesthetic Experience Activates the Default Mode Network". In: *Frontiers in Human Neuroscience* 6, p. 66. DOI: 10.3389/fnhum.2012.00066.

Wang, Sheng-Yu et al. (2020). "CNN-Generated Images Are Surprisingly Easy to Spot... For Now". In: pp. 8695–8704.

Chapter 6

GENERAL DISCUSSION

This dissertation has examined human decision-making at the intersection of behavioral science, artificial intelligence, and neuroscience. Across five studies spanning experimental asset markets, large language model agents, online investor communities, and neuroimaging of synthetic media perception, a set of interconnected findings emerges about how cognition, emotion, and social context shape choices under uncertainty, and how AI both diverges from and illuminates human behavior.

6.1 Summary of Contributions

Chapter 2 provides the conceptual foundation for the dissertation. Rather than presenting a new empirical study, the chapter develops a framework for understanding how AI agents can enter social science research. It identifies three roles for AI systems: as interaction partners for human subjects, as synthetic substitutes for human participants, and as analytical tools for processing large-scale social data. The chapter also emphasizes a central caution that motivates the empirical work that follows: LLMs may appear human-like in some settings, but their behavior can diverge from human behavior in domain-specific and theoretically important ways.

This framing organizes the rest of the dissertation. Chapter 4 tests the substitution question directly by asking whether LLM agents reproduce human behavior in experimental asset markets. Chapter 5 examines the human side of the human-AI boundary by studying neural responses to AI-generated and real images. In this sense, the chapter functions as the dissertation's conceptual entry point: it argues that AI will expand the methodological frontier of social science, while also showing why those methods require empirical validation rather than uncritical adoption.

Chapter 3 investigated the relationship between physiological arousal, belief formation, and asset price bubbles in controlled experimental markets with a constant fundamental value. The main result is that electrodermal activity is not a strong period-by-period predictor of returns once standard market variables are included. Traditional measures, especially lagged returns and participant forecasts, explain short-horizon price dynamics substantially better than contemporaneous tonic arousal, and machine-learning models reach the same conclusion: adding

tonic EDA provides little incremental improvement over market and forecast variables alone.

The physiological signal does, however, appear to matter more at slower timescales and in predicting or explaining individual behavior. A five-period moving average of tonic EDA is associated with forward returns, suggesting that slower arousal trends may capture something closer to a market-level affective state (or "mood") than a momentary emotional reaction. At the individual level, EDA has modest predictive value for behavior: lower arousal variability is associated with higher earnings, and elevated arousal is associated with de-risking and revealed risk-preference measures. These results suggest that physiology is not best understood as a direct high-frequency price signal, but as a slower and noisier marker of market state, emotional regulation, and risk adjustment.

The chapter also provides counterevidence against a simple time-varying-risk-premia account of bubbles in this setting. The risk-elicitation measures do not explain return dynamics in the way such an account would predict, while forecasts and return momentum remain more informative. This null result is useful because it narrows the set of plausible mechanisms: bubbles in these markets appear less likely to arise from measured period-by-period shifts in risk preferences alone, and more likely from structured belief formation, extrapolative expectations, and feedback between market prices and trader behavior. In this sense, the chapter shows that experimental bubbles are driven by interacting behavioral mechanisms rather than by a single physiological or risk-preference variable.

Chapter 4 examined how LLM agents behave in the same experimental asset-market paradigm studied with human traders in Chapter 3. The chapter asked whether off-the-shelf LLM agents can be treated as substitutes for human participants in endogenous financial markets, and whether agent-only markets reproduce the bubble-crash dynamics that are robustly observed in human experiments. The answer was no. Across six frontier models, LLM markets traded closer to fundamental value, displayed less strategic heterogeneity, and produced substantially weaker bubble dynamics than human markets. Some models generated modest run-ups or erratic deviations, but none reproduced the magnitude, volatility, and crash structure characteristic of human trading.

This result is important not only as a comparison between humans and models, but as evidence about agent-only systems themselves. As LLM agents are increasingly deployed inside software products, automated workflows, financial tools, and multi-

agent protocols, their behavior in environments without direct human interaction becomes an object of study in its own right. The chapter shows that such systems can generate coherent market behavior, respond to dividend shocks, and adapt forecasts to changing conditions, but that their dynamics are not simply human behavior translated into artificial agents. Their strategies were more fundamental-value oriented, their forecasts were less biased, and their reasoning text showed less of the heuristic, heterogeneous, and momentum-driven behavior seen in human traders.

The chapter therefore makes two contributions to the dissertation. Substantively, it shows that the behavioral richness of human markets, belief heterogeneity, trend extrapolation, speculative momentum, and bubble-crash dynamics, does not automatically emerge from frontier LLM agents placed into the same experimental setting. Methodologically, it establishes experimental finance as a useful testbed for studying autonomous AI agents as economic actors. As agent-only systems become more common, controlled experiments of this kind will be necessary for understanding when LLM agents behave rationally, when they behave erratically, and when they produce genuinely emergent collective dynamics.

Chapter 5 examined whether fMRI responses contain information about the distinction between real and AI-generated faces. Whole-brain GLM analyses showed that image origin was associated with measurable differences in neural activity: real faces elicited greater activity in early visual cortex, whereas AI-generated faces elicited greater activity in regions including lingual gyrus, precuneus, cingulate cortex, and frontal pole. The perceived-realness contrast partially converged with this pattern, while the Correct versus Incorrect contrast produced a broader pattern consistent with task difficulty, uncertainty, effort, or response-related processing rather than a focal neural signature of successful AI detection.

The decoding analyses provided weaker but more direct evidence that these signals contained image-level information. An upper-bound MVPA analysis using circular feature selection achieved a mean ROC-AUC of approximately 69.8%, but this should be interpreted only as an optimistic benchmark. In the non-circular within-group analyses, training-derived masks produced ROC-AUC values ranging from approximately 0.58 to 0.64, with a mean AUC of 0.617 across usable masks. Some masks reached conventional or marginal significance, and the most consistent mask overlap appeared in precuneus/retrosplenial cortex, with additional overlap in cingulate, occipital fusiform, and right precuneus regions. However, performance varied

across masks, one split produced no surviving mask, and cross-group generalization was unreliable.

The chapter provides cautious evidence that neural responses carry information relevant to distinguishing real from AI-generated faces, but that this information is weak, distributed, and difficult to decode robustly. The results are therefore best interpreted as a proof of concept rather than evidence for a practical neuroimaging-based detector.

6.2 Limitations and Future Directions

Each chapter acknowledges study-specific limitations in detail; here I note broader limitations that apply across the dissertation.

The experimental markets in Chapters 3 and 4 use a specific constant-fundamental-value design. While this design reliably generates bubbles and simplifies the identification of mispricing, it represents one point in the space of possible market structures. Whether the psychobiological patterns documented here generalize to declining-fundamental designs, continuous double auctions, or markets with richer information structures remains an open question.

The biometric data in Chapter 3 were collected from a subset of in-lab participants (approximately 21% of the total sample), all of whom were Caltech students. This raises questions about external validity. The finding that in-lab participants significantly outperformed online participants suggests meaningful population differences that future work should address with larger and more representative biometric samples.

The LLM results in Chapter 4 are necessarily specific to the models tested (GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro, Mistral-Large, GPT-3.5, and Grok 2) at a particular point in time. As frontier models evolve rapidly, these results represent a snapshot rather than a permanent characterization. Future work should examine whether next-generation models with improved reasoning capabilities exhibit more human-like trading behavior, and whether agent prompt optimization can bridge the gap between LLM and human market dynamics.

The neuroimaging results in Chapter 5 are constrained by modest sample sizes and a specific stimulus set of Midjourney-generated faces. The fragility of cross-group generalization suggests that current decoding approaches may overfit to sample-specific features. Scaling to larger cohorts, diverse generators, and preregistered

analytic pipelines will be essential for establishing whether the neural signatures observed here are robust and generalizable.

6.3 Concluding Remarks

The studies in this dissertation converge on a view of human decision-making as fundamentally shaped by the interaction of cognitive processes, emotional states, and social environments, and as fundamentally distinct from the decision-making of artificial agents. Asset bubbles arise not from individual irrationality alone but from the collective dynamics of heterogeneous beliefs, emotional arousal, and adaptive expectation formation. AI agents trained on the same information and facing the same incentives produce qualitatively different market outcomes, highlighting the irreducibly human character of speculative behavior. Social platforms create new channels through which non-pecuniary incentives can shape risk-taking and disclosure. And the human perceptual system, while imperfect, retains a measurable sensitivity to the authenticity of visual stimuli even as synthetic media grow more convincing.

As artificial intelligence becomes increasingly embedded in financial markets, social platforms, and information environments, understanding these human-AI differences is not merely an academic exercise. It is essential for designing markets, institutions, and technologies that account for how people actually think, feel, and decide.