Financial Stability Implications of Generative AI: Taming the Animal Spirits

Anne Lundgaard Hansen[a,b] and Seung Jung Lee[a]

[a]Board of Governors of the Federal Reserve System [b]Federal Reserve Bank of Richmond

November 24, 2025 | 4th CEMLA/Dallas Fed Financial Stability Workshop

Introduction

- Generative AI is reshaping workflows across institutional and individual traders.

What could be the implications for financial stability?

Introduction

- Generative AI is reshaping workflows across institutional and individual traders.

What could be the implications for financial stability?

- Previous work has emphasized that AI-powered trading could lead to more correlated behaviors, e.g., stemming from:
  - Collusion (Dou et al., 2025).
  - Market concentration (Bank of England Financial Policy Committee, 2025).
  - Model monoculture (Danielsson & Uthemann, 2024; Financial Stability Board, 2024).

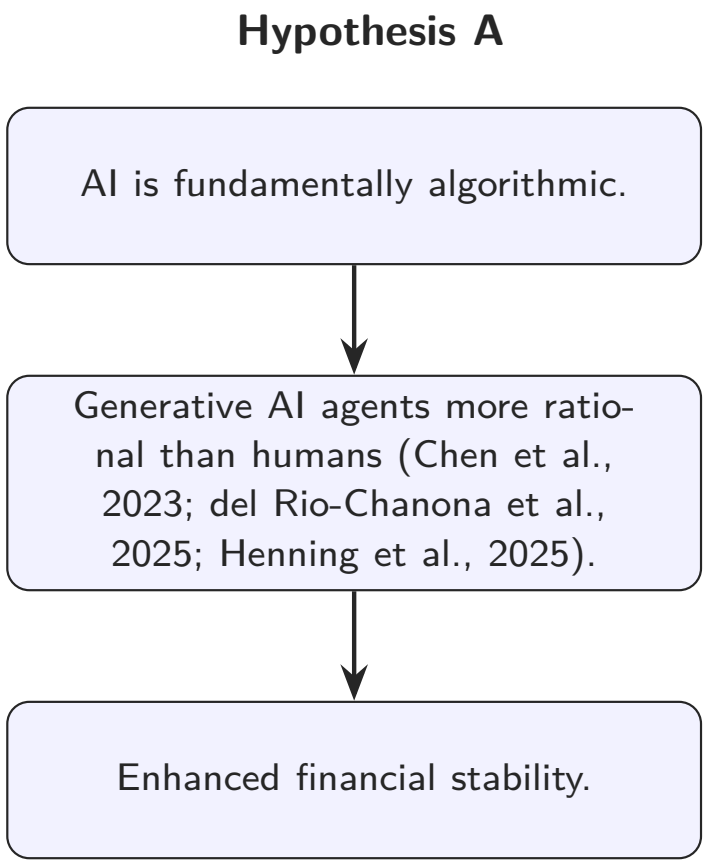
Introduction

- Generative AI is reshaping workflows across institutional and individual traders.

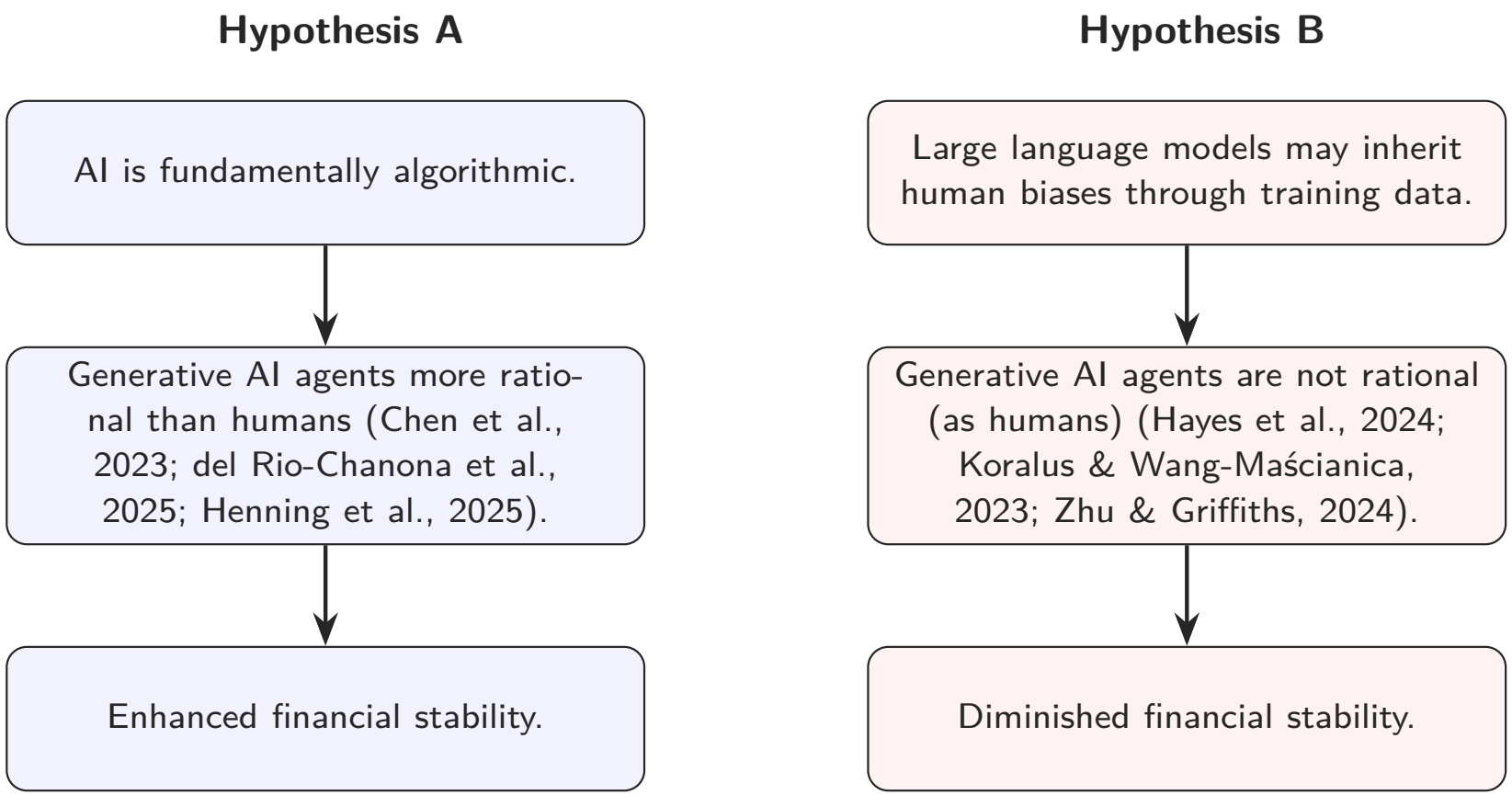
What could be the implications for financial stability?

- Previous work has emphasized that AI-powered trading could lead to more correlated behaviors, e.g., stemming from:
  - Collusion (Dou et al., 2025).
  - Market concentration (Bank of England Financial Policy Committee, 2025).
  - Model monoculture (Danielsson & Uthemann, 2024; Financial Stability Board, 2024).
- We focus on another source of correlation: Herding behavior.
  - Herding = disregarding private information to follow market trends.
  - IMF outreach: Herding was among the top cited risks of generative AI adoption in capital markets (International Monetary Fund, 2024).

Two competing hypotheses



Two competing hypotheses



Methodology in a nutshell

- We conduct laboratory-style experiments to detect herding behavior within large language models (LLMs).
  - Compare AI decisions with human decisions from identical experiment conducted with financial market professionals (Cipriani & Guarino, 2009).
  - Zoom in on differences in how humans and AI make decisions, in a controlled setting.

Methodology in a nutshell

- We conduct laboratory-style experiments to detect herding behavior within large language models (LLMs).
  - Compare AI decisions with human decisions from identical experiment conducted with financial market professionals (Cipriani & Guarino, 2009).
  - Zoom in on differences in how humans and AI make decisions, in a controlled setting.
- Build upon two strands of literature:
  - Behavioral econonomics: Studies human decisions in controlled laboratory settings to provide microfoundations.
  - Agentic AI: LLMs can be treated as agents that can be studied like humans (Horton, 2023).

The Experiment

The model

- Experiment based on the Avery and Zemsky (1998) model:
  - Financial market with one risky asset traded sequently over discrete periods ($T = 8$).
  - At the outset, $v = 50$, but an information event may occur in which case $v \in \{0, 100\}$.
  - If an information event occurred, informed traders receive private signals on the value: White signal: $v = 100$ (blue signal: $v = 0$) with 70% probability.

The model

- Experiment based on the Avery and Zemsky (1998) model:
  - Financial market with one risky asset traded sequently over discrete periods ($T = 8$).
  - At the outset, $v = 50$, but an information event may occur in which case $v \in \{0, 100\}$.
  - If an information event occurred, informed traders receive private signals on the value: White signal: $v = 100$ (blue signal: $v = 0$) with 70% probability.
- Question:
  - Do informed traders make decisions based on their private signals (*rational* behavior)?
  - Or do they act the same regardless of the signal they receive (*cascade* behavior)?

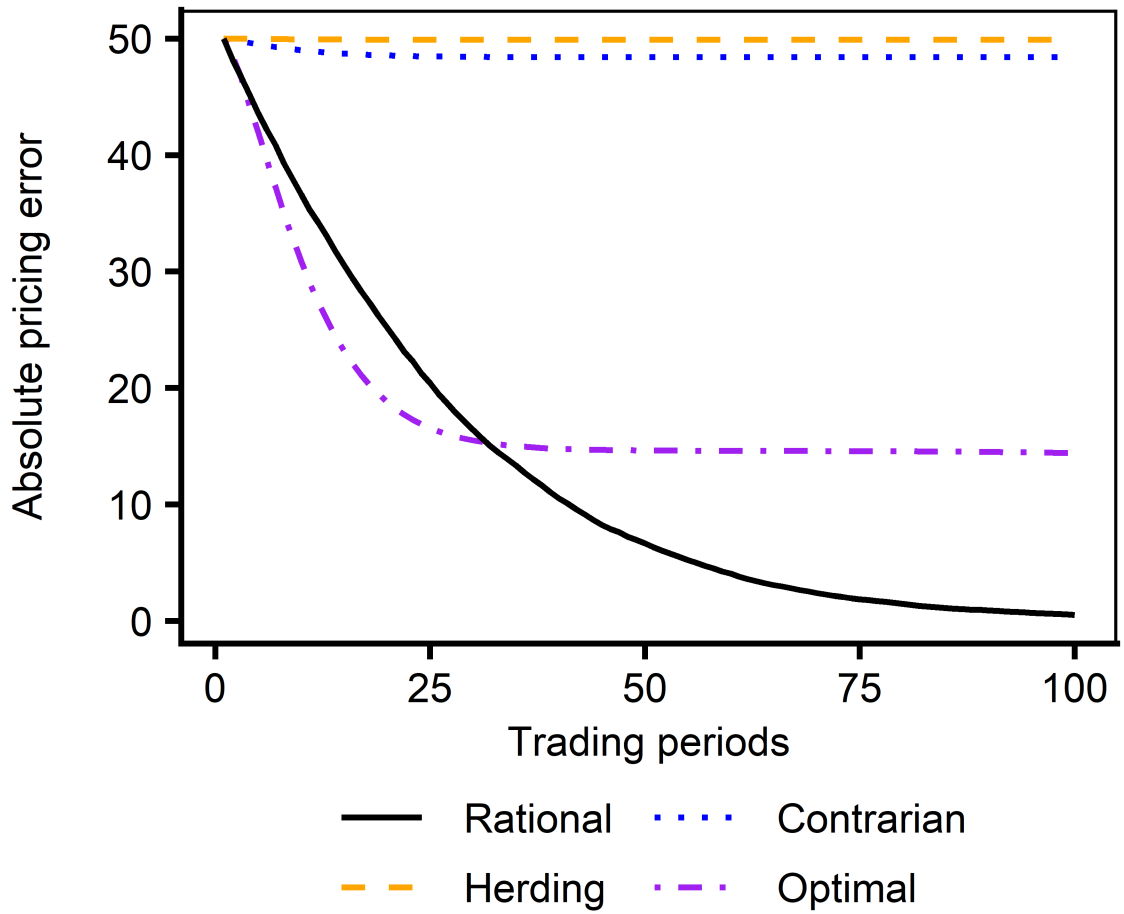
Types of decisions

- **Rational**: Buy on white signal, sell on blue signal.
- **Partial rational**: Rational behavior on one signal, no trade on the other signal.
- **Cascade trading**: Same action (buy or sell) regardless of the signal.
  - Herding: Action follows the market (majority action in the trading history).
  - Contrarian behavior: Action goes against the market.
  - Undetermined: Zero trade imbalance.
- **Cascade no trading**: No trade regardless of the signal.
- **Error**: Buy on blue signal, sell on white signal.

Types of decisions

- **Rational**: Buy on white signal, sell on blue signal.
- **Partial rational**: Rational behavior on one signal, no trade on the other signal.
- **Cascade trading**: Same action (buy or sell) regardless of the signal.
  - Herding: Action follows the market (majority action in the trading history). Appendix
  - Contrarian behavior: Action goes against the market.
  - Undetermined: Zero trade imbalance.
- **Cascade no trading**: No trade regardless of the signal.
- **Error**: Buy on blue signal, sell on white signal.

Types of decisions and financial stability



Laboratory setups

- Set up AI laboratory to mimic human laboratory from Cipriani and Guarino (2009) as closely as possible.

Laboratory setups

- Set up AI laboratory to mimic human laboratory from Cipriani and Guarino (2009) as closely as possible.
- **Participants:** 32 AI agents simulated as the average response of 4 LLMs (Claude 3.7 Sonnet with extended thinking, Claude 3.5 Sonnet, Llama 3 70B, Nova Pro) with temperature 0.7.

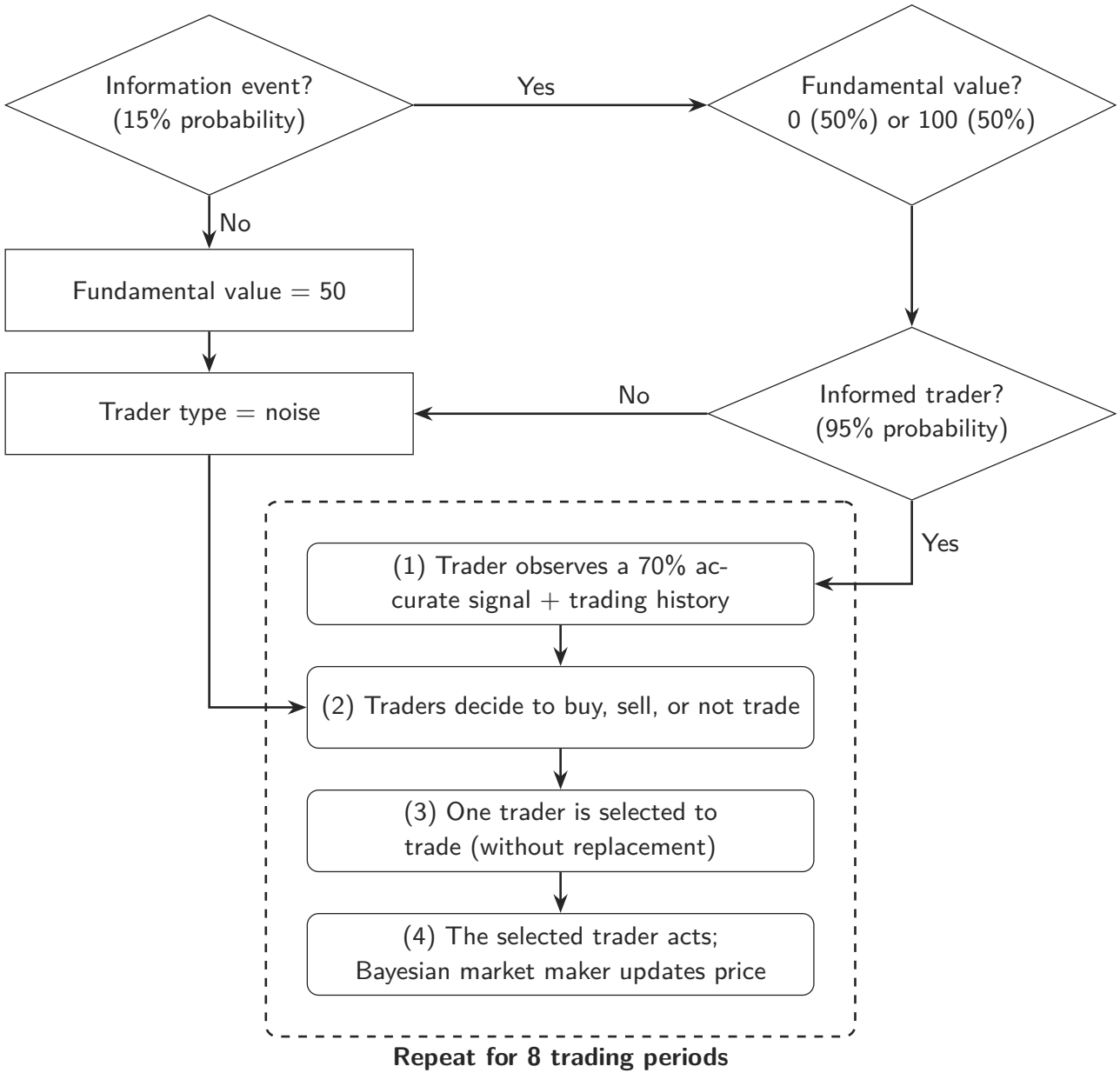
Laboratory setups

- Set up AI laboratory to mimic human laboratory from Cipriani and Guarino (2009) as closely as possible.
- **Participants:** 32 AI agents simulated as the average response of 4 LLMs (Claude 3.7 Sonnet with extended thinking, Claude 3.5 Sonnet, Llama 3 70B, Nova Pro) with temperature 0.7.
- **Setup:** 4 sessions with 8 trading rounds.

Laboratory setups

- Set up AI laboratory to mimic human laboratory from Cipriani and Guarino (2009) as closely as possible.
- **Participants:** 32 AI agents simulated as the average response of 4 LLMs (Claude 3.7 Sonnet with extended thinking, Claude 3.5 Sonnet, Llama 3 70B, Nova Pro) with temperature 0.7.
- **Setup:** 4 sessions with 8 trading rounds.
- **Instructions:** Written instructions through prompting.
  - **System prompt:** Initial instructions.
  - **User prompt:** Information for each trading period (price, trading history, agent-specific memory) and requests (trading decisions for each signal + reasoning).

Flow diagram



Results

AI vs. human decisions

	Human	AI
Rational	50.90%	97.36%
Partial Rational	20.10%	2.64%
Cascade Trading	12.00%	0.00%
Optimal Herding	+	0.00%
Suboptimal Herding	+	0.00%
Contrarian	+	0.00%
Undetermined	+	0.00%
Cascade No Trading	16.50%	0.00%
Error	0.05%	0.00%
Optimal Herding Opportunities	+	36.56%

Human: Results from Cipriani and Guarino (2009). AI: Average decisions of experiments run with four LLMs.

AI vs. human decisions

	Human	AI	Claude 3.7	Claude 3.5	Llama 3	Nova Pro
Rational	50.90%	97.36%	100.00%	100.00%	100.00%	89.45%
Partial Rational	20.10%	2.64%	0.00%	0.00%	0.00%	10.55%
Cascade Trading	12.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Optimal Herding	+	0.00%	0.00%	0.00%	0.00%	0.00%
Suboptimal Herding	+	0.00%	0.00%	0.00%	0.00%	0.00%
Contrarian	+	0.00%	0.00%	0.00%	0.00%	0.00%
Undetermined	+	0.00%	0.00%	0.00%	0.00%	0.00%
Cascade No Trading	16.50%	0.00%	0.00%	0.00%	0.00%	0.00%
Error	0.05%	0.00%	0.00%	0.00%	0.00%	0.00%
Optimal Herding Opportunities	+	36.56%	30.61%	46.88%	21.88%	46.88%

Human: Results from Cipriani and Guarino (2009). AI: Average decisions of experiments run with four LLMs.

AI reasoning: Why no optimal herding?

- Models fail to acknowledge the trading history when forming expected values.
- **Example**: Session 2, $t = 7$, $h_7 = \{\text{buy}, \text{buy}, \text{sell}, \text{no trade}, \text{buy}, \text{buy}\}$, $p_7 = 62$, Claude 3.7:
  - Reasoning: *"With a White signal, the expected value is 70 (70% chance of 100, 30% chance of 0), which exceeds the current price of 61.77, giving an expected profit of about 8.23 lire from buying. With a Blue signal, the expected value is 30 (30% chance of 100, 70% chance of 0), which is below the current price of 61.77, giving an expected profit of about 31.77 lire from selling."*
  - The AI makes rational decisions because:

$$\mathbb{E}(v|s_t = \text{white}) = 70 > p_t = 62 > \mathbb{E}(v|s_t = \text{blue}) = 30.$$

  - But, it is optimal to buy regardless of signal (herd), because:

$$\mathbb{E}(v|h_t, s_t = \text{white}) = 96 > \mathbb{E}(v|h_t, s_t = \text{blue}) = 83 > p_t = 62.$$

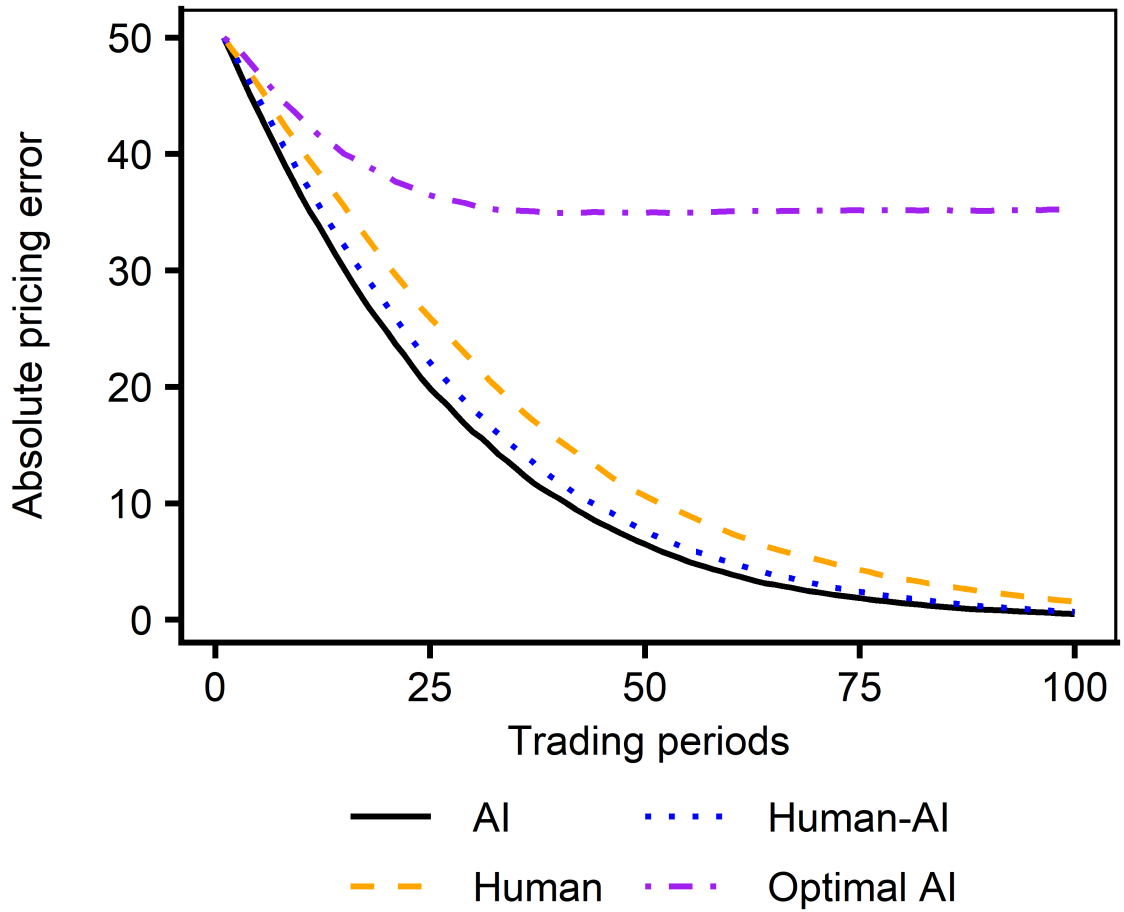
Appendix: Reasoning analysis

Prompting AI to make optimal decisions

	Human	AI	Optimal AI
Rational	50.90%	97.36%	18.65%
Partial Rational	20.10%	2.64%	21.88%
Cascade Trading	12.00%	0.00%	59.48%
Optimal Herding	+	0.00%	47.43%
Suboptimal Herding	+	0.00%	0.00%
Contrarian	+	0.00%	6.60%
Undetermined	+	0.00%	5.44%
Cascade No Trading	16.50%	0.00%	0.00%
Error	0.05%	0.00%	0.00%
Optimal Herding Opportunities	+	36.56%	81.52%

Appendix: Expected payoffs

Simulated implications for financial stability



Exploring variations to the experiment

- Conducting experiments in the human lab is expensive $\Rightarrow$ infeasible to explore many variations.
- The AI lab is a lot cheaper and available from the couch - let's explore some options!
  - **Temperature**: Robustness to model temperature. Temperature
  - **Payoff structure**: How are LLMs incentivized by "pay"? Payoffs
  - **Personality profiles**: Can AI agents role play generate different results? Profiles
  - **Experiment length**: What happens if the experiment is run over longer periods or more sessions? Length
  - **Signals**: Are LLMs truly rational, or do they respond differently to different signal colors?

Relabeling signal color codes

- Signal colors matter when using counterintuitive coding: AI agents are not purely rational!

	Good: Green, Bad: Red	Good: Red, Bad: Green
Rational	98.54%	50.78%
Partial Rational	1.46%	11.72%
Cascade Trading	0.00%	12.50%
Optimal Herding	0.00%	7.32%
Suboptimal Herding	0.00%	1.56%
Contrarian	0.00%	0.00%
Undetermined	0.00%	3.61%
Cascade No Trading	0.00%	0.00%
Error	0.00%	25.00%
Optimal Herding Opportunities	52.94%	42.93%

Concluding Remarks

Financial stability implications of generative AI

- **Reduced herding:** AI-influenced trading may make markets less prone to self-reinforcing cycles $\rightarrow$ Fewer herding-driven asset price bubbles.

- **Diversified market responses:** AI's reliance on private information may introduce greater heterogeneity in market reactions to new information $\rightarrow$ Further reduce market correlation.

Some caveats

(1) If AI agents can be successfully instructed to engage in optimal herding, financial stability implications are more nuanced.

Some caveats

(1) If AI agents can be successfully instructed to engage in optimal herding, financial stability implications are more nuanced.

(2) When exposed to counter-intuitive information, AI exhibits surprisingly human-like behavior.

Some caveats

(1) If AI agents can be successfully instructed to engage in optimal herding, financial stability implications are more nuanced.

(2) When exposed to counter-intuitive information, AI exhibits surprisingly human-like behavior.

(3) Interaction of AI and humans could create new market dynamics with unpredictable outcomes.

Some caveats

(1) If AI agents can be successfully instructed to engage in optimal herding, financial stability implications are more nuanced.

(2) When exposed to counter-intuitive information, AI exhibits surprisingly human-like behavior.

(3) Interaction of AI and humans could create new market dynamics with unpredictable outcomes.

(4) Findings based on today's LLMs may not fully predict the behavior of future generations of financial AI.

Thank you!

Bibliography

Avery, C., & Zemsky, P. (1998).Multidimensional Uncertainty and Herd Behavior in Financial Markets. *American Economic Review*, *88*(4), 724–748.

Bank of England Financial Policy Committee. (2025). *Financial Stability in Focus: Artificial Intelligence in the Financial System* (tech. rep.). Bank of England.

Chen, Y., Liu, T. X., Shan, Y., & Zhong, S. (2023).The Emergence of Economic Rationality of GPT. *Proceedings of the National Academy of Sciences (PNAS)*, *120*(51), 1–9.

Cipriani, M., & Guarino, A. (2009).Herd Behavior in Financial Markets: An Experiment With Financial Market Professionals. *Journal of the European Economic Association*, *7*(1), 206–233.

Danielsson, J., & Uthemann, A. (2024).AI Financial Crises. *VoxEU.org*.

del Rio-Chanona, R. M., Pangallo, M., & Hommes, C. (2025).Can Generative AI Agents Behave Like Humans? Evidence From Laboratory Market Experiments. *arXiv Preprint arXiv:2505.07457v1*.

Dou, W. W., Goldstein, I., & Ji, Y. (2025).AI-Powered Trading, Algorithmic Collusion, and Price Efficiency. *SSRN Working Paper*.

Financial Stability Board. (2024). *The Financial Stability Implications of Artificial Intelligence* (tech. rep.). Financial Stbility Board.

Hayes, W. M., Yax, N., & Palminteri, S. (2024).Relative Value Biases in Large Language Models. *arXiv Preprint arXiv:2401.14530*.

Henning, T., Ojha, S. M., Spoon, R., Han, J., & Camerer, C. F. (2025).LLM Trading: Analysis of LLM Agent Behavior in Experimental Asset Markets. *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*.

Horton, J. J. (2023).Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? *NBER Working Paper, No. 31122*.

Appendix A: Additional Details

Optimal decision-making

- Herding *can* be (but is not always) optimal (= profit-maximizing) in this model.
  - Informed traders know that the trading history comes from another informed trader with 95% probability.
  - Market maker never receives signals and thinks history reflects informed trades with 14% ($= 95\% \cdot 15\%$) probability.
  - The price is therefore updated more conservatively than trader's expectations.

  $\Rightarrow$ Investors may earn profits from herding by exploiting the build-up of other investors' private information in the trading history.

Back

Appendix B: Analysis of LLM Reasoning

Analyzing LLM reasoning

- Analyze the reasoning provided by the AI agents for each decision (baseline).
- Two approaches: LDA and Claude 3.7.
- LDA: Test 2-5 topics and identify 3 unique topics.
- Claude 3.7: Prompt the model to read each sentence of reasoning and ask:
  - Question 1: Is the trader comparing the price to the expected fundamental value of the asset? (True/False).
  - Question 2: Is the expected value computed using only the signal accuracy and the signal, e.g., 0.7*100+0*0.3=70 or 0.7*0+0.3*100=30? (True/False).
  - Question 3: Does the trader consider the market trend or the trading history in their reasoning? (True/False).
  - Question 4: How does the trader characterize the attractiveness of the investment?
  - Question 5: On a scale from 0-100 (where 100 represents purely emotional and 0 represents purely rational or logical), how much is the investor driven by emotions in their assessment?

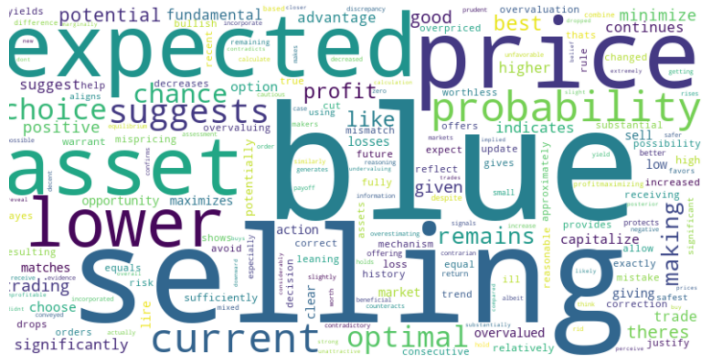
Back

LDA results

	AI	Claude 3.7	Claude 3.5	Llama 3	Nova Pro
Topic 0	51.93%	72.26%	50.00%	0.20%	94.73%
Topic 1	21.27%	27.74%	50.00%	9.57%	0.78%
Topic 2	26.80%	0.00%	0.00%	90.23%	4.49%

Topic 0

Topic 1



Topic 2

Back

Claude 3.7 results

	AI	Claude 3.7	Claude 3.5	Llama 3
Question 1: Price expected value comparison?	99.01%	100.00%	100.00%	100.00%
Question 2: Expected value given signal only?	63.09%	99.27%	99.41%	4.49%
Question 3: Consider market trends?	9.50%	0.00%	0.00%	30.66%
Question 4: Attractiveness of investment				
VERY ATTRACTIVE	1.88%	4.01%	2.93%	1.56%
ATTRACTIVE	69.39%	68.61%	45.31%	83.79%
REASONABLE	6.08%	9.85%	2.93%	4.88%
LESS ATTRACTABLE	3.65%	4.38%	0.59%	5.66%
NO INCENTIVE	19.01%	13.14%	48.24%	4.10%
Question 5: Rate on a scale from 0 (logic) to 100 (emotional)				
Mean	4.93%	0.13%	0.06%	12.72%
Bottom decile	0.00%	0.00%	0.00%	5.00%
Median	0.00%	0.00%	0.00%	10.00%
Top decile	15.00%	0.00%	0.00%	20.00%

Back

Appendix C: Additional Results

Expected payoffs

	Treatment I		Treatment II		Treatment III	
	AI	Optimal AI	AI	Optimal AI	AI	Optimal AI
Mean	2.57	2.72	3.80	14.95	5.07	7.79
Median	2.74	2.74	6.67	19.53	6.67	11.49
Min	-6.67	-6.67	-11.44	-28.28	-16.19	-16.19
Max	6.67	6.67	11.55	28.35	16.46	16.63
Std Dev	3.90	3.57	6.47	14.20	8.83	7.87

Back

Robustness to model temperature

	T=0.0	T=0.7 (baseline)	T=1.0
Rational	97.27%	97.36%	88.48%
Partial Rational	2.73%	2.64%	11.52%
Cascade Trading	0.00%	0.00%	0.00%
Optimal Herding	0.00%	0.00%	0.00%
Suboptimal Herding	0.00%	0.00%	0.00%
Contrarian	0.00%	0.00%	0.00%
Undetermined	0.00%	0.00%	0.00%
Cascade No Trading	0.00%	0.00%	0.00%
Error	0.00%	0.00%	0.00%
Optimal Herding Opportunities	41.25%	36.56%	45.15%

Back

Payoff structure

	0 GBP per lire	1M GBP per lire	3 lire per USD
Rational	97.27%	95.21%	97.07%
Partial Rational	2.73%	3.91%	2.93%
Cascade Trading	0.00%	0.88%	0.00%
Optimal Herding	0.00%	0.39%	0.00%
Suboptimal Herding	0.00%	0.00%	0.00%
Contrarian	0.00%	0.00%	0.00%
Undetermined	0.00%	0.49%	0.00%
Cascade No Trading	0.00%	0.00%	0.00%
Error	0.00%	0.00%	0.00%
Optimal Herding Opportunities	39.04%	34.49%	43.50%

Back

Personality profiles

	Human	Professional Trader	Robo-Advisor	Rational	C&G Characteristi
Rational	89.68%	67.30%	54.21%	59.22%	59.35%
Partial Rational	7.69%	29.51%	37.41%	30.66%	31.30%
Cascade Trading	2.63%	2.32%	5.93%	7.88%	9.35%
Optimal Herding	0.00%	0.00%	0.00%	0.00%	0.00%
Suboptimal Herding	0.00%	0.00%	0.00%	0.00%	0.00%
Contrarian	2.63%	2.32%	5.93%	7.88%	9.35%
Undetermined	0.00%	0.00%	0.00%	0.00%	0.00%
Cascade No Trading	0.00%	0.88%	2.44%	2.25%	0.00%
Error	0.00%	0.00%	0.00%	0.00%	0.00%
Optimal Herding Opportunities	0.00%	0.00%	0.00%	0.00%	0.00%

Back

Experiment length

	Baseline (4 sessions of 8 rounds)	10 sessions of 8 rounds	4 sessions of 20 rounds
Rational	97.36%	89.43%	94.45%
Partial Rational	2.64%	6.48%	5.55%
Cascade Trading	0.00%	4.04%	0.00%
Optimal Herding	0.00%	0.33%	0.00%
Suboptimal Herding	0.00%	0.00%	0.00%
Contrarian	0.00%	3.67%	0.00%
Undetermined	0.00%	0.04%	0.00%
Cascade No Trading	0.00%	0.03%	0.00%
Error	0.00%	0.02%	0.00%
Optimal Herding Opportunities	36.56%	65.73%	37.19%

Back