Ex Machina: Financial Stability in the Age of Artificial Intelligence

Kartik Anand[1] Sophia Kazinnik[2] Agnese Leonello[3] Ettore Panetti[4]

[1] Deutsche Bundesbank [2] Stanford University [3] European Central Bank [4] University of Naples Federico II

This paper represents the authors' personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank, the European Central Bank or the Eurosystem.

Role of AI in financial decision-making is expanding rapidly

- Release of R1 by DeepSeek has spurred an AI arms race in China to automate processing of market data and the generation of trading signals (Reuters, 2025)
- AI-enabled applications that autonomously generate financial advice will be the leading source for investors by 2027 (MIT Sloan, 2024; Deloitte, 2024)
- Fully agentic AI, capable of executing tasks, adapting to new information, and coordinating raises further prospects for automation (Nvidia, 2025)
- **Progression** towards more **sophisticated autonomous AI** raises **key questions:**
  - How will AI-agents behave in environments with strategic and economic uncertainty?
  - What are the implications for financial stability?

We study how AI-agents behave in a stylized model of financial fragility

- Canonical model of mutual fund redemptions (Chen et al., 2010)
  - Parsimonious variant of Diamond and Dybvig (1983)
  - Characterize equilibria under different assumptions regarding the environment
- **Simulation-based experimental setup:** Test model with **AI-based investors**
  1. Q-learning (QL)-based investors: Update strategies via reinforcement-learning
  2. LLM-based investors: Perform context-based inference via chain-of-thought reasoning
- Our choices are intended to capture and contrast **how AI-based investors learn** vs. **how they reason**, and the consequences of these for **financial stability**

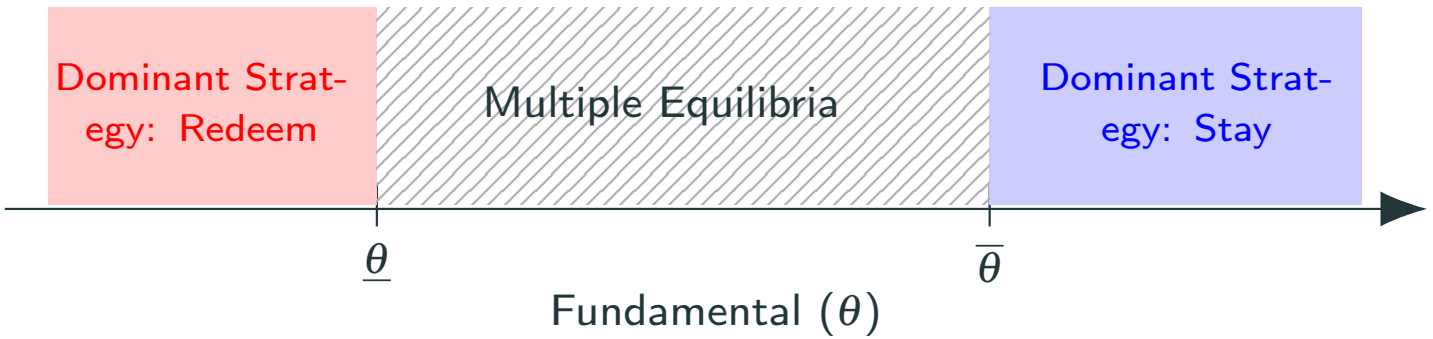
Results in a nutshell

- **Reinforcement-learning** encourages **coordination** between AI-based investors
  - But limited experimentation biases them to redeeming → exacerbates fragility
- **Chain-of-though reasoning** allows AI-based investors to **interpret the environment** → avoid redemption bias
  - But, belief heterogeneity renders the outcomes less predictable
- **Take-away:** AI architecture (learning vs reasoning) matters for financial stability

Model of mutual fund redemptions

Model setup

- N risk-neutral investors in a mutual fund choose **redeem** or **stay**
- The mutual fund services redemptions by **liquidating assets**, which is costly $\rightarrow$ reduces the available funds to re-invest and, thus, returns for those who stayed
- Each investor's **best response** depends on their **beliefs about others' actions**



- Equilibrium outcomes unchanged by introducing **mutual fund default risk**
- **Fragility:** excess redemptions relative to first-best, i.e., redeeming when $\theta > \underline{\theta}$

Testable predictions

Prediction

Without fundamental uncertainty, all investors: (i) redeem when $\theta < \underline{\theta}$ and (ii) stay when $\theta > \overline{\theta}$. In the intermediate range, both "all redeem" and "all stay" are equilibria

Prediction

Mutual fund default risk is irrelevant for investors' redemption decisions

Experimental Setup

QL-based investors

- Q-learning is a **reinforcement learning paradigm**
  - Investors learns by interacting with the environment
  - Receives rewards or penalties
  - Individually maximize their cumulative reward through repeated interactions
- Q(quality)-functions that map investors' actions to rewards (Watkins, 1989)
- QL-based investors balance **exploration** vs. **exploitation**
  - Exploration: trying different actions to learn Q-values
  - Exploitation: choosing actions with highest known Q-values
- **Key assumption:** Exploration rate decreases over time – training is costly

LLM-based investors

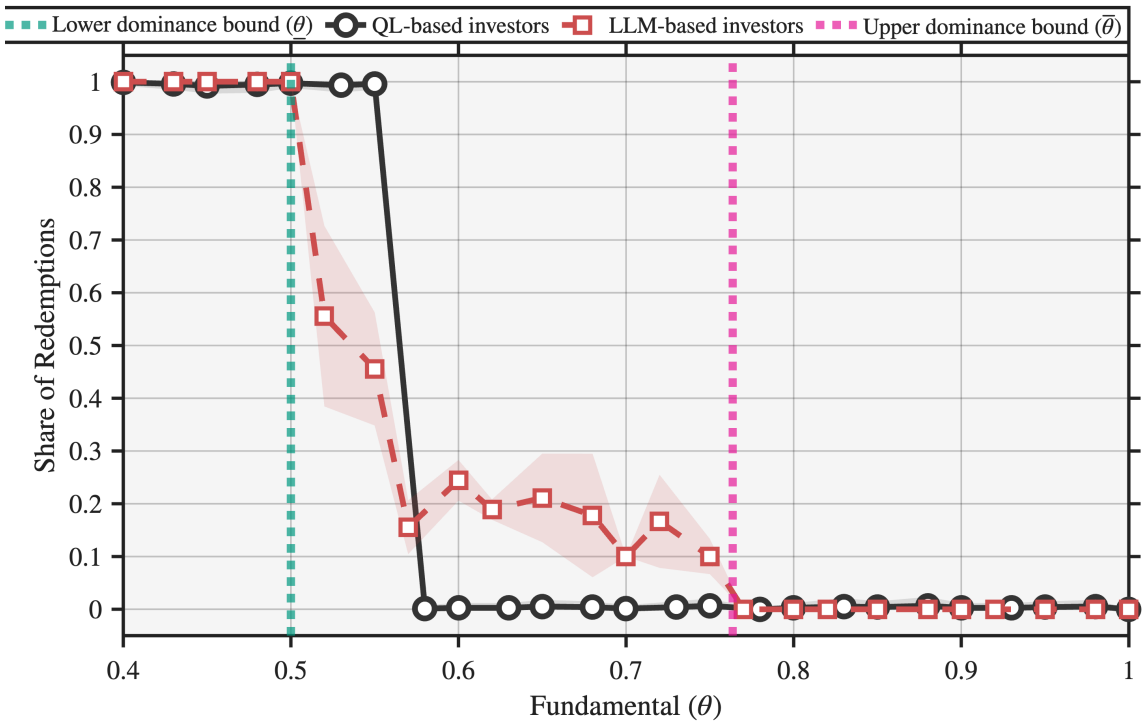
- **LLMs** are sophisticated **next-word prediction** engines
  - Use context to figure out the meaning of prompts and predict the next-word(s)
  - Draw on patterns learned from corpora of words, structures & relations
  - Can thinking out loud to breakdown and solve problems
- Experimental setup
  1. Each LLM-based investor is given a prompt that summarizes the economic problem
  2. We record their decisions (redeem or stay) and all reasoning tokens

◂ Inspect prompts

Experimental Results

Prediction 1: Coordination and multiple equilibria

Testing for coordination and multiple equilibria

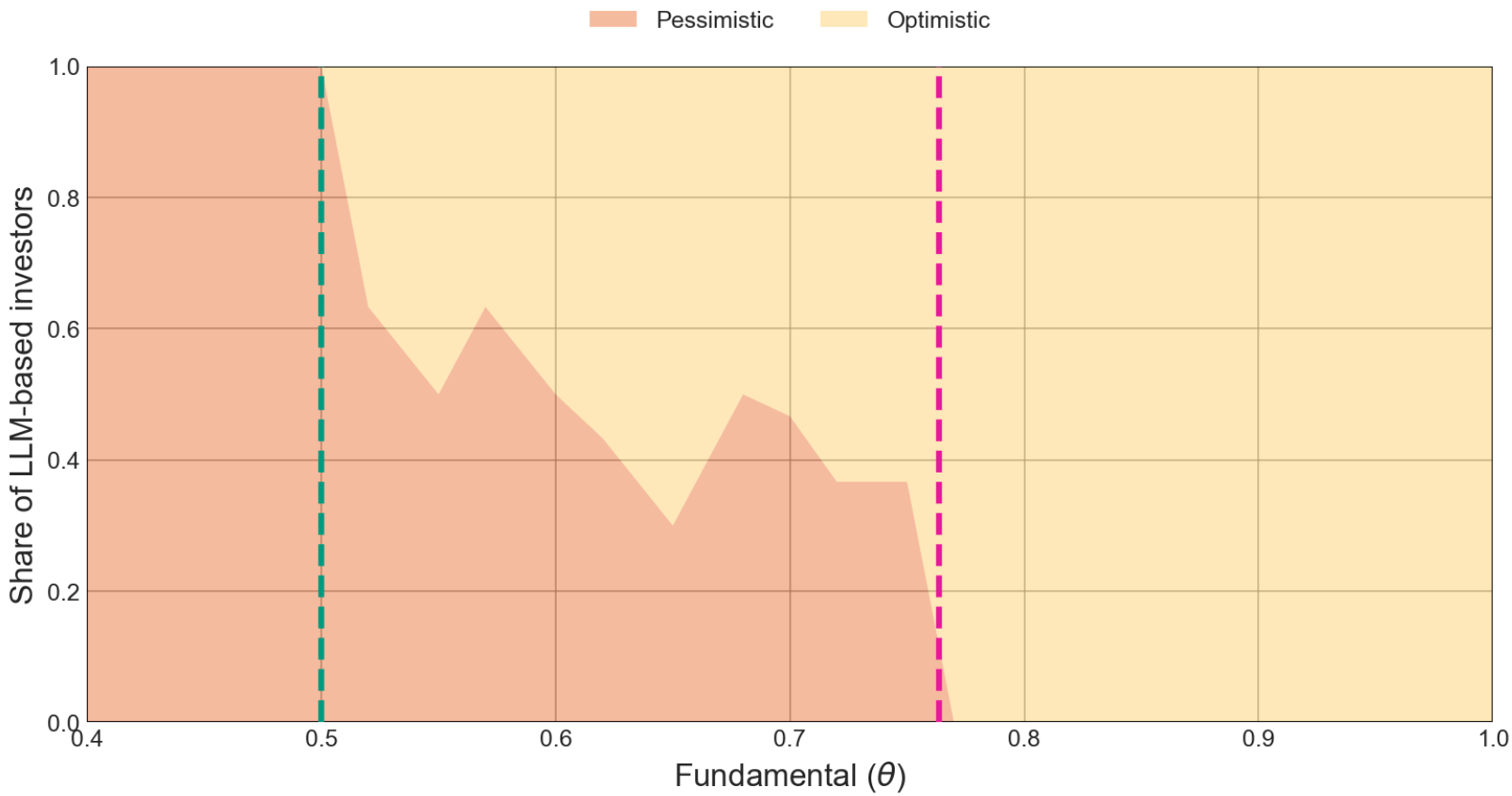


- In the **dominance regions**, both types of AI-based investors **coordinate on the Nash equilibria predicted by theory**
- **QL-based investors** converge to a **common switching threshold**, $\tilde{\theta}$ – investors acts as if all others play at random, and best-responds (**Level-K Reasoning**)
- **LLM-based investors** struggle to coordinate in the intermediate region – **outcomes are less predictable**

Classifying LLM-based investors' beliefs

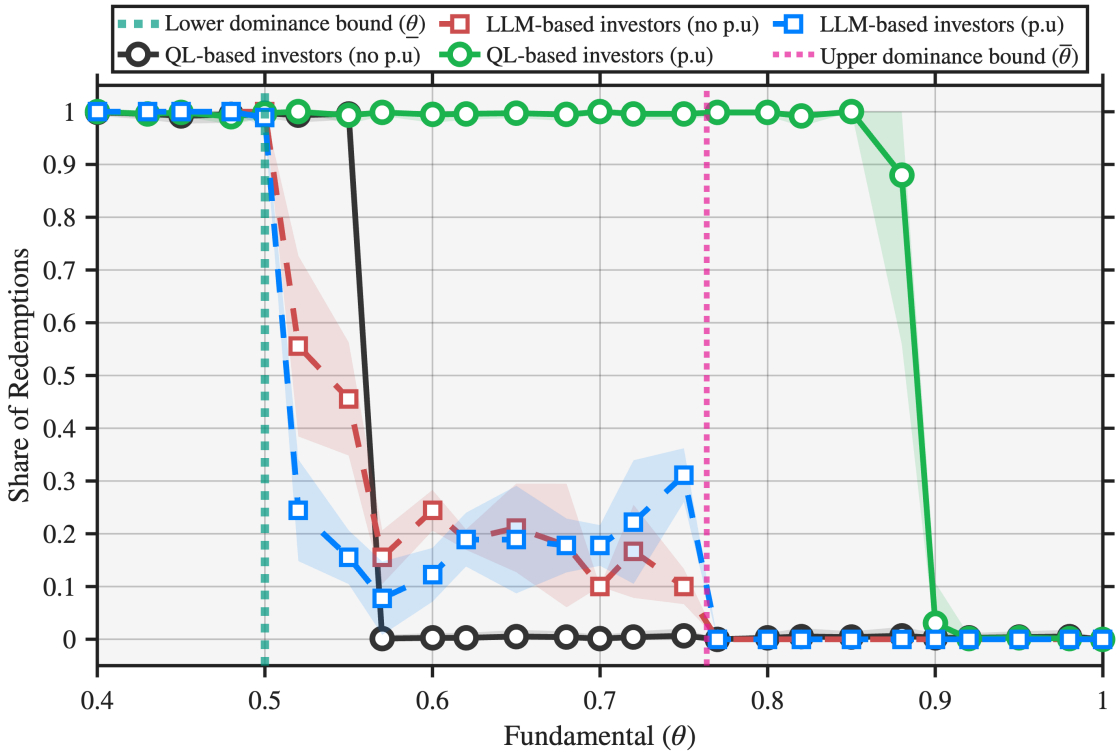
- **Reasoning tokens** contain information on the **beliefs of LLM-based investors**
- We use a LLM to **extract these beliefs** in a **theory-consistent way**
- **Classification** follows the **canonical structure** of coordination models:
  - Pessimistic belief, i.e., all others redeem
  - Optimistic belief, i.e., all others stay

Example for the heterogenity in LLM-based investors' beliefs



Prediction 2: Irrelevance of mutual fund default risk

Testing for the irrelevance of mutual fund default risk



- **LLM-based investors** remain largely uneffected by **mutual fund default risk**
- **QL-based investors** experience a strong **bias towards redeeming**

Understanding LLM-based investors

- **Analyze reasoning tokens** to **classify decisions** as being based on:
  - Expected value
  - Risk-aversion
  - Ambiguity-aversion
- Results suggest that
  - Risk aversion and ambiguity aversion play no role
  - Expected value calculations dominate the reasoning
- **LLM-based investors** behave **risk-neutral**, though they were not prompted to be!

Rationalizing the behavior of QL-based investors: The Hot Stove Effect

- The payoff from **staying** is **uncertain**
  - Sometimes the fund performs well → positive return
  - Sometimes the fund defaults → zero return
- **Redeeming** has a **fixed** and **certain payoff**
- Under **reinforcement learning** investors **learn from each experience**
  - Every zero-return experience makes staying seem worse → investors become pessimistic about the value of staying
- **Self-confirming bias:** a redeeming investor gets a safe payoff and never learns the fundamental → error is not corrected and is exacerbated by limited exploration
- Like a cat sitting on a hot stove lid, only to get burnt, and learning to never sit on a stove lid again, even a cold one (Mark Twain, 1897; Denrell and March, 2001)

Conclusion

Policy discussion: issues with reinforcement learning

- **Key finding:** reinforcement learning encourages coordination but with a systemic bias towards redeeming, which exacerbates fragility
- **Policy challenges:** The excessive fragility is not a bug, but a rational adaptation
- **Policy solution:** Insofar that algorithm design (regulating internal model parameters) is not viable, we should think about market design approaches
  - Is an AI-agent 'kill-switch' to prevent panics viable and sensible?

Policy discussion: issues with reasoning LLMs

- **Key finding:** Since LLM-based investors act on context and reasoning, they are more resilient, but belief heterogenity makes their outcomes less predictable
- **Policy challenges:**
  - Despite being given the same information, LLM-based investors hold diverse beliefs
  - Who is the representative LLM-based investor for regulators and supervisors?
- **Policy suggestion:** develop "evals" to audit belefs / detect emerging narratives

Thank you!

Backup Slides

Baseline prompt

```
1  You are one of A=\%d active investors in a mutual fund out of a
       total of N=\%d investors. Each investor holds one share.
     Your goal is to maximize your return.
2  If you redeem your share, then you earn R1.
3  If you do not redeem your share, i.e., you stay, then you earn
     fraction * R1 * R * \theta, where:
4   - fraction = (N - W * (1 + λ)) / (N - W) is the fraction of
       assets remaining in the fund after serving redemptions;
5   - W ≤ A-1 is the number of other active investors who redeem
       (the remaining N - A investors are passive and never
       redeem);
6   - λ (lambda) = \%f is the illiquidity parameter;
7   - R1 = \%f is the value of the share if you redeem;
8   - R = \%f is the return earned by the fund from managing its
       portfolio;
9   - θ (theta) = \%f is the fundamental, which measures the fund'
       s performance.
10 Do you choose to redeem your share or stay? State your decision
      with exactly one word: ''redeem'' or ''stay'' using the
     XML tag <decision>...</decision>.
```

Prompt 1: Baseline prompt without mutual fund default risk

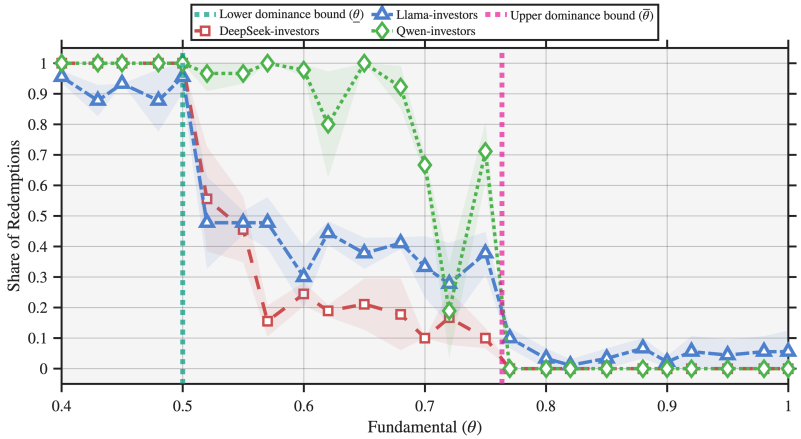
Modification to the baseline prompt

```
3 If you do not redeem your share (i.e., you stay), then with
      probability θ you earn fraction * R1 * R, and otherwise
      you get 0, where:
```

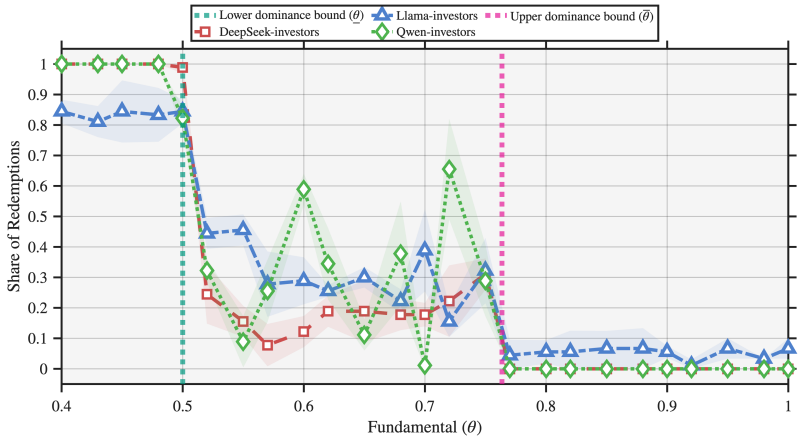
Prompt 2: Introducing mutual fund default risk

return

Robustness: different LLMs



(a) No payoff uncertainty



(b) With payoff uncertainty

return