Ex Machina: Financial Stability in the Age of AI

Kartik Anand[*] Sophia Kazinnik[†]
Agnese Leonello[‡] Ettore Panetti[§]

Abstract

We compare two types of AI decision systems, reinforcement learning algorithms and large language models (LLMs). We study them in a standard mutual fund redemption model where investors' decisions depend on what they expect others to do, so the same economic conditions can lead to either a stable or a fragile outcome. Both systems can steer the market away from the efficient outcome and trigger financial fragility, but for different reasons. Reinforcement learning (Q-learning) agents redeem too aggressively because of a systematic learning bias, while LLM agents struggle to coordinate because they hold different beliefs about how others will act. We identify why each system fails and examine how these failures can be corrected. Our results show that the design of AI systems can shape collective outcomes and financial stability.

Keywords: artificial intelligence; financial stability; strategic uncertainty; coordination games; Q-learning; large language models.
JEL classifications: G01, G23, C63, C72.

[*]Deutsche Bundesbank; kartik.anand@bundesbank.de
[†]Stanford University, HAI; kazinnik@stanford.edu
[‡]European Central Bank, DG-Research and CEPR; agnese.leonello@ecb.europa.eu
[§]University of Naples Federico II and CSEF; ettore.panetti@unina.it

We thank Wilko Bolt (discussant), Richard Berner, Markus Brunnermeier, Emilio Calvano, Jean-Edouard Colliard, Mehmet Ekmekci, Robert Engle, Thierry Foucault, Itay Goldstein, Michael Gofman, John Horton, Bryan Kelly, Shumiao Ouyang (discussant), Alessandro Previtero, Tamas Vadasz (discussant), and seminar audiences at the Deutsche Bundesbank, the European Central Bank, the BIS/CEPR/ECB Conference "Technological Innovations in Financial Markets", the Frankfurt/Mannheim Workshop on Digital Finance, the ECONDATA Knowledge Session, De Nederlandsche Bank, IIM-Bangalore, the NYU Seventh Annual Volatility and Risk Institute Conference, the Barcelona Summer Forum, the 2026 ESIF Economics and AI+ML Meeting, Banco de Portugal, the 20th CSEF-IGIER Symposium on Economics and Institutions, and the Bristol-Essex-UCL Workshop on Decentralised Financial Markets for useful comments. The views expressed here are the authors' and do not reflect those of the European Central Bank, the Deutsche Bundesbank, or the Eurosystem.

I. Introduction

Financial decisions are increasingly being delegated to AI systems. Reinforcement learning algorithms already execute a large share of institutional order flow, with algorithmic trading estimated to account for 60–70 percent of U.S. equity trading volume (Foucault et al., 2025). AI is also rapidly expanding into retail investment (Even-Tov et al., 2025).[1] As adoption grows, AI agents increasingly interact with each another and their decisions become interdependent. In this context, when strategic interactions shape financial outcomes, episodes of instability such as fire sales (Brunnermeier and Pedersen, 2009), financial market runs (Bernardo and Welch, 2004), and bank runs (Goldstein and Pauzner, 2005) may emerge as a result of coordination failures. Yet, despite the growing role of AI in trading and investment, we still know little about how AI systems interact and what those interactions mean for financial stability. Our paper sets to address this gap.

We study two canonical AI decision architectures, namely, reinforcement learning (specifically, Q-learning) algorithms, and reasoning-based large language models (LLMs). We place both in the same mutual fund redemption model with strategic complementarities and multiple equilibria. Because payoffs and information are held fixed, any difference in outcomes can be attributed to the architecture itself. Neither architecture consistently selects the efficient equilibrium, leading instead to financially fragile outcomes.

Both architectures depart from the efficient benchmark, but through different mechanisms. Q-learning investors redeem too much and converge on an outcome that leaves each of them worse off. LLM investors, by contrast, fail to coordinate, making their individual decisions difficult to predict. The two types of "misbehavior" have different sources. Q-learning is driven by a systematic learning bias, while LLM coordination breaks down because agents form different beliefs despite receiving the same prompt. Showing Q-learning investors the payoff from the action they did not take reduces the learning bias and over-redemption, but prevents them from settling on an equilibrium. For LLM investors, noisy signals about fundamentals and records of past outcomes bring beliefs closer together and improve coordination.

To our knowledge, this is the first paper to compare different AI decision architectures in a setting with multiple equilibria. We make three contributions. First, simulation experiments show that architecture is not merely an implementation choice. It determines which equilibrium emerges. Second, we formally prove three properties of the Q-learning bias: (i) the role of realized losses in driving these agents to over redeem, (ii) how algorithmic explo-

[1] Recent survey evidence indicates that globally around one in five retail investors use AI tools to select or adjust holdings (eToro, 2025). Moreover, large investment platforms now allow investors to delegate trading directly to autonomous AI agents (Assia, 2026).

ration shapes the bias, and (iii) why counterfactual learning engenders coordination. Third, we develop a framework for studying how LLM investors reason. By treating their generated reasoning as data, we recover their beliefs and causal explanations, making their otherwise opaque decisions easier to interpret.[2]

We ground our analysis in the canonical mutual fund redemption model of Chen et al. (2010), which is a parsimonious variant of the models by Diamond and Dybvig (1983) and Goldstein and Pauzner (2005). In the model, investors choose whether to redeem their shares before the fund's investment project matures. Their payoffs depend on both economic fundamentals and the actions of other investors. Strategic complementarities imply that the incentive to redeem increases when more investors redeem as well. Stronger fundamentals reduce the incentives to redeem. We model the fundamental as the probability with which the fund's investment yields a positive return and that investors who stay are repaid. It therefore captures default risk in the fund's portfolio.

Under complete information, the model produces three regions. When the fundamental is weak, it is a strictly dominant action to redeem, and the unique equilibrium is that all redeem. When the fundamental is strong, staying is strictly dominant, and all investors stay. For intermediate fundamentals, both outcomes are equilibria, so investors may either all redeem or all stay. The efficient equilibrium prescribes investors to redeem only when the fundamental is weak, and stay otherwise.

To explore the role of delegating redemption decisions to AI, we first develop a simulation-based experimental framework where we replace investors with Q-learning algorithms and LLMs. These systems represent two broad approaches to AI decision-making, learning from experience and reasoning from information. We define misbehavior as a departure from the efficient equilibrium that increases financial fragility. The two architectures misbehave in different ways: Q-learning investors over-redeem, while LLMs investors struggle to coordinate. These misbehaviors emerge through mechanisms specific to both architectures, which we discuss next.

First, the Q-learning bias operates through two effects. The first is a *hot stove effect*: the accumulation of losses following the materialization of default risk, combined with limited algorithmic exploration, biases the learning dynamics and pushes Q-learning investors toward inefficient redemptions.[3] We formally prove that this bias is unavoidable in the presence of limited exploration, stochastic payoffs and constant learning gains. The result is not limited

[2]LLMs, and deep learning models in general, are often described as "black boxes" because the processes behind their decisions are not well understood. This lack of transparency is considered an important source of risk (Cecchetti et al., 2025).

[3]The expression "*hot stove effect*" was coined by Denrell and March (2001) to refer to situations where agents over-react to negative experiences, for example, by entirely avoiding similar future situations.

to settings with multiple equilibria. It also holds when there is a unique dominant strategy Nash equilibrium.

The second channel through which the Q-learning bias operates is an *exploration effect*, which we expose in an exercise removing default risk. Using expected rather than realized payoffs reduces the hot stove effect, but the algorithm still fails to reach the efficient outcome. Instead, Q-learning converges to a strategy that best-responds to other investors randomizing between redeeming and staying. We formally prove that this behavior emerges when the exploration phase is sufficiently long relative to how quickly learning converges. Through these two effects, Q-learning investors coordinate, but on a privately and socially *suboptimal* action. This sharply contrasts with the algorithmic collusion literature, where algorithms coordinate on privately optimal actions and earn supra-competitive profits (Calvano et al., 2020; Colliard et al., 2025; Dou et al., 2025a). Here, coordination instead leads them to a harmful outcome.

LLM investors, by contrast, reason from the information provided in the prompt rather than update their behavior through realized payoffs. They are therefore not subject to the same hot stove effect.[4] However, the theoretical indeterminacy that gives rise to multiple equilibria in the intermediate region of fundamentals manifests itself as *belief heterogeneity*: identically prompted LLMs form different beliefs about how other investors will act. In this region, the same information can support several internally consistent beliefs about how other investors will behave. Because no single belief clearly dominates, independent runs can form different expectations and choose different actions. This variation in beliefs makes it difficult for LLM investors to coordinate.[5]

Having shown that the two architectures fail for different reasons, we next ask whether targeted changes can improve their behavior. This matters because safeguards that work for one type of AI may fail for another, or may replace one problem with a different one. For Q-learning, we introduce synchronous updating, which allows the algorithm to update its estimate of both actions after each round. We prove that this eliminates the hot stove effect. However, it also makes the estimated value of staying highly volatile, preventing investors from settling on a common equilibrium. This complements Banchio and Mantegazza (2025), who show that synchronous learning eliminates collusive and socially inefficient behavior in a game with a unique Nash equilibrium. In our setting with multiple equilibria, it removes the learning bias but creates a new coordination problem.

[4]LLMs generate text by predicting the next token based on the context provided and patterns learned during training.

[5]Importantly, this result is inconsistent with mere retrieval of known solutions from training data, suggesting that LLM investors are engaging in genuine strategic reasoning rather than behaving as "stochastic parrots" (Bender et al., 2021).

For LLMs, we test two interventions designed to reduce differences in beliefs. The first introduces uncertainty about the fundamental. Each LLM receives a noisy private signal, giving agents a common basis for forming expectations while preserving some disagreement. We find that their decisions converge around the theoretical global games threshold (Carlsson and Van Damme, 1993; Morris and Shin, 1998). The introduction of noisy signals induces LLM investors to hold a common belief on the actions of others, which, together with their private signals, drives their decisions to redeem or stay. Thus, imperfect private signals emerge as a focal point for the individual LLMs and, as such, improves coordination and stability. Second, we allow the LLMs to observe previous outcomes in a repeated game set up. In this case, we show that history serves as a focal point and so coordination emerges (Argenziano and Gilboa, 2012). Whether LLMs coordinate on the good (no redemption) or bad (universal redemption) equilibrium depends on the exact history: observing fewer redemptions leads to coordination on the Pareto optimal equilibrium, while the opposite to the socially and privately suboptimal one.

Taken together, our results show that financial stability depends not only on economic incentives but also on the type of AI system making the decisions. Learning-based systems can coordinate on inefficient outcomes, while reasoning-based systems can fail to coordinate when no clear focal point exists. The main takeaway is that AI does not simply implement standard equilibrium predictions, it changes how equilibria are selected, and hence how financial fragility emerges.

Outline. The rest of the paper proceeds as follows. Section II summarizes the related literature. In Section III, we present a stylized theoretical framework of mutual fund redemptions. In Section IV, we describe how we implement the two AI architectures. The baseline results highlighting the misbehaviors are reported in Section V. Remedies are analyzed in Section VI. Section VII concludes.

II. Related Literature

Our paper contributes to a growing body of work on the economics of AI, and in particular on strategic behavior in algorithmic systems. A central finding in this literature is that algorithms can learn to sustain tacit collusion and earn supracompetitive profits without communicating explicitly. This result was first shown by Calvano et al. (2020) in simulation experiments with Q-learning algorithms in a repeated pricing oligopoly and has since been documented in other settings, including auctions (Banchio and Skrzypacz, 2022). Algorithmic collusion and the mechanisms behind it have also been explored in financial market

applications, where asymmetric information, noise, and stochastic payoffs add further complexity. Collusion has been shown to emerge among algorithmic market makers (Colliard et al., 2025; Guarino et al., 2025), and AI-powered informed speculators (Dou et al., 2025a). More recently, Dou et al. (2025b) extend this literature by showing that algorithms can also learn trading strategies that manipulate prices.

Our analysis differs from this literature in two important ways. First, we study both Q-learning algorithms and LLMs, which represent two distinct approaches to AI decision-making. Second, we examine a static coordination game with strategic complementarities, placing the problem of equilibrium selection at the center of the analysis. Our contribution to the AI collusion literature is twofold. First, we show that the payoff dominant equilibrium, akin to the tacit collusion in Calvano et al. (2020), is not always selected by algorithms, and that, as a result, in our setting coordination is a double-edged sword; AI investors' coordination leads them to select a stable but privately inefficient equilibrium. Second, we complement simulation results with formal proofs of the mechanisms driving Q-learning algorithms to coordinate on a specific equilibrium. In line with recent analyses (Banchio and Mantegazza, 2025; Guarino et al., 2025; Dou et al., 2025c), which provide an analytical characterization of the algorithms' behavior and the equilibria they converge to, our paper methodologically advances the understanding of how Q-learning algorithms behave and how their learning is influenced by specific characteristics of the economic environment they operate in.

Horton et al. (2023) introduced the idea of using LLMs as simulated economic agents, or *homo silicus*, whose information, preferences, endowments, and incentives can be varied across economic settings. Our framework builds on this approach by recovering the beliefs and reasoning behind LLM decisions, not just their observed choices. Similarly, Cook et al. (2026) use revealed preference methods and structural estimation to identify LLM preferences in allocation and job search tasks, showing that these preferences are systematic but sensitive to framing and model interventions.[6] Building on the broader idea of using LLMs as economic agents, a growing literature simulates the behavior of investors (Fedyk et al., 2024), bank depositors (Kazinnik, 2023), and loan officers (Cook and Kazinnik, 2025). LLMs can also serve as proxies for survey respondents (Hansen et al., 2024) and reproduce human-like macroeconomic expectations (Bybee, 2023; Zarifhonarvar, 2024). Within finance, Lopez-

[6]The approach espoused here is distinct from agent-based modeling (ABM), which has been used in economics and finance to study the emergent behavior of heterogeneous, boundedly rational interacting agents that cannot be captured using a representative agent framework (Farmer and Foley, 2009; Battiston et al., 2012). Whereas ABMs specify behavioral rules for interacting agents to study aggregate dynamics, the approach here treats the LLM itself as an object of inquiry and seeks to understand how AI agents form beliefs, reason, and make decisions.

Lira (2025) introduce an open-source framework that pits LLM trading agents against value investors, momentum traders, market makers, and contrarians, while Gao et al. (2024) show that heterogeneous LLM investors generate price dynamics resembling those observed in financial markets. Finally, Bhagwat et al. (2025) elicit beliefs from LLM investor personas and find that disagreement about firm news both tracks abnormal trading volume and predicts a subsequent return premium.

We contribute to this literature in two ways. First, we compare LLM and Q-learning investors in a canonical coordination game, clarifying how explicit reasoning, rather than learning, shapes equilibrium selection. Second, we introduce a framework for analyzing LLMs' latent reasoning. By treating chain-of-thought text as data, we recover beliefs about others' actions, the justifications for those beliefs, and the causal links from justifications to beliefs and decisions. Because we infer beliefs from the reasoning that produces decisions, rather than through direct elicitation, our approach is less vulnerable to post-hoc rationalization and helps explain why LLM investors behave as they do.

Our analysis shows that belief heterogeneity prevents coordination when multiple equilibria exist: without a shared focal point, identically prompted LLMs form different but internally consistent beliefs about others' behavior. With noisy private signals, however, they converge on a common belief and behave in line with the global-games equilibrium. This parallels Heinemann et al. (2004), who find that private information moves human behavior closer to the global-games solution, providing a useful benchmark for interpreting LLM responses to strategic uncertainty. Together with our reasoning analysis, these results suggest that LLMs can apply sophisticated game theoretic logic when the information structure guides belief formation.

Within the nascent literature on LLMs as decision-making agents, our paper is closest to Fish et al. (2024). Our results on LLMs complement their analysis, which shows that LLM-based pricing agents learn to collude in a repeated Bertrand competition game. We show that learning is not the only path to coordination. With fundamental uncertainty, i.e., when LLM investors receive private signals, they coordinate even in a one-shot setting. This suggests that the information environment itself can serve as a coordination device for reasoning-based AI. Furthermore, collusion (i.e., coordination on the privately optimal outcome) is not automatically achieved when extending the environment into a dynamic setting. Equilibrium selection is obtained by allowing LLMs to observe past history, but history itself determines which equilibrium LLMs coordinate on.

Our paper also contributes to the emerging literature on AI and financial stability. This topic has attracted much attention from policymakers (Shabsigh and Boukherouaa, 2023; Aldasoro et al., 2024; Financial Stability Board, 2024; Leitner et al., 2024; Bank of England,

2025), discussing several channels through which AI may destabilize the financial system, including procyclicality, unknown unknowns, herding, increased exposure to cyber risks. Danielsson and Uthemann (2025) formalize some of these ideas using a global games framework, showing that AI's superior information processing, reliance on common data sources, and speed advantages can make transitions between stable and crisis states more abrupt. We advance this literature by studying AI behavior in a canonical coordination game grounded in economic theory. This setting allows us to isolate the forces shaping Q-learning and LLM decisions and to show how they affect equilibrium selection and financial fragility. In this respect, our study is also related to Yang (2024), who shows that Q-learning can trigger currency attacks in a two-agent speculative attack setting. We instead study a mutual fund redemption game with many interacting investors. We formally characterize Q-learning behavior and compare it directly with LLM behavior, allowing us to show how AI architecture and the economic environment jointly shape financial fragility.

Finally, we also contribute to the literature on financial fragility. Following the seminal work of Diamond and Dybvig (1983), this literature has extended to other financial institutions subject to strategic uncertainty, such as mutual funds (Chen et al., 2010), hedge funds (Liu and Mello, 2011), credit markets (Bebchuk and Goldstein, 2011), life insurers (Foley-Fisher et al., 2020) and stablecoin issuers (Gorton et al., 2025), among others. Despite rich theory, credible empirical evidence remains scarce, with Goldstein et al. (2017) and Chen et al. (2024) as notable exceptions. Our contribtion here is that we use AI agents to test theories of financial fragility in a setting where information, payoffs, and strategic interactions can be precisely controlled. This allows us to isolate how economic fundamentals and AI design shape financial risk.

III. Theoretical Framework

We build on Chen et al. (2010) and characterize the equilibria in a stylized mutual fund redemption game. The economy extends over two dates, $t = 1$ and $t = 2$, and consists of a mutual fund and an integer number, $N > 2$, of risk-neutral investors. Each investor holds one share in the mutual fund. Thus, the initial size of the fund is N.

At $t = 1$, the book value of the fund's investments is $R_1\, N$, which is common knowledge, with $R_1 \geq 1$. At $t = 2$, the per-unit investment return is instead:

$$R_2(\theta) = \begin{cases} R & \text{with prob. } \theta, \\ 0 & \text{with prob. } 1-\theta, \end{cases} \tag{1}$$

where $\theta \in [0,1]$ is a uniformly distributed random variable drawn at the start of $t = 1$ and represents the strength of the economic fundamentals. Stronger fundamentals (higher θ) imply a higher probability that the fund's investments yield a high return and hence lower default risk for investors.[7]

At the beginning of $t = 1$, a number A of active investors observe θ and choose whether to redeem their shares or hold them until $t = 2$. Investors who redeem receive the current value R_1. However, in order to service redemptions, the fund must liquidate its investments, which is costly. Specifically, to raise R_1, the fund must liquidate $(1 + \lambda)R_1$ worth of assets, where $\lambda > 0$ is a measure of illiquidity. The higher is λ, the more assets must be liquidated to service redemptions. Following Chen et al. (2010), we assume that $A \leq \frac{N}{1+\lambda}$, implying that the mutual fund has enough resources to meet all redemptions.

After servicing redemptions at $t = 1$, the fund re-invests its remaining resources. Denoting the number of investors who redeem at $t = 1$ by $W \leq A$, an investor who decides to stay invested expects to receive a pro-rata share of the fund's return at $t = 2$, which is given by:

$$R_2(\theta)\, R_1\, \frac{N - W(1+\lambda)}{N - W}. \tag{2}$$

Thus, an investor's payoff from staying is increasing in the strength of economic fundamentals, θ, and decreasing in both asset illiquidity, λ, and the number of redemptions, W. It is useful to define the critical mass of redemptions $\widehat{W}(\theta)$ at which an investor is just indifferent in expectations between staying and redeeming, as the solution to:

$$1 = R\theta \frac{N - \widehat{W}(1+\lambda)}{N - \widehat{W}}. \tag{3}$$

The threshold $\widehat{W}(\theta)$ is increasing in θ, as stronger fundamentals increase the value of staying and therefore requires a higher critical mass of redemptions to make investors indifferent between staying and redeeming.

We focus on characterizing equilibria in pure strategies. Following standard lines of reasoning (Morris and Shin, 1998), we can partition the support for the fundamental into three intervals, which we report in Lemma 1 and Figure 1.

Lemma 1: *The equilibrium outcomes depend on the realization of θ. For sufficiently extreme values of θ, there exist a unique Nash equilibrium in pure strategies, while for intermediate values multiple equilibria arise. Specifically,*

[7] In Chen et al. (2010), the variable θ is a measure of the performance of the mutual fund. For the purpose of our exercise, the precise definition is immaterial.



Figure 1: Equilibria without fundamental uncertainty.

- *if* $\theta < \underline{\theta} = \frac{1}{R}$, *the unique equilibrium is that all investors redeem at* $t = 1$*;*
- *if* $\theta > \overline{\theta} = \frac{1}{R}\frac{N-(A-1)}{N-(A-1)(1+\lambda)}$, *the unique equilibrium is that all investors stay until* $t = 2$*;*
- *if* $\theta \in [\underline{\theta}, \overline{\theta}]$, *both all-redeem and all-stay are equilibria.*

An important prerequisite for multiple equilibria is costly liquidation. When the fund's investments are perfectly liquid, i.e., $\lambda = 0$, the expected return from staying is independent of early redemptions. This implies that redeeming at $t = 1$ is the strictly dominant action for $\theta < \underline{\theta}$, while staying is strictly dominant for $\theta > \underline{\theta}$. Only negative net present value investments are unwound. Proposition 1 describes this first-best efficient benchmark. In contrast, when $\lambda > 0$, the expected payoff difference between staying until $t = 2$ and redeeming at $t = 1$ decreases monotonically with the number of investors who redeem. Investors' decisions are strategic complements, i.e., the incentive to redeem is increasing in the belief that others will redeem as well. Crucially, for $\theta \in [\underline{\theta}, \overline{\theta}]$ the belief that all others will redeem and all others will stay are both rationalizable leading to multiple self-fulfilling equilibria.

Proposition 1: *The efficient equilibrium outcome is for all investors to redeem at* $t = 1$ *if* $\theta < \underline{\theta}$ *and stay if* $\theta \geq \underline{\theta}$.

In what follows, we use the result in Proposition 1 as the benchmark against which we measure algorithmic behavior when AI agents play the game.

IV. Experimental Design

We replace the investors of the theoretical model with algorithmic counterparts. We consider two distinct types of AI-agents: Q-learning investors, who optimize through trial-and-error reward updating, and LLM investors, who reason contextually via chain-of-thought inference. The two classes differ not only in how they behave but also in how we can characterize that behavior. For Q-learning investors, the learning dynamics are analytically tractable: we establish their limiting behavior in closed form and use simulation to confirm the result and probe its robustness. For LLM investors, no such characterization is available, and simulation is the primary evidence. Both classes are evaluated under a common economic environment.

There are $A = 30$ active investors out of a total of $N = 50$. The parameters governing investment returns are $R_1 = 1$ and $R = 2$. For the liquidation cost, we set $\lambda = 0.25$. We consider a range of different values for the fundamental, θ, drawn from an equally spaced twenty-five-point grid spanning the interval $[0.4, 1]$, that covers both the efficient redemption bound $\underline{\theta} = 0.5$ and the no redemption bound $\overline{\theta} = 0.77$.

A. Q-learning

Q-learning is a transparent reinforcement learning algorithm. With no prior knowledge of the payoff structures, agents learn the values of choosing different actions through trial-and-error interaction with the environment and by receiving rewards and penalties. An agent's optimal action is the one that maximizes the cumulative reward via repeated interactions.

In our set-up, the algorithm has four components:

1. **States**: The state θ_i for each Q-learning investor i is derived from the information that they receive, i.e., either the true fundamental θ or the noisy signal s_i. As the Q-learning algorithm requires a finite state space, this continuous information is mapped to a discrete classification, which constitutes the effective state of the algorithm. The set of all states is denoted $\Theta \subset \mathbb{N}$.

2. **Actions**: The action set is binary, $\mathcal{A} = \{a_R, a_S\}$, where a_R denotes "redeem" and a_S denotes "stay".

3. **Rewards**: The reward function $\pi(\theta, a)$ assigns a payoff based on the realization of the fundamental θ and the chosen action a.[8] Crucially, Q-learning investors do not observe the underlying mapping $(\theta, a) \mapsto \pi$, but only the realized rewards.

4. **Episodes**: Learning unfolds over T episodes. In each episode, Q-learning investors observe their state, choose an action, and update their Q-values based on the realized payoff.

Each Q-learning investor $i = 1, \ldots, A$ starts episode $t = 1, \ldots, T$ with a Q-matrix $Q_{i,t} \in \mathbb{R}^{|\Theta| \times 2}$, with rows for states and columns for actions. Following the realization of the state $\theta_{i,t}$, the action is determined according to an ε-greedy policy: with probability $\varepsilon_t = \beta^t$, the Q-learning investor explores by randomizing between a_R and a_S, while with the complementary probability $1 - \varepsilon_t$, the Q-learning investor exploits by choosing the action that maximizes $Q_{i,t}(\theta_{i,t}, a)$.

[8] Note that $\pi(\theta, a)$ implicitly depends on the actions of other investors through Equation (2), but we suppress this dependence in the notation since Q-learning treats other agents' actions as part of the environment.

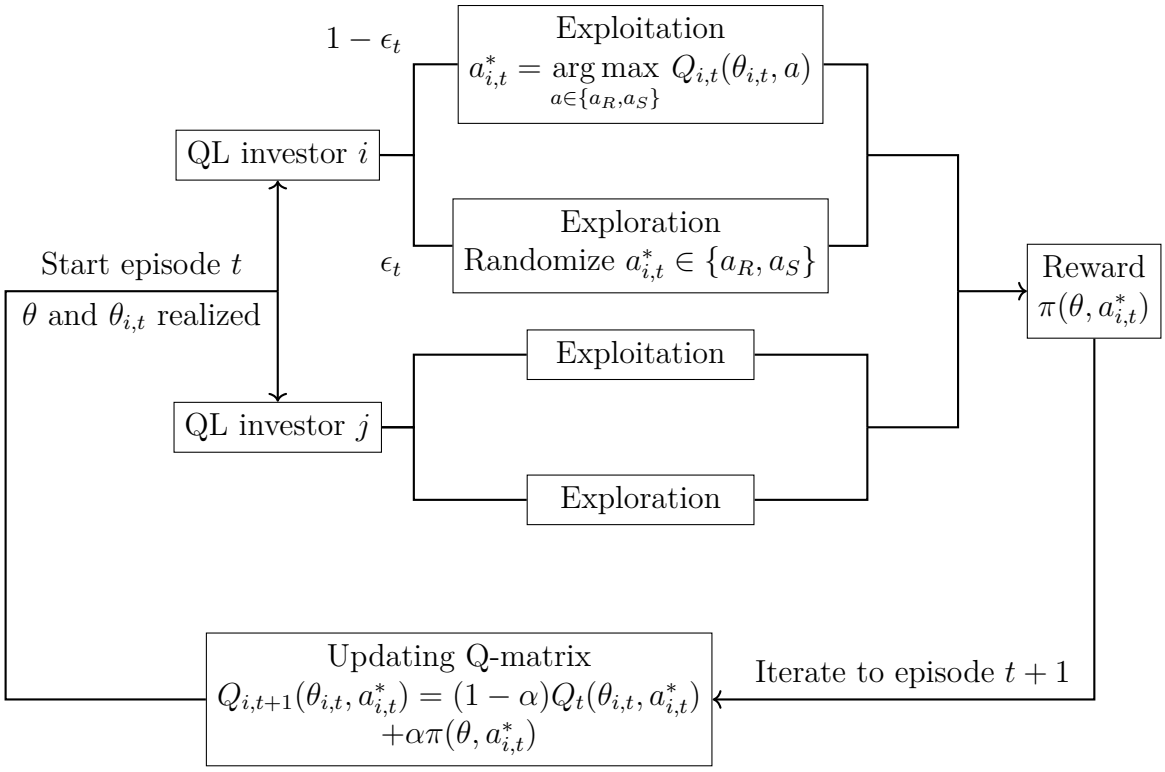


Figure 2: Iterative process of Q-learning.

If the chosen action is $a^*_{i,t} = a_R$, the reward is $\pi(\theta, a_R) = R_1$. If instead $a^*_{i,t} = a_S$, the reward depends on θ (the true fundamental) and the other Q-learning investors' actions and is given by Equation (2). The Q-matrix is then updated as

$$Q_{i,t+1}(\theta_{i,t}, a^*_{i,t}) = (1-\alpha)Q_{i,t}(\theta_{i,t}, a^*_{i,t}) + \alpha\,\pi(\theta_{i,t+1}, a^*_{i,t}), \tag{4}$$

where $\alpha \in [0,1]$ is the learning rate: higher α places more weight on new rewards, while lower α smooths learning over past experience. Figure 2 summarizes the iterative process.

In our simulations, we set $\beta = 0.99999$, $\alpha = 0.1$ and $|\Theta| = 75$. We run 25 independent rounds of training, each with $T = 500,000$ episodes.[9] During each training round, the initial Q-values are drawn at random from the unit interval. For these hyperparameters, each Q-learning investor experiments approximately 100,000 times.Assuming a decaying exploration rate is standard in the literature (e.g., Colliard et al., 2025) and reflects the idea that exploration corresponds to training, which is costly and therefore limited.

[9]The results remain qualitatively unchanged if we use a convergence criterion instead, i.e., the variance in the share of investors redeeming over a rolling window of episodes is sufficiently small.

B. *Large Language Models*

LLMs are sophisticated next-word predictors. Drawing on the vast corpora of text used for training, these models attempt to determine the meaning and context of prompts, or input text, given to them in order to improve their predictions and responses. In addition, some models, like OpenAI's o3, Google's Gemini 2.5 Pro, and DeepSeek's R1 can "think out loud", i.e., produce a sequence of latent reasoning steps linking the prompts given to them, to assumptions, intermediate steps of calculations to the final output. As such, LLMs are capable of instantaneously adapting their behavior when the information they are presented with changes. In our experiments, we use DeepSeek's R1-0528 to model the LLM investors. A key advantage of using this LLM is that the full chain-of-thought reasoning can be extracted and analyzed, unlike with other reasoning-based LLMs. DeepSeek's models are also open-weight, which supports transparency and replicability.[10]

The experiment unfolds in two stages. First, we design the prompts. Each prompt distills the economic environment into a concise, textual description that specifies the investor's objective, the structure of payoffs, and the information available.

Second, the prompt is instantiated with the relevant parameters and submitted to the LLM via API calls. Every investor is represented by a separate, independent API call. To encourage consistent behavior, we set the temperature hyperparameter to zero, reducing randomness in the outputs.[11] From each call, we record both the latent chain-of-thought reasoning and the final decision. We run three independent rounds of inference to account for any residual randomness and construct confidence intervals around the outcomes.

Finally, to explain the decision-making of LLM investors, we treat the chain-of-thought reasoning text as unstructured data, which we then analyze using an unsupervised LLM (we also use DeepSeek R1-0528) to extract beliefs and causal reasoning. Our approach proceeds in two steps. First, we prompt the LLM to extract from each investor's reasoning text: (i) the belief they held about the behavior of others, and (ii) the justification underlying that belief. Second, we ask the same unsupervised LLM to cluster these into canonical game-theoretic categories. The resulting causal graph maps justifications to beliefs to decisions, revealing the logical structure of AI-investor reasoning.[12] To ensure consistency in the labels for the justifications and beliefs, we jointly cluster the DAGs for all investors across all values of the fundamental.

[10] For a broader discussion on the use and evaluation of open-weight LLMs, see Cook et al. (2023).

[11] The temperature setting controls the degree of randomness in the model's responses: a value near zero yields deterministic and focused outputs, while higher values introduce more variation.

[12] The prompts used to extract beliefs and causal reasoning are shown in Online Appendix B.

```
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.
If you redeem your share, then you earn R1.
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:
- fraction = (N - W * (1 + λ)) / (N - W) is the fraction of assets remaining in the fund after serving redemptions;
- W ≤ A-1 is the number of other active investors who redeem (the remaining N - A investors are passive and never redeem);
- λ (lambda) = \%f is the illiquidity parameter;
- R1 = \%f is the value of the share if you redeem;
- R = \%f is the return earned by the fund from managing its portfolio;
- θ (theta) = \%f is the fundamental, which measures the fund's performance.
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.

V. Baseline results: Algorithmic misbehaviors

Under the efficient equilibrum, all investors should redeem if and only if the fundamental is below the lower dominance bound, i.e., $\theta < \underline{\theta}$. Building on this, we classify an algorithm as misbehaving if the realized outcome of redemptions deviates from the efficient equilibrium. Those are the instances in which individual behaviour collectively trigger financial fragility.

Figure 3 plots the share of Q-learning and LLM investors who redeem, as a function of the economic fundamental, θ. Both algorithms adhere to the efficient benchmark when the fundamental is weak, i.e., all Q-learning and LLM investors redeem when $\theta < \underline{\theta}$. But above this threshold, they misbehave and their behaviors differ significantly. Q-learning investors over-redeem. They coordinate on redemption even in ranges of fundamentals where no redemptions should be observed, i.e., when $\theta > \overline{\theta}$. In contrast, when the fundamental is strong, $\theta > \overline{\theta}$, all LLM investors stay and adhere to the efficient benchmark. However, in the intermediate region, they do not coordinate on the all-stay Nash equilibria. Instead, for

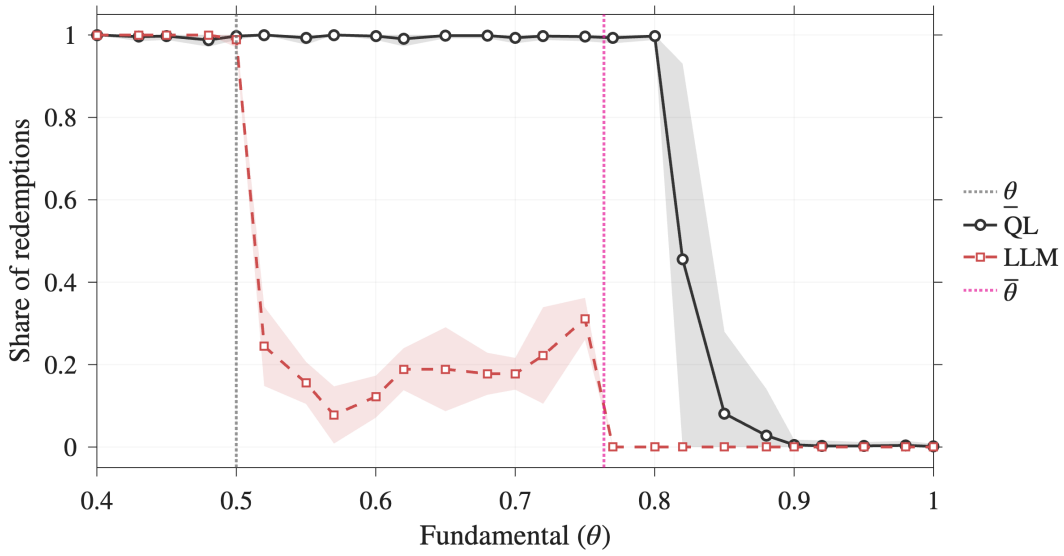


Figure 3: Share of redemptions as a function of the fundamental θ. The line labelled QL refers to the outcome with Q-learning investors, while the line labeled LLM is with the LLM investors. The vertical lines $\underline{\theta} = 0.5$ and $\overline{\theta} = 0.78$ are the bounds defined in Lemma 1.

a given θ, some LLM investors stay while others redeem.

Thus, Q-learning and LLMs exhibit different misbehaviors that lead them to deviate from the efficient equilibrium benchmark. Next, we explain the mechanisms behind these results. For ease of exposition, we start with exploring over-redemption with Q-learning investors and then subsequently turn to the LLM investors' inability to coordinate.

A. Q-learning: Learning biases

A.1. Hot stove effect

The redemption bias stems from how the underlying learning dynamics operate. Consider the case where θ is just above $\overline{\theta}$. Since the fundamental is strong, the probability that the fund's investments default is low. However, each time a Q-learning investor chooses to stay and default risk materializes, they obtain a zero reward, which revises down the Q-value of staying. Since the reward from redeeming is fixed and certain, following a sufficiently long sequence of consecutive zero-rewards, the investors switch to favoring redeeming. Once an investor switches, they stop updating the Q-value for staying. As a consequence, they incorrectly conclude that the expected reward from staying is below that of redeeming, and this error is never corrected. The "hot stove effect" has previously been identified to explain

other algorithmic biases (Denrell and March, 2001).[13]

As the description above suggests, the hot stove effect requires both a sufficiently long sequence of consecutive default events and that Q-learning investors do not continue to update. the Q-values for stay once they switch. While the second property is a consequence of the standard assumption of asynchronous updating (i.e., only the Q-value of the action taken is updated) the first one arises from vanishing exploration. Proposition 2 shows that the hot stove effect is highly pervasive and biases Q-learning investors towards redeeming for all $\theta < 1$.

Proposition 2: *Q-learning investors almost surely reach a date in finite time after which they redeem forever for any* $\theta < 1$.

In the special case $\theta = 1$, the fund's investments never default, so staying yields a strictly positive payoff with certainty and emerges as the outcome in line with the efficient equilibrium benchmark. For every other value of θ, however, the hot stove effect dominates and Q-learning investors lock in on redeeming, even where staying is the unique dominant-strategy equilibrium. The mechanism is therefore not one of equilibrium selection: strategic uncertainty and equilibrium multiplicity play little role. This is a learning bias: correct payoff signals are mishandled under updating.

Our findings are broadly consistent with other papers in the literature (e.g., Colliard et al., 2025; Dou et al., 2024), that show Q-learning algorithms are unable to play the Nash equilibrium as predicted by theory. There is, however, an important caveat. These papers find that Q-learning algorithms tacitly collude on outcomes that deliver higher payoffs and rewards, i.e., they choose actions that are privately beneficial. Instead, in our framework Q-learning investors choose the action that gives them the lowest (and sure) payoff, that is privately suboptimal.

A.2. Exploration effect

To further understand the forces behind the over-redemption by Q-learning investors, we remove default risk. In this case, we assume that with probability one the payoff from staying is equal to its expected value, i.e., $R_2(\theta) = R\theta$. Importantly, this does not change the efficient benchmark, since risk-neutral investors base their decisions on the expected payoffs. Figure 4 plots the share of redemptions as a function of the fundamental for Q-learning investors. The green line is our baseline result, while the black line represents the case where the reward for staying is the expected payoff. We readily note that there is a sharp switch in the share

[13] The name of the effect is in deference to Mark Twain who once wrote, "*Like a cat sitting on a hot stove lid, only to get burned, and learning to never sit on a stove lid again, even a cold one*".

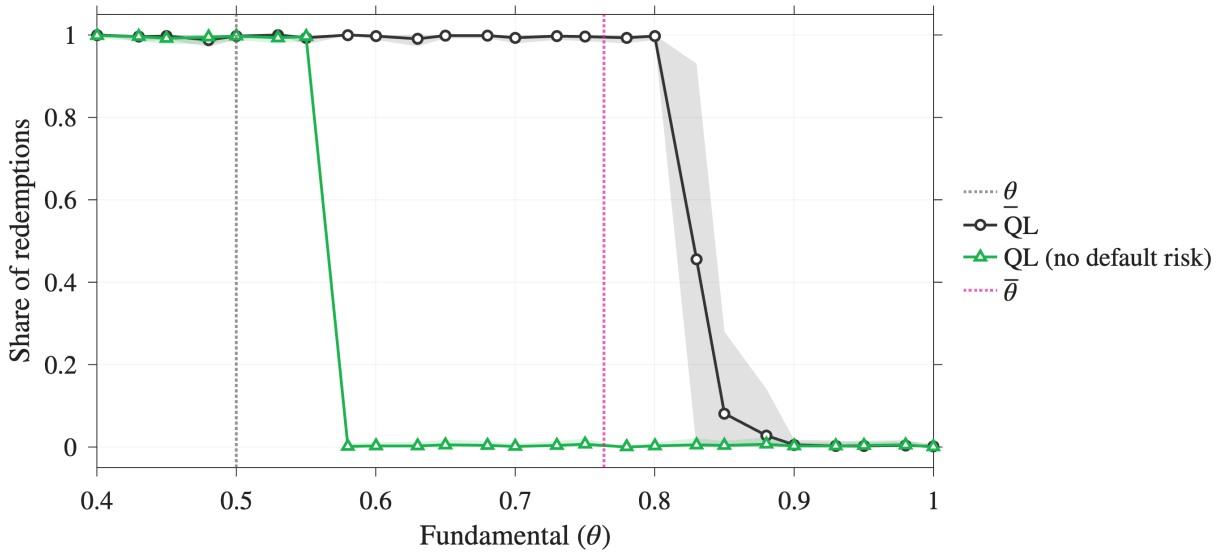


Figure 4: Share of redemptions as a function of the fundamental θ. The line labelled QL refers to the outcome with Q-learning investors, while the line labeled QL (no default risk) refers to the ablation exercise where we replace the realized rewards with the expected rewards. The vertical lines $\underline{\theta} = 0.5$ and $\overline{\theta} = 0.78$ are the bounds defined in Lemma 1.

of redemptions around $\tilde{\theta} = 0.558$: for $\theta < \tilde{\theta}$, all Q-learning investors redeem, while for larger values of θ, they all stay.

The threshold $\tilde{\theta}$ aligns with the solution to the indifference condition between redeeming and staying, where the expected payoff for staying is calculated assuming that all other investors randomize between redeem and stay with equal probability. Formally,

$$\tilde{\theta} = \frac{R_1}{\left(\frac{1}{2}\right)^{A-1} \sum_{W=0}^{A-1} \binom{A-1}{W} \left(\frac{N-W(1+\lambda)}{N-W}\right) R_1 R}. \tag{5}$$

Crucially, $\tilde{\theta}$ lies strictly above the efficient redemption bound $\underline{\theta} = 0.5$. So, while removing default risk eliminates the hot stove effect, it exposes a residual and more subtle mechanism, which we call the *exploration effect*: the algorithms learn to best-respond to the random play generated by their exploration.

During exploration, every investor randomizes with equal probability, so the environment each Q-learning investor faces is the equal-probability mixing distribution underlying $\tilde{\theta}$. Since the Q-values update at rate α, the investors learn the expected payoffs of this environment provided two conditions hold. First, the influence of the initial Q-matrix must dissipate, which the constant learning rate guarantees. Second, learning must converge faster than exploration vanishes. If the exploration rate ε_t decays too quickly, investors lock in on half-

learned values shaped by initial conditions rather than on the best response to randomizing opponents. When learning is fast relative to the decay of exploration, there is a window in which exploration remains high and the Q-values have already settled onto the equal-randomization best response. Proposition 3 formalizes this result.

Proposition 3: *Suppose that the learning rate, α, and exploration rate, $\varepsilon_t = \beta^t$, satisfy:*

$$-\log \beta \ll \alpha. \tag{6}$$

Then, absent default risk, Q-learning investors converge to the threshold policy that redeems for $\theta < \tilde{\theta}$ and stays for $\theta > \tilde{\theta}$, where $\tilde{\theta}$ is given by Equation (5).

Proposition 3 isolates the residual exploration effect, i.e., with the hot stove removed, the Q-learning investors converge to outcomes governed entirely by the exploration distribution, which we characterize in closed form. The result explains the findings in Abada and Lambin (2023) who show via simulations that replacing uniform exploration with "rational" exploration on the Nash action eliminates the non-Nash outcomes.

B. LLM investors: Belief heterogeneity

Figure 3 shows that LLM investors fail to coordinate when the economic fundamentals fall in the range $\left(\underline{\theta}, \overline{\theta}\right)$. By analyzing the chain-of-thought produced by each individual LLM investor, we trace this behavior to belief heterogeneity.

Figure 5 presents the causal graph for $\theta = 0.75$, obtained by analyzing the chain-of-thought reasoning text produced by each LLM investor in one round. Investors exhibit substantial heterogeneity in both justifications and beliefs. Two justification categories were identified: (i) *payoff comparison*, where investors compare payoffs across equilibria under particular beliefs, and (ii) *equilibrium analysis*, where investors reason using formal equilibrium concepts. These justification categories map into three belief categories: (i) *focal point (efficient)*, expecting others to coordinate on the Pareto-optimal equilibrium; (ii) *focal point (safe)*, expecting others to coordinate on the safe outcome; (iii) *probabilistic beliefs*, assuming others redeem or stay with specific probabilities. The mapping from beliefs to actions suggests that the investors correctly play best-responses, i.e., when expecting others to coordinate on the efficient equilibrium they choose to stay, while when they expect others to coordinate on the safe equilibrium, use a uniform prior, or use worst-case reasoning, they choose to redeem.

This heterogeneity is striking because all investors receive identical prompts and the prompts do not give any guidance about how to deal with beliefs and multiple equilibria.

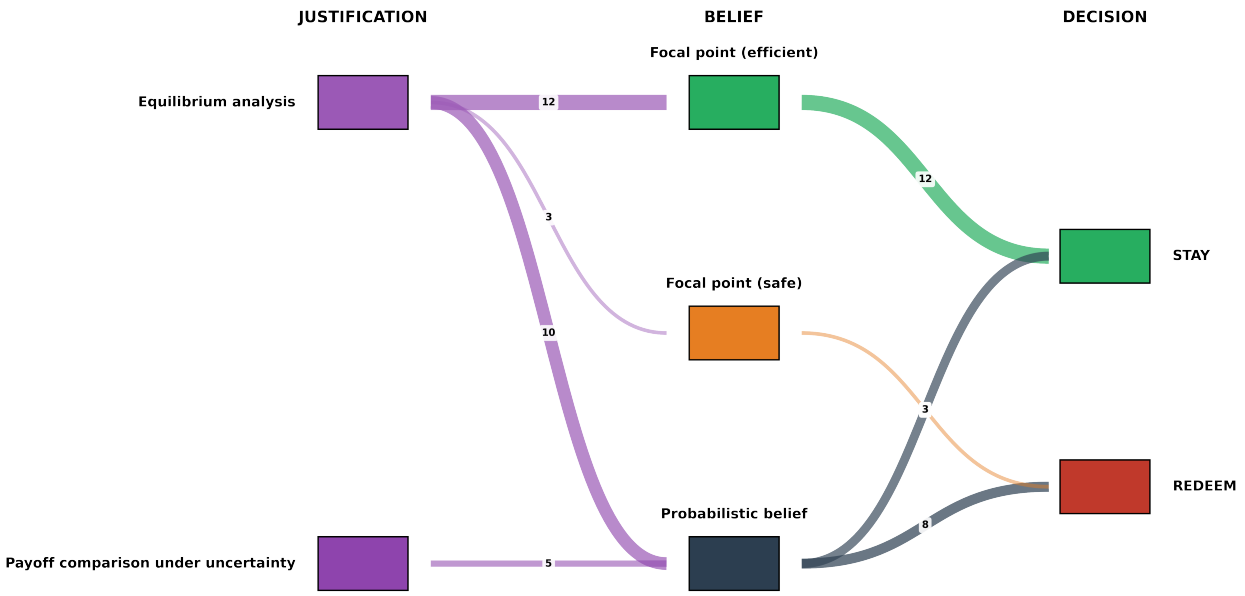


Figure 5: Causal mapping from justifications to beliefs and decisions, for LLM investors at $\theta = 0.75$. Link widths and numbers indicate investor counts.

While we set temperature to zero to minimize randomness, prompting each investor independently leaves minor numerical variation in the outputs. That such small perturbations send investors down divergent reasoning paths is surprising but economically intuitive. For intermediate realizations of the fundamental, the prompt does not pin down a reasoning path. Instead, several internally consistent frameworks and beliefs are admissible. Thus, while the prompt fixes the support and distribution over reasoning paths, it does not select a realized path. Each independent query resolves this near-tie into one path, and because the queries are independent, different investors resolve it differently. The outcome is a population of agents that receive identical prompts yet hold heterogeneous beliefs — not because any investor reasons incorrectly, but because the prompt underdetermines which valid framework to apply, and independent realizations select different ones. Importantly, this emerging heterogeneity is also inconsistent with retrieval of memorized solutions from training data, suggesting that LLM investors are not merely acting as "stochastic parrots" (Bender et al., 2021).

Figure 6 reveals three regimes in how LLM investors reason about coordination as the fundamental varies. For low θ (below $\underline{\theta}$), nearly all investors report "no belief", consistent with redeeming being the dominant strategy and requiring no strategic reasoning. For high θ (above $\overline{\theta}$), "no belief" again dominates, now because staying is the dominant strategy.

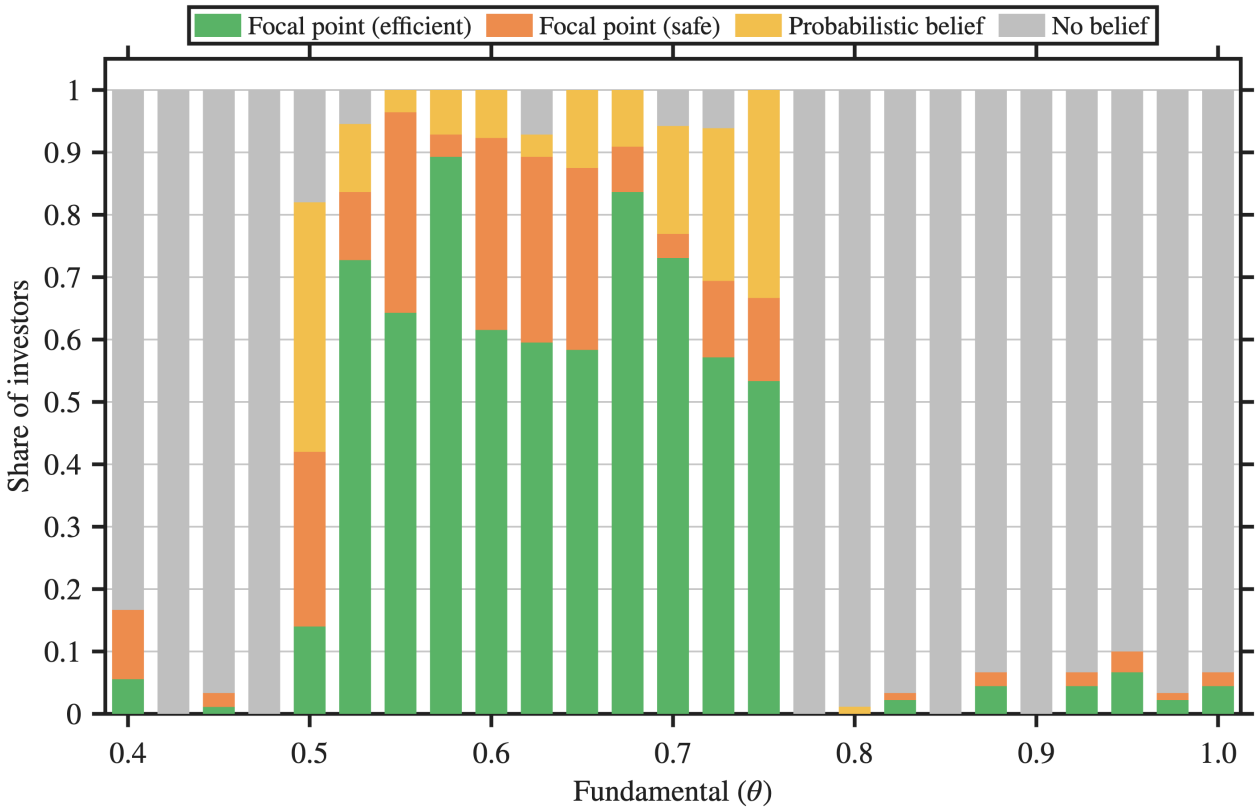


Figure 6: Evolution of LLM investors' beliefs about the behavior of other investors as a function of the fundamental θ.

In the intermediate region, beliefs become heterogeneous. As θ rises within this region, the share of investors expecting coordination on the efficient equilibrium increases. This aligns with the game's structure: higher fundamentals reduce the critical mass redeemers required for staying to be optimal, enlarging the basin of attraction of the all-stay equilibrium. This agreement on the threshold holds despite disagreement on beliefs. As Table 1 shows, the cutoff $\widehat{W}$ for the number of investors who redeem from the indifference condition inferred from the chain-of-thought matches the theoretical values across the interval with only minor cross-investor variation. In other words, LLM investors agree on where the threshold lies even where they disagree on what others will do.[14] The LLM investors, thus, suffer from a *belief heterogeneity*.

A natural concern is that these results reflect the LLM retrieving a memorized solution to a familiar redemption game rather than reasoning from the economic primitives. If the model were simply pattern-matching to a canonical bank-run problem in its training corpus, belief heterogeneity would be an artifact of the framing, not a genuine feature of strategic reasoning. To address this, in Appendix C we consider three prompt variations: (i) we remove default risk, (ii) strip the fund-redemption framing entirely in favor of a neutral action-choice problem, and (iii) re-express the identical payoffs in natural language, requiring investors to compute each action's return from a narrative scenario description. Across all three, the belief-heterogeneity pattern is preserved and localizes to the intermediate region exactly where theory predicts multiple equilibria. That behavior survives the removal of both default risk and the financial framing indicates that LLM investors reason from the strategic structure rather than pattern-match to a familiar problem.

VI. Treatments

The misbehaviors of Q-learning and LLM investors stem from distinct mechanisms: Q-learn investors suffer from learning biases while LLM investors exhibit belief heterogeneity. In this section, we suggest various informational remedies that augment what the algorithms observe, either about the fundamental or about the outcome of play across episodes to address the shortcomings and issues identified.

We consider three remedies, where the first two target the belief heterogeneity of LLM investors while the third addresses the learning bias – specifically the hot stove effect – of

[14] This pattern indicates that LLM investors rely on simple expected-value comparisons. It follows that default risk plays no role in LLM behavior not because such risk is immaterial in principle, but because LLM investors' decision rules do not appear to incorporate concerns about risk or ambiguity. In Appendix C we report the equivalents of Figures 5 and 6 in the absence of default risk, confirming that LLM behavior is unchanged.

Table 1: LLM investors' critical cutoff estimates in the presence of default risk.

θ	Mean($\widehat{W}$)	Var($\widehat{W}$)	Theoretical ($\widehat{W}$)
0.50	0.000	0.000	0.000
0.55	13.556	1.432	13.333
0.60	20.000	0.000	20.000
0.65	24.000	0.000	24.000
0.70	26.667	0.000	26.667
0.75	28.571	0.000	28.571

The means and variances of the critical cutoff $\widehat{W}$ are inferred from the chain-of-thought for each individual LLM investor.

Q-learning investors. In the first treatment we replace common knowledge of θ with a noisy private signal for the investors. The standard global games logic (Carlsson and Van Damme, 1993; Morris and Shin, 1998) suggests that this should resolve the indeterminacy in investors' beliefs and deliver a unique equilibrium. The second treatment instead involves providing LLM investors with a history of previous play that should work as an anchor to beliefs (Argenziano and Gilboa, 2012). Finally, to address the hot stove effect of Q-learning investors, we consider synchronous updating where the Q-value for both actions are updated after each episode (Banchio and Mantegazza, 2025).

A. Remedying belief heterogeneity with private signals

This remedy introduces fundamental uncertainty. Instead of directly observing θ, each investor i receives a private signal $\theta_i = \theta + \eta_i$ where the η_i are i.i.d. noise drawn from $[-\eta, +\eta]$. The parameter $\eta > 0$ is a measure of information dispersion.

With private signals the redemption game becomes a global game with a *unique* equilibrium in threshold strategies (Carlsson and Van Damme, 1993; Morris and Shin, 1998). In this equilibrium, each investor redeems if and only if their signal falls below a common cutoff $\theta^* \in (\underline{\theta}, \overline{\theta})$. Under this threshold strategy, for a given level of the fundamental, θ, the fraction of investors who redeem is a function $w^*(\theta, \theta^*)$ that declines smoothly in θ. Details on the derivations are provided in Appendix D.

We incorporate private signals by modifying the prompts for the LLM investors as shown in Prompt 2. The investors observe their private signal and also know the generic signal structure and distributions for the fundamental and noise term. They also are told that

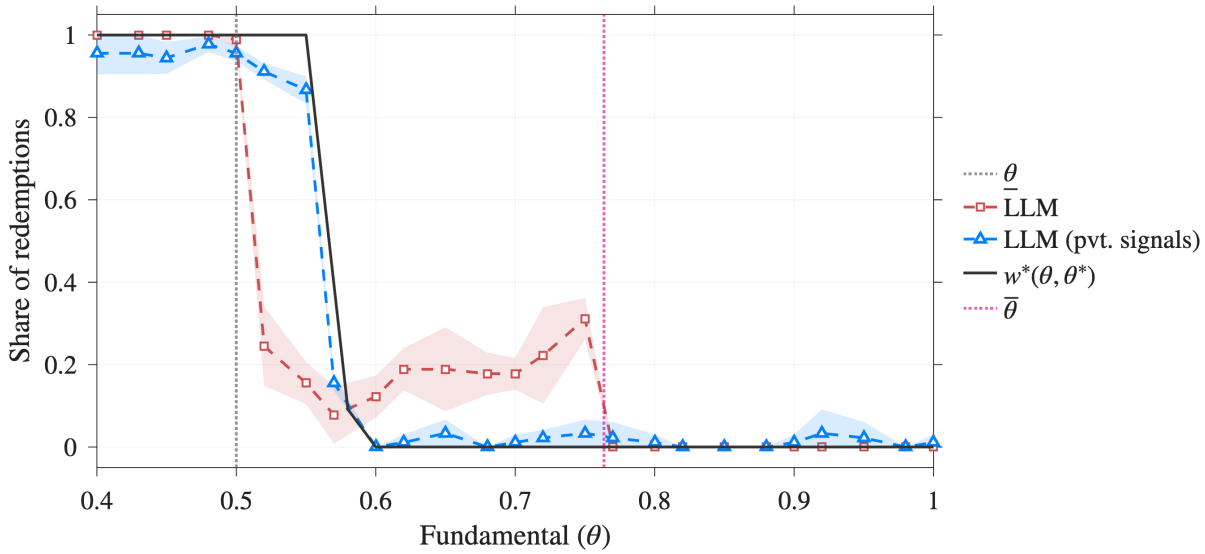


Figure 7: Share of redemptions as a function of the fundamental θ. The line labelled LLM refers to the baseline case with LLM investors; LLM (pvt. signals) refers to the case where LLM investors receive private signals about the fundamental, where the signal dispersion is $\eta = 0.01$; $w^*(\theta, \theta^*)$ is the theoretical prediction for the share of investors who redeem given the threshold θ^* and the current value of the fundamental. The vertical lines $\underline{\theta} = 0.5$ and $\overline{\theta} = 0.78$ are the bounds defined in Proposition 1.

other investors receive private signals.

```
10 The fundamental value θ (theta) of the fund is randomly drawn from
     the interval [0,1] but you do not directly observe it. All values
      of θ are equally likely.
11 Instead you receive a private signal x_i defined as x_i = θ + ϵ_i,
     where ϵ_i is drawn uniformly from [−η, η] with η = \%f. Signals of
     different investors are drawn independently. Your private signal
     is x_i = \%f.
12 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 2: Introducing fundamental uncertainty: The baseline prompt is supplemented with additional information after line 9.

Figure 7 compares the simulated outcomes with LLM investors to the theoretical benchmark in the presence of fundamental uncertainty.

We readily note that LLM investors closely align with the theoretical prediction, coordinating their redemption decisions around θ^* and therefore closely matching the average redemption rate $w(\theta^*, \theta)$ predicted by the theory.

By adding additional context to the prompt about the signal structures, LLM investors seem to recognize the structure of the underlying game and thus converge on the global games equilibrium. Hence, they achieve coordination under fundamental uncertainty by converging on the global games equilibrium. Figure 8 presents the causal graph for $\theta = 0.75$ and $\eta = 0.01$. The contrast with Figure 5 is striking: whereas investors exhibited substantial heterogeneity in both justifications and beliefs without fundamental uncertainty, the introduction of noisy signals induces convergence on a shared reasoning framework.

The justification for all LLM investors is equilibrium analysis. Each one derives the critical signal $x^* \approx 0.558$ via the standard global-games argument and compares its private signal against it. Given the high signal precision ($\eta = 0.01$) and a high fundamental ($\theta = 0.75$), every investor's posterior places all rival signals above the threshold, so each forms the belief that no other active investor redeems, yielding an expected payoff $\mathbb{E}[\text{stay}] \approx 2\theta \approx 1.49 > 1$. Twenty-eight investors act on this reasoning and stay. The two redemptions (investors 7 and 29) are *not* driven by adverse signals or divergent beliefs. Both draw high signals ($x_7 = 0.745$, $x_{29} = 0.752$), both derive the same threshold, and both explicitly conclude that staying is optimal. Yet each is recorded as redeeming, an action inconsistent with its own derived conclusion. We read these two cases not as genuine redemption decisions but as computation errors introduced when the LLMs' decisions are recorded.

B. Remedying belief heterogeneity with revealed history

A second route to coordinating beliefs is to give investors a common record of past play. Unlike private signals, history does not alter the game's payoffs or provide information about the fundamental: the multiplicity of the baseline remains intact in principle. What history supplies is a shared, public basis for forming beliefs about how others will act. By the logic of inductive, case-based reasoning (Argenziano and Gilboa, 2012), a commonly observed record of previous outcomes lets players coordinate on the equilibrium that past play has established. History functions here not as data from which investors learn an unknown payoff (as payoffs are already known) but purely as a coordination device: it tells each investor what others have done, and hence what others are likely to do. This generates a prediction that distinguishes history from private signals. Whereas signals select a *unique* threshold equilibrium pinned down by the information structure, history selects whichever equilibrium past play has favored: coordination is inherited from the realized record rather than determined

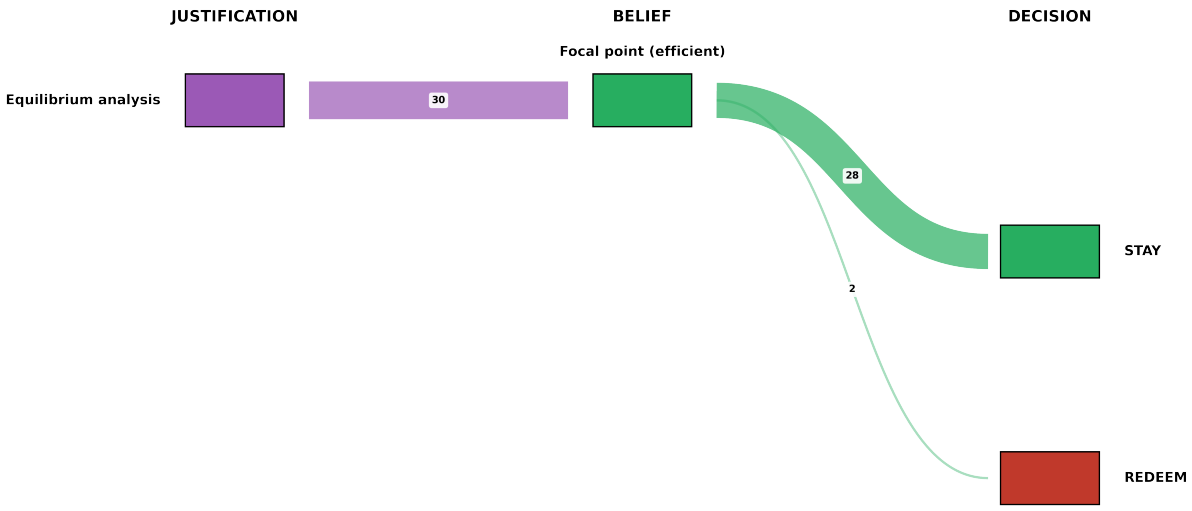


Figure 8: Causal mapping from justifications to beliefs and decisions where LLM investors receive private signals about the fundamental. The true level of the fundamental is $\theta = 0.75$ and the dispersion in signals is $\eta = 0.01$. The width of the links indicate investor counts.

by theory.

To implement this, we extend the interaction from the three rounds of the baseline to twenty, holding the fundamental θ fixed throughout. The first ten rounds are played blind, i.e., each investor decides without any record of past play, exactly as in the baseline. From the eleventh round onward, every investor is shown the full history of the preceding rounds. This includes the complete profile of actions and realized payoffs in each round. This information is appended to the prompt, with the record updating after every subsequent round.[15] Prompt 3 shows the appended block.

```
Previous rounds with this same θ value:
Round 1: Actions=[...], Payoffs=[...]
⋮
Round 10: Actions=[...], Payoffs=[...]
```

Prompt 3: Revealing history. From the eleventh round onward, the baseline prompt is supplemented with the record of past play at the same fundamental.

[15]The history reveals the full profile of actions and payoffs, not a summary statistic, so investors observe the actual distribution of play rather than an aggregate.

Figure 9 reports the outcome for $\theta = 0.75$.The top panel shows the share of redeeming investors across rounds. Over the first ten rounds, played blind, behavior reproduces our baseline belief heterogeneity result. The investors fail to coordinate, and the redemption share sits at an intermediate level. From the eleventh round, once the history is revealed, the investors coordinate. The bottom panel shows the mechanism: the common record selects a single belief, and the population converges on the corresponding equilibrium, consistent with selection by induction over past play (Argenziano and Gilboa, 2012).

The direction of coordination is inherited from the revealed history. When the average redemption over the blind phase falls below the critical $\widehat{W}$, i.e., the number of redemptions from the indifference condition, investors coordinate on all stay, while when it falls above, they coordinate on all redeem. The cutoff $\widehat{W}$ thus determines each investor's best response and also selects the equilibrium after history is revealed. Since $\theta = 0.75$, default events are relatively infrequent. And so by revealing the history over ten blind rounds, a few unlucky default events are smoothed over and investors coordinate on an unambiguous history.

C. Treating the Q-learning hot stove effect with synchronous updating

The source of the hot stove effect in Proposition 2 was asynchronous updating, i.e., each Q-learning investor updates only the value of the action it takes. So, once a sequence of defaults drives the Q-value of stay below that of redeem, the investor switches to redeeming. Crucially, the Q-value of stay remains *frozen* and the error is never corrected.

The hot stove effect can be addressed by providing Q-learning investors with counterfactual feedbacks. In each episode an investor observes not only the realized payoff of the action it took in the previous episode but also the payoff the unchosen action would have earned, and updates both action-values accordingly. Following Banchio and Mantegazza (2025) we call this *synchronous updating*, since both Q-values are revised every period rather than only the value of the action played.

As Figure 10 shows, synchronous updating does treat the hot stove effect. But a side effect is that the variance of outcomes is significantly higher, suggesting that coordination has been impaired. Proposition 4 formalizes both these points.

Proposition 4: *Under synchronous updating, for any $\theta < 1$, in the limit $t \to \infty$, the Q-value of stay is characterized by a stationary distribution with strictly positive variance. As a consequence, neither the all-stay nor the all-redeem equilibria are absorbing.*

The intuition behind this result is as follows. Under synchronous updating the Q-value for stay is updated every period regardless of the action taken. So it is an exponentially-weighted moving average of the i.i.d. payoffs. It is characterized by a stationary non-degenerate dis-

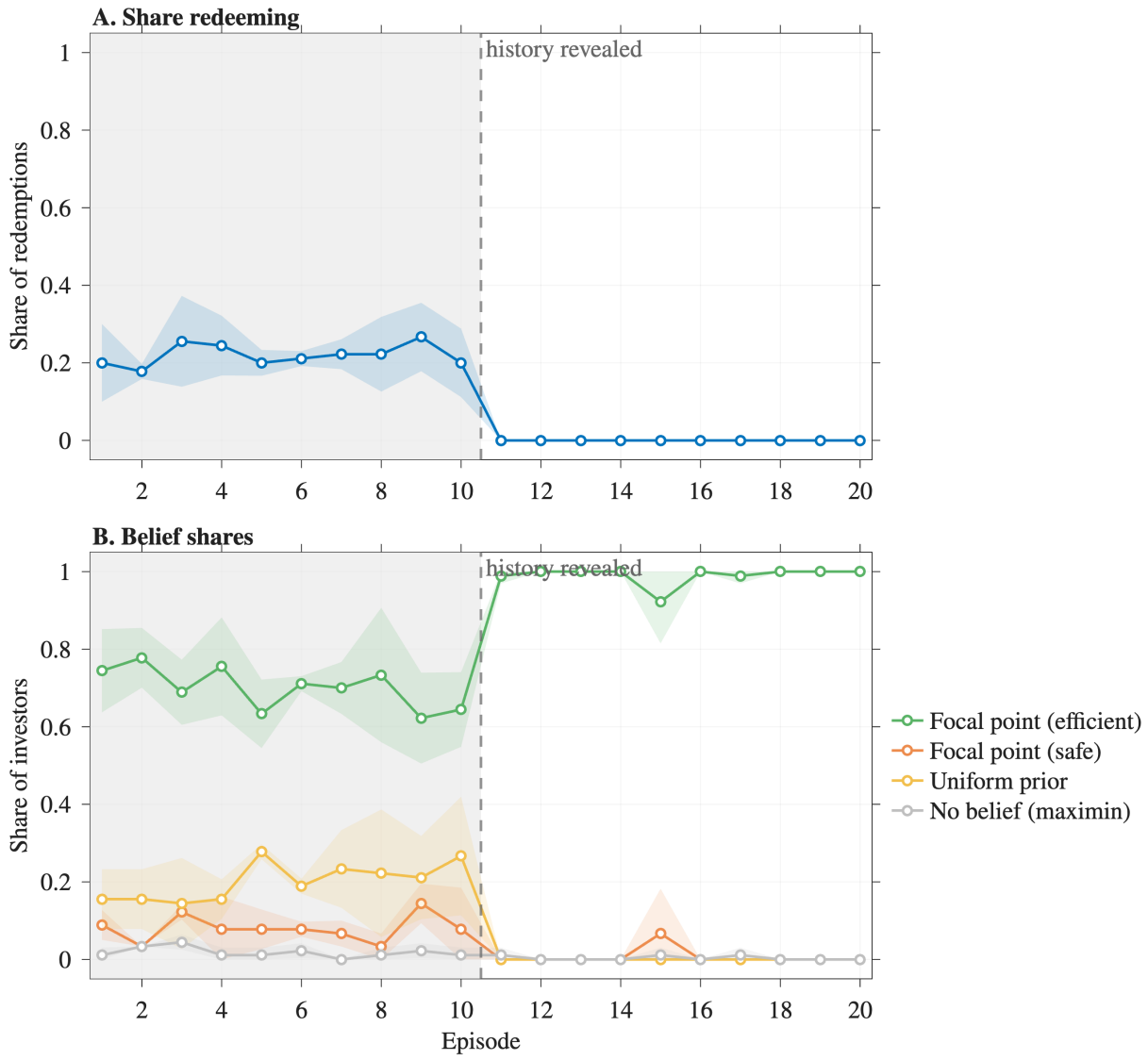


Figure 9: LLM investors with revealed history at $\theta = 0.75$. The top panel shows the share of redeeming investors across rounds; the bottom panel shows the share of investors holding each belief. For the first ten rounds investors play blind; from the eleventh round onward, the full record of past play is revealed.

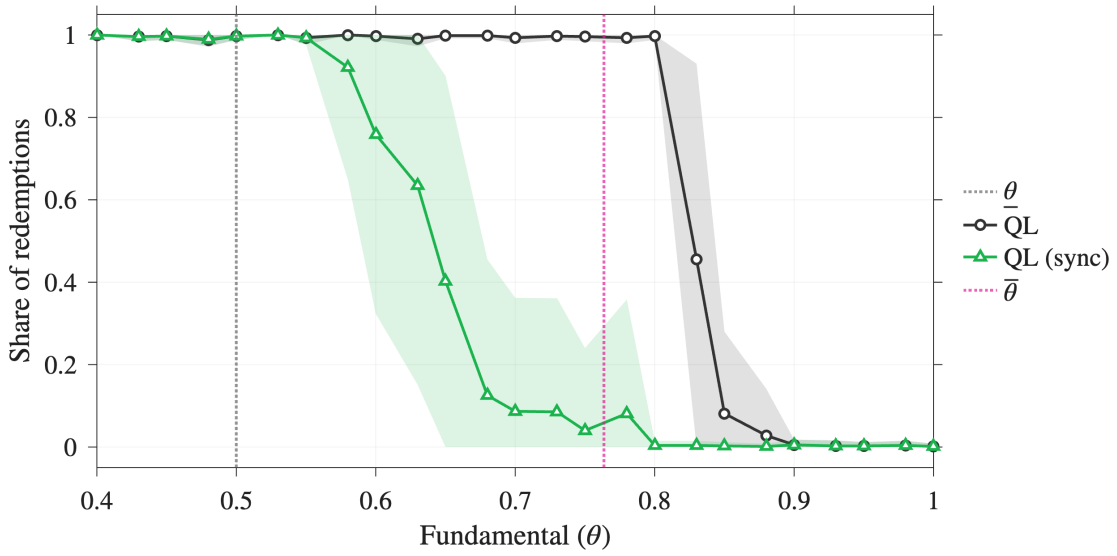


Figure 10: Share of redemptions as a function of the fundamental θ. The line labelled QL refers to the baseline case with Q-learning investors and asynchronous updating, while the line labelled QL(sync) represents the outcome using synchronous updating. The vertical lines $\underline{\theta} = 0.5$ and $\overline{\theta} = 0.78$ are the bounds defined in Proposition 1.

tribution with positive variance. An important consequence of this is that the Q-learning investors never converge to either the all-stay or all-redeem equilibria. For example, if all investors find the all-redeem equilibrium, there is now a positive probability that the Q-value of stay for some investors will rise above R_1 and thus switch to stay. And this switch is guaranteed to happen in finite time. A similar argument applies when starting from the all-stay outcome.

Thus, while synchronous updating helps to mitigate the hot stove effect by ensuring that all-redeem is no longer an absorbing state, it still does not lead to an efficient outcome. In fact, the outcome is one where Q-learning investors never settle on any outcome and keep oscillating between all-stay and all-redeem. We see this clearly in Figure 11 which shows, for several different values of θ, how the share of redemptions evolves across episodes. In particular, for values of θ in the intermediate region, we see that across different runs the share of redemptions wildly fluctuates and never settles. But as θ increases the extent of these fluctuations decreases.

Our result contrasts that of Banchio and Mantegazza (2025). In their setting, with a single dominant-strategy equilibrium, correct counterfactual estimation resolves the selection problem entirely. But, in our setting with multiple equilibria it cannot. This is because under default risk, the stationary distribution for the stay payoff has strictly positive variance. As a

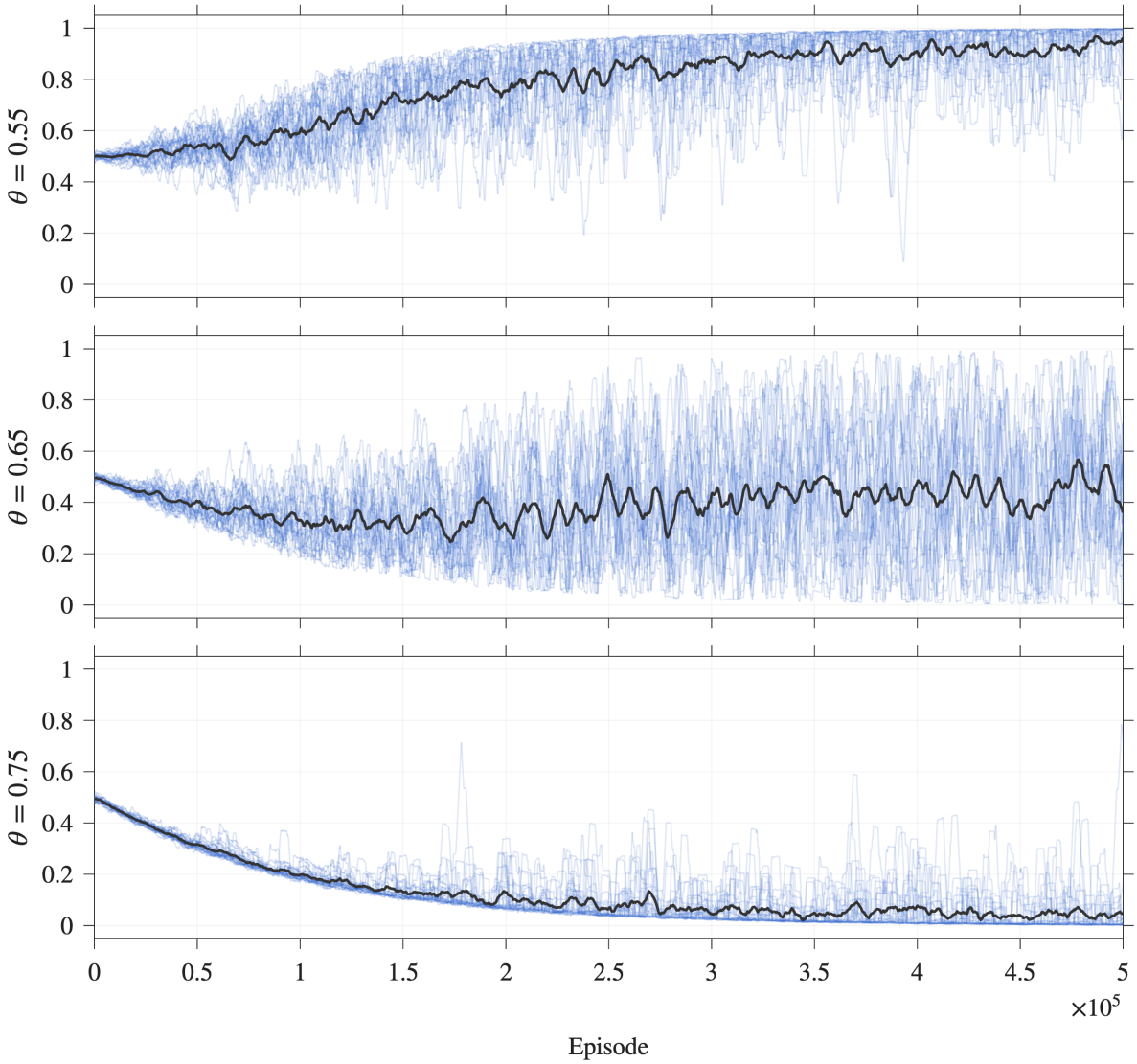


Figure 11: Evolution of the share of redemptions across episodes for three values of the fundamental (top panel $\theta = 0.55$; middle panel $\theta = 0.65$; bottom panel $\theta = 0.75$), with synchronous updating. Each blue line is the 5,000-episode rolling average of redemptions for one of the 25 training rounds; the thick black line is the mean of these rolling averages across all 25 rounds.

consequence, there is a strictly positive probability of deviating from an equilibrium outcome.

VII. Conclusions

We study how AI agents behave in settings where outcomes depend not only on economic fundamentals but also on what other agents are expected to do. In these settings, the same economic fundamental is consistent with multiple possible outcomes. We use a mutual fund redemption game as a canonical example of a framework with these characteristics. Fund redemptions depend on the state of the economy, but also on the number of investors who redeem. Holding payoffs, information, and incentives fixed, we find that AI architecture alone select equilibrium outcomes. Q-learning agents coordinate successfully, but frequently on the wrong outcome: they redeem even when fundamentals are strong. LLM agents redeem less overall, but do not converge on a common action because they form different beliefs about what others will do.

The broader lesson is that AI architecture is not just an implementation choice: even with the same fundamentals, payoffs, and information, different systems can produce different (and often inefficient) aggregate outcomes. Investors and institutions, therefore, need to understand how the AI tools they use shape financial decisions.

For policymakers, the key implication is that financial stability may depend not only on financial incentives, but also on which AI systems investors and institutions adopt. For example, investor protection tools (such as MiFID suitability questionnaires) could assess whether investors understand not only financial risk but also the AI systems guiding their decisions. Stress tests could examine how different AI architectures behave under pressure, while disclosure rules could require institutions to explain how those systems shape investor behavior.

Market safeguards should also work across architectures. Circuit breakers, trading halts, and related measures can limit destabilizing feedback loops (Breeden, 2026), especially when AI agents react rapidly and in correlated ways. Traditional prudential tools remain useful, but they may need to be supplemented by safeguards designed for the speed and scale at which AI agents trade.

More broadly, AI coordination failures may arise wherever strategic interaction matters, including autonomous vehicles, supply chains, and resource allocation. Traditional safeguards may therefore need to be complemented by tools that monitor and stress test interactions among AI agents. Even when each agent is aligned with its own objective, groups of agents can still produce inefficient outcomes (Lorè and Heydari, 2024). Our findings thus highlight the importance of multi-agent alignment, not just individual system alignment.

REFERENCES

Abada, Ibrahim, and Xavier Lambin, 2023, Artificial intelligence: Can seemingly collusive outcomes be avoided?, *Management Science* 69, 5042–5065.

Aldasoro, Iñaki, Leonardo Gambacorta, Anton Korinek, Vatsala Shreeti, and Merlin Stein, 2024, Intelligent Financial System: How AI is Transforming Finance, BIS Working Paper No. 1194.

Argenziano, Rossella, and Itzhak Gilboa, 2012, History as a coordination device, *Theory and Decision* 73, 501–512.

Assia, Yoni, 2026, Introducing agent portfolios, eToro, Insights from the CEO, Accessed 30 June 2026.

Banchio, Martino, and Giacomo Mantegazza, 2025, Artificial intelligence and spontaneous collusion, *arXiv preprint arXiv:2202.05946* .

Banchio, Martino, and Andrzej Skrzypacz, 2022, Artificial Intelligence and Auction Design.

Bank of England, 2025, Financial Stability in Focus: Artificial Intelligence in the Financial System, Technical report, Bank of England.

Battiston, Stefano, Michelangelo Puliga, Rahul Kaushik, Paolo Tasca, and Guido Caldarelli, 2012, Debtrank: Too central to fail? financial networks, the fed and systemic risk, *Scientific reports* 2, 541.

Bebchuk, Lucian, and Itay Goldstein, 2011, Self-Fulfilling Credit Market Freezes, *Review of Financial Studies* 24, 3519–3555.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, 2021, On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623 (Association for Computing Machinery).

Bernardo, Antonio E., and Ivo Welch, 2004, Liquidity and financial market runs, *The Quarterly Journal of Economics* 119, 135–158.

Bhagwat, Vineet, J Anthony Cookson, Chukwuma Dim, and Marina Niessner, 2025, The Market's Mirror: Revealing Investor Disagreement with LLMs, FEB-RN Research Paper No. 107.

Breeden, Sarah, 2026, Agents of change, Speech given at the ECB Sintra Forum 2026.

Brunnermeier, Markus K., and Lasse Heje Pedersen, 2009, Market liquidity and funding liquidity, *Review of Financial Studies* 22, 2201–2238.

Bybee, J. Leland, 2023, The Ghost in the Machine: Generating Beliefs with Large Language Models, arXiv preprint arXiv:2305.02823.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolo, and Sergio Pastorello, 2020, Artificial Intelligence, Algorithmic Pricing, and Collusion, *American Economic Review* 110, 3267–3297.

Carlsson, Hans, and Eric Van Damme, 1993, Global Games and Equilibrium Selection, *Econometrica* 61, 989–1018.

Cecchetti, Stephen G, Robin L Lumsdaine, Tuomas Peltonen, and Antonio Sánchez Serrano, 2025, Artificial intelligence and systemic risk, *ESRB: Advisory Scientific Committee Reports* 16.

Chen, Qi, Itay Goldstein, Zeqiong Huang, and Rahul Vashishtha, 2024, Liquidity Transformation and Fragility in the US Banking Sector, *Journal of Finance* 79, 3985–4036.

Chen, Qi, Itay Goldstein, and Wei Jiang, 2010, Payoff Complementarities and Financial Fragility: Evidence from Mutual Fund Outflows, *Journal of Financial Economics* 97, 239–62.

Colliard, Jean-Edouard, Thierry Foucault, and Stefano Lovo, 2025, Algorithmic Pricing and Liquidity in Securities Markets, CEPR Discussion Paper No. 17606.

Cook, Thomas R, and Sophia Kazinnik, 2025, Social Group Bias in AI Finance, arXiv preprint arXiv:2506.17490.

Cook, Thomas R., Sophia Kazinnik, Anne Lundgaard Hansen, and Peter McAdam, 2023, Evaluating Local Language Models: An Application to Financial Earnings Calls, *Federal Reserve Bank of Kansas City Research Working Paper* no. 23-12.

Cook, Thomas R, Sophia Kazinnik, Zach Modig, and Nathan M Palmer, 2026, What do llms want? .

Danielsson, Jon, and Andreas Uthemann, 2025, Artificial Intelligence and Financial Crises, Journal of Financial Stability, forthcoming.

Denrell, Jerker, and James G March, 2001, Adaptation as information restriction: The hot stove effect, *Organization science* 12, 523–538.

Diamond, D., and P. Dybvig, 1983, Bank Runs, Deposit Insurance and Liquidity, *Journal of Political Economy* 91, 401–19.

Dou, Winston, Jiantao Wu, and Yan Jiang, 2024, AI-Powered Trading, Algorithmic Collusion, and Price Efficiency, Jacobs Levy Equity Management Center Working Paper.

Dou, Winston Wei, Itay Goldstein, and Yan Ji, 2025a, AI-powered Trading, Algorithmic Collusion, and Price Efficiency, NBER Working Paper No. 34054.

Dou, Winston Wei, Itay Goldstein, and Yan Ji, 2025b, Financial market fragility in the era of ai planning, *Available at SSRN 5763222* .

Dou, Winston Wei, Itay Goldstein, Zigang Li, and Liyan Yang, 2025c, Human edge, machine limits: Ai-human competition in financial markets, *Machine Limits: AI-Human Competition in Financial Markets (November 18, 2025)* .

eToro, 2025, Retail investors flock to AI tools, with usage up 46% in one year, Retail Investor Beat, press release, eToro, Survey of 11,000 retail investors across 13 countries, conducted by Opinium, 5–19 August 2025. Accessed 30 June 2026.

Even-Tov, Omri, Ben Lourie, Kimberlyn Munevar, and Alexander Nekrasov, 2025, The effect of AI on retail investor behavior: Evidence from account-level trading data, Working paper, SSRN, Available at SSRN.

Farmer, J. Doyne, and Duncan Foley, 2009, The economy needs agent-based modelling, *Nature* 460, 685–686.

Fedyk, Anastassia, Ali Kakhbod, Peiyao Li, and Ulrike Malmendier, 2024, AI and Perception Biases in Investments: An Experimental Study, Available at SSRN 4787249.

Financial Stability Board, 2024, The Financial Stability Implications of Artificial Intelligence, Technical report, Financial Stability Board, Report to the G20.

Fish, Sara, Yannai A Gonczarowski, and Ran I Shorrer, 2024, Algorithmic collusion by large language models, *arXiv preprint arXiv:2404.00806* 7.

Foley-Fisher, Nathan C., Borghan Narajabad, and Stephane H. Verani, 2020, Self-fulfilling Runs: Evidence from the U.S. Life Insurance Industry, *Journal of Political Economy* 128, 3520–3678.

Foucault, Thierry, Leonardo Gambacorta, Wei Jiang, and Xavier Vives, 2025, Artificial Intelligence in Finance, Technical report, CEPR and IESE Banking Initiative.

Gao, Shen, Yuntao Wen, Minghang Zhu, Jianing Wei, Yuhan Cheng, Qunzi Zhang, and Shuo Shang, 2024, Simulating Financial Market via Large Language Model Based Agents, arXiv preprint arXiv:2406.19966.

Goldstein, Itay, Hao Jiang, and David T Ng, 2017, Investor Flows and Fragility in Corporate Bond Funds, *Journal of Financial Economics* 126, 592–613.

Goldstein, Itay, and Ady Pauzner, 2005, Demand Deposit Contracts and the Probability of Bank Runs, *Journal of Finance* 60, 1293–1327.

Gorton, Gary B., Elizabeth C. Klee, Chase P. Ross, Sharon Y. Ross, and Alexandros P. Vardoulakis, 2025, Leverage and Stablecoin Pegs, *Journal of Financial and Quantitative Analysis* .

Guarino, Antonio, Philippe Jehiel, and James Symons-Hicks, 2025, Q-Learning and Algorithmic Market Making: Loss-free, Collusive, or Competitive Prices?, CEPR Discussion Paper No. 20461.

Hansen, Anne Lundgaard, John J Horton, Sophia Kazinnik, Daniela Puzzello, and Ali Zarifhonarvar, 2024, Simulating the Survey of Professional Forecasters, SSRN Working Paper.

Heinemann, Frank, Rosemarie Nagel, and Peter Ockenfels, 2004, The theory of global games on test: experimental analysis of coordination games with public and private information, *Econometrica* 72, 1583–1599.

Horton, John J, Apostolos Filippas, and Benjamin S Manning, 2023, Large language models as simulated economic agents: What can we learn from homo silicus?, Technical report, National Bureau of Economic Research.

Kazinnik, Sophia, 2023, Bank Run, Interrupted: Modeling Deposit Withdrawals with Generative AI, SSRN Working Paper.

Leitner, Georg, Jaspal Singh, Anton van der Kraaij, and Balazs Zsamboki, 2024, The Rise of Artificial Intelligence: Benefits and Risks for Financial Stability, Technical report, European Central Bank, Financial Stability Report.

Liu, Xuewen, and Antonio S. Mello, 2011, The Fragile Capital Structure of Hedge Funds and the Limits to Arbitrage, *Journal of Financial Economics* 102, 491–506.

Lopez-Lira, Alejandro, 2025, Can Large Language Models Trade? Testing Financial Theories with LLM Agents in Market Simulations, SSRN Working Paper.

Lorè, Nunzio, and Babak Heydari, 2024, Strategic Behavior of Large Language Models and the Role of Game Structure versus Contextual Framing, *Scientific Reports* 14, 18490.

Morris, Stephen, and Hyun Song Shin, 1998, Unique Equilibrium in a Model of Self-Fulfilling Currency Attacks, *American Economic Review* 88, 587–597.

Morris, Stephen, and Hyun Song Shin, 2002, Measuring Strategic Uncertainty, Manuscript.

Ouyang, Shumiao, Hayong Yun, and Xingjian Zheng, 2025, Ai as decision-maker: Risk preferences of llms, *Available at SSRN 4851711* .

Shabsigh, Ghiath, and El Bachir Boukherouaa, 2023, Generative Artificial Intelligence in Finance: Risk Considerations, Fintech Note 2023/006, International Monetary Fund.

Yang, Hao, 2024, AI Coordination and Self-fulfilling Financial Crises, Working Paper.

Zarifhonarvar, Ali, 2024, Experimental Evidence on Large Language Models, SSRN Working Paper.

Online Appendix A

Proof of Proposition 1. For extreme values of the fundamental θ, investors have a dominant action. We start with the low values of θ. Irrespective of what other investors choose, redeeming at $t = 1$ is optimal for an investor when even the highest payoff they can accrue at $t = 2$, which corresponds to the payoff accrued if no other investors redeem, is smaller than R_1. Formally, this is the case when $R_1 R\theta < R_1$, that is, when $\theta < \underline{\theta} \equiv \frac{1}{R}$.

Symmetrically, staying until $t = 2$ is a dominant action when even the worst payoff that an investor expects to receive at the final date, which corresponds to the payoff accrued if all $A-1$ investors redeem, is larger than R_1. Formally, this is the case when $R_1 R\theta \frac{N-(A-1)(1+\lambda)}{N-(A-1)} > R_1$ that is when $\theta > \overline{\theta} = \frac{1}{R}\frac{N-(A-1)(1+\lambda)}{N-(A-1)}$.

When $\theta \in [\underline{\theta}, \overline{\theta}]$, both redeeming at $t = 1$ and not redeeming are equilibria. If an investor expects others to redeem, it is optimal for them to redeem as well, since staying until $t = 2$ would yield a lower payoff than R_1. Conversely, if no one else redeems at $t = 1$, it is optimal not to redeem either, as the payoff from staying, $R_1 R\theta$, exceeds R_1 for all $\theta > \underline{\theta}$. This completes the proof. □

Proof of Proposition 2. We prove the result for a single Q-learning investor and a fixed state $\theta < 1$. The results directly extend to all states since Q-values are updated state-by-state. We discuss how the result extends to A investors is discussed at the end.

Throughout we suppress the investor index i and the state argument θ, writing $Q_t(a)$ for the Q-value of action $a \in \{a_R, a_S\}$ at episode t and $\pi_t(a)$ for its realized reward. In this notation the update is $Q_{t+1}(a) = (1-\alpha)Q_t(a) + \alpha\, \pi_t(a)$, with $\pi_t(a_R) = R_1$ and $\pi_t(a_S) = Z_t B_t$, where $B_t = R\, R_1\, (N - W(1+\lambda))/(N-W)$ and $Z_t \in \{0, 1\}$ is the default indicator with $\Pr(Z_t = 0) = 1 - \theta$, drawn independently across dates; exploration follows $\varepsilon_t = \beta^t$ and updating is asynchronous.

Two invariants. Each Q-value stays within the interval spanned by its initialization and the range of its rewards, since the update is a convex combination of the previous value and a new reward. For the redeem value, the reward is the constant R_1, so $Q_{t+1}(a_R) = (1-\alpha)Q_t(a_R) + \alpha R_1$ is a convex combination of $Q_t(a_R)$ and R_1; by induction,

$$Q_t(a_R) \;\geq\; \underline{r} := \min\{Q_0(a_R),\, R_1\} \qquad \text{for all } t. \tag{A1}$$

For the stay value, the reward satisfies $0 \leq \pi_t(a_S) \leq \overline{B}$ with $\overline{B} := R\, R_1$ (the best case $W = 0$, $Z_t = 1$), so $Q_{t+1}(a_S)$ is a convex combination of $Q_t(a_S)$ and a quantity at most $\overline{B}$; by induction,

$$Q_t(a_S) \;\leq\; \overline{Q} := \max\{Q_0(a_S),\, \overline{B}\} \qquad \text{for all } t. \tag{A2}$$

Since $\underline{r} < \overline{Q}$ in general, staying may be preferred initially. The question is whether a run of defaults can overturn this permanently.

A finite run of defaults locks in redemption. If the investor stays and the fund defaults ($Z_t = 0$), then $\pi_t(a_S) = 0$ and the stay value contracts geometrically, $Q_{t+1}(a_S) = (1-\alpha)Q_t(a_S)$. Iterating over K consecutive defaults and bounding the starting value by the

ceiling $\overline{Q}$ gives $Q_t(a_S) \leq (1-\alpha)^K \overline{Q}$. The condition $(1-\alpha)^K \overline{Q} < \underline{r}$ is equivalent, taking logarithms and noting $\log(1-\alpha) < 0$, to $K > \log(\underline{r}/\overline{Q})/\log(1-\alpha)$; hence any run of at least

$$K^* := \left\lfloor \frac{\log(\underline{r}/\overline{Q})}{\log(1-\alpha)} \right\rfloor + 1 \tag{A3}$$

consecutive defaults drives $Q_t(a_S) < \underline{r} \leq Q_t(a_R)$. Once this happens while the investor is greedy, the greedy action is a_R. Under asynchronous updating, redeeming updates only $Q_t(a_R)$ and leaves $Q_t(a_S)$ unchanged; since redeeming pays the constant R_1, the value $Q_t(a_R)$ remains a convex combination of quantities at least $\underline{r}$ and so stays at or above $\underline{r} > Q_t(a_S)$. The greedy action therefore remains a_R forever, and the stay value is never re-sampled to correct the gap: redemption is an absorbing state of the greedy dynamics. It remains to show that the investor eventually becomes permanently greedy and that a run of K^* defaults does occur.

Exploration ceases in finite time. The exploration probabilities are summable, $\sum_{t\geq 0} \varepsilon_t = \sum_{t\geq 0} \beta^t = 1/(1-\beta) < \infty$. Letting E_t be the event that the investor explores at date t, so $\Pr(E_t) = \varepsilon_t$, the first Borel–Cantelli lemma gives $\Pr(E_t \text{ infinitely often}) = 0$. Thus almost surely only finitely many E_t occur, and $T_E := 1 + \sup\{t : E_t \text{ occurs}\}$ is almost surely finite; for $t \geq T_E$ the investor never explores and always plays $\arg\max_a Q_t(a)$.

A lock-in run is inevitable. Fix a realization with $T_E < \infty$, an almost sure event. From T_E onward the investor is greedy. If it ever redeems at some $t \geq T_E$, then by the absorption argument above its stay value is frozen while its redeem value stays at or above $\underline{r}$, so it redeems forever and we are done. Otherwise the investor stays at infinitely many dates $t \geq T_E$. At each such date the return carries an independent default draw with $\Pr(\text{default}) = 1-\theta > 0$, since $\theta < 1$. Partition these stay-dates into consecutive disjoint blocks of size K^* and call block j *bad* if all K^* of its draws default, so the blocks are independent with $\Pr(A_j) = (1-\theta)^{K^*} =: p > 0$. The probability that none of the first J blocks is bad is $(1-p)^J \to 0$ as $J \to \infty$, so almost surely some block consists entirely of defaults.[16] By the second step, this run of K^* consecutive defaults while staying and greedy drives $Q_t(a_S) < \underline{r} \leq Q_t(a_R)$ and locks in redemption from that date on.

Extension to A investors. The argument above conditions only on a single investor's own default draws and own actions. Three observations extend the conclusion to the full system. First, the ceiling invariant is unaffected: $B_t \leq \overline{B} = R\,R_1$ for every $W \geq 0$, so $\overline{Q}$ remains a valid upper bound on each investor's stay value whatever the others do. Second, the contraction-on-default $Q_{t+1}(a_S) = (1-\alpha)Q_t(a_S)$ and the count K^* in (A3) depend on no other investor, so the absorption mechanism operates investor-by-investor. Third, defaults occur whenever the realized return is zero, an event of probability $1-\theta > 0$ at every date for every investor as long as $\theta < 1$; hence, conditional on any action path of the others, each investor almost surely experiences a run of K^* defaults while staying and greedy, and locks

[16] Equivalently, since the A_j are independent and $\sum_j \Pr(A_j) = \infty$, the second Borel–Cantelli lemma gives $\Pr(A_j \text{ infinitely often}) = 1$; one bad block suffices.

into redemption. Applying this to each investor in turn, the whole system reaches all-redeem in almost surely finite time.

Strategic complementarity tightens the result rather than weakening it. The stay reward B_t is decreasing in W: as other investors lock into redemption and W rises, the realized stay value of any remaining investor contracts from a lower level, so the threshold $\underline{r}$ is crossed after fewer defaults. Lock-in is therefore reached weakly earlier than in the single-investor benchmark. □

Proof of Proposition 3. As before, we fix a state θ, suppress the investor index, and write $Q_t(a)$ for the value of action $a \in \{a_R, a_S\}$ at episode t.

We characterize the dynamics under a mean-field approximation, which is exact in the small-gain limit $\alpha \to 0$ and for a large investor population. Concretely, we replace each investor's stochastic update with its conditional expectation, and summarize the actions of the other symmetric investors through the redemption propensity, p via $\Phi(p) = \mathbb{E}_{W\sim \mathrm{Bin}(A-1,p)}[\phi(W)]$. for finite α and A used in our simulations, we treat $\tilde{\theta}$ as an approximation.

We consider a mean-field approximation of the Q-learning dynamics, whereby the representative investor's Q-values evolve under the expected (mean) update, with the other investors taken to be symmetric. Denote $\phi(W) = (N - W(1+\lambda))/(N-W)$ and its mean $\Phi(p) = \mathbb{E}_{W\sim \mathrm{Bin}(A-1,p)}[\phi(W)]$, where p is the independent redemption probability for each of the other investors. It follows that $\tilde{\theta} = 1/(R\,\Phi(\frac{1}{2}))$.

Fix $\theta \in (\underline{\theta}, \overline{\theta})$ with $|\theta - \tilde{\theta}| \geq \eta$, where the excluded width $\eta = \eta(\alpha, \beta)$ is pinned down in Step 2 and vanishes as $-\ln\beta/\alpha \to 0$. Let $g_t = \mathbf{1}\{Q_t(a_R) > Q_t(a_S)\}$ indicate that redeem is greedy; tie-breaking is immaterial away from indifference. Under uniform exploration, the probability that the representative investor redeems is

$$p_t = \tfrac{1}{2}\varepsilon_t + (1-\varepsilon_t)g_t, \tag{A4}$$

which is a convex combination of the exploration value $\frac{1}{2}$ and the current greedy action. By symmetry the expected stay reward is $R_1 R\theta\,\Phi(p_t)$ and the redeem reward is R_1.

Step 1: learning erases initial conditions during high exploration. An action's Q-value is updated only when that action is sampled. Denote by $s_t, r_t \in [\varepsilon_t/2, 1]$ the sampling probabilities of stay and redeem, respectively. So, with probability s_t, stay is played at the Q-value for stay moves by a fraction α toward the reward. While with the converse probabilty, $1 - s_t$, redeem is played and so the Q-value for stay is not updated. So, averaging over the sampling outcomes, we get

$$\mathbb{E}[Q_{t+1}(a_S) \mid Q_t] = (1-\alpha s_t)Q_t(a_S) + \alpha s_t\, R_1 R\theta\,\Phi(p_t) \tag{A5}$$

$$\mathbb{E}[Q_{t+1}(a_R) \mid Q_t] = (1-\alpha r_t)Q_t(a_R) + \alpha r_t\, R_1. \tag{A6}$$

Each time the Q-values update, they contracts toward the conditional expected reward. With fixed α, the realized Q-values fluctuate at order $\sqrt{\alpha}$ around the mean path. These fluctuations do not affect the sign of the expected payoff difference since we characterize behavior a distance η away from the indifference point $\tilde{\theta}$. After T steps the initial Q-value

for stay is scaled by $\prod_{t<T}(1-\alpha s_t)$ (resp. for redeem with $1-\alpha r_t$). Define

$$T_\star \equiv \frac{1}{\sqrt{\alpha\,(-\ln\beta)}}, \tag{A7}$$

such that under the assumption $-\log\beta \ll \alpha$, it follows that

$$\alpha T_\star = \sqrt{\tfrac{\alpha}{-\ln\beta}} \gg 1, \qquad (-\ln\beta)T_\star = \sqrt{\tfrac{-\ln\beta}{\alpha}} \ll 1. \tag{A8}$$

The second inequality ensures that exploration remains high, i.e., $1-\beta^t = 1-e^{t\ln\beta} \le t(-\ln\beta) \le (-\ln\beta)T_\star \ll 1$ for $t \le T_\star$, so $s_t, r_t \ge \varepsilon_t/2 \ge c$ for a constant $c > 0$. The cumulative contraction is then at most $(1-\alpha c)^{T_\star} \le \exp(-c\,\alpha T_\star) = \exp(-c\sqrt{\alpha/(-\ln\beta)})$, small by the first inequality in (A8). This implies that, since the initial Q-values are drawn from the unit interval, their contribution becomes vanishly small by $T_\star$, implying that initial conditions are washed out.

Step 2: the value learned is the equal-randomization target, giving the cutoff $\tilde\theta$. Over the same window, $t \le T_\star$, the propensity to redeem, p_t, stays near $\frac{1}{2}$. We can see this from $|p_t - \frac{1}{2}| = (1-\varepsilon_t)|g_t - \frac{1}{2}| \le \frac{1}{2}(1-\beta^t) \le \frac{1}{2}(-\ln\beta)T_\star \ll 1$. So the conditional expected reward that the mean-field equations contract towards are

$$Q_{T_\star}(a_S) = R_1 R\theta\,\Phi(\tfrac{1}{2}) + O(\eta), \qquad Q_{T_\star}(a_R) = R_1 + O(\eta), \qquad \eta(\alpha,\beta) = O\Big(\sqrt{\tfrac{-\ln\beta}{\alpha}}\Big), \tag{A9}$$

where the big-O correction terms reflect the residues of the intialization of the Q-values. Hence

$$Q_{T_\star}(a_S) - Q_{T_\star}(a_R) = c_\theta + O(\eta), \qquad c_\theta \equiv R_1 R\theta\,\Phi(\tfrac{1}{2}) - R_1. \tag{A10}$$

Since $\Phi(\frac{1}{2})$ is a positive constant, c_θ is strictly increasing in θ that crosses zero at exactly $\theta = \tilde\theta$. It thus follows that $|c_\theta| = \kappa|\theta - \tilde\theta|$, for some constant $\kappa > 0$. Thus, as long as θ is at least η away from $\tilde\theta$, the sign of c_θ indicates the greedy action for the investor. In the limit $-\ln\beta/\alpha \to 0$, where $\eta \to 0$, the switch occurs at exactly $\tilde\theta$.

Step 3: the selected action persists as exploration fades. Steps 1–2 fix the greedy action at $T_\star$. We now check it does not reverse once the investor stops exploring. Take $\theta > \tilde\theta$, so stay is greedy at $T_\star$. After the window the redeem propensity falls from $\frac{1}{2}$ toward 0 since the greedy action is stay and $\varepsilon_t \to 0$. The mean value for stay reaches a maximum $R_1 R\theta\,\Phi(0)$, while the mean value for redeem stays at R_1. Since the stay value was already above the redeem value at $T_\star$, the gap has only gotten wider and hence stay remains the greedy choice and the population settles into all-stay. The case $\theta < \tilde\theta$ is symmetric, with $p_t \uparrow 1$, the stay target falling, and all-redeem selected. □

Proof of Proposition 4. We fix a state $\theta < 1$ and write $Q_t(a)$ for the Q-value of action $a \in \{a_R, a_S\}$ at episode t, suppressing the investor index. The only change from the baseline is the updating rule. Under *synchronous* updating both action-values are revised every period, using the realized payoff of the action taken and the counterfactual payoff of the action

not taken. We first establish the stationary law of the stay Q-value at a fixed number of redemptions, W (Step 1), and then use it to show neither corners are absorbing (Steps 2–3).

Throughout, the redeem reward is the constant R_1, so $Q_t(a_R) \to R_1$ and we treat the redeem-value as fixed at R_1 in the limit. The stay reward at redemption count W is $\pi_t(a_S) = Z_t B(W)$, where $B(W) = R\,R_1\,(N - W(1+\lambda))/(N-W)$ and $Z_t \in \{0,1\}$ is the common default indicator with $\Pr(Z_t = 1) = \theta$, drawn independently across episodes.

Step 1: stationary law of the stay-value. Because synchronous updating revises the stay-value every period regardless of the action taken, and the redemption count is held at W, the stay-value evolves as

$$Q_{t+1}(a_S) = (1-\alpha)Q_t(a_S) + \alpha\, Z_t B(W), \tag{A11}$$

which is an exponentially-weighted moving average of the i.i.d. stay payoffs. Iterating,

$$Q_t(a_S) = (1-\alpha)^t Q_0(a_S) + \alpha \sum_{k=1}^{t} (1-\alpha)^{t-k} Z_k B(W). \tag{A12}$$

The first term $(1-\alpha)^t Q_0(a_S) \to 0$ becomes vanishingly small since $\alpha < 1$. The weighted sum converges in distribution to a stationary limit $Q_\infty(a_S)$ with

$$\begin{aligned} \mathbb{E}[Q_\infty(a_S)] &= \theta B(W) && \text{(A13)} \\ \mathrm{Var}(Q_\infty(a_S)) &= \alpha^2 B(W)^2\,\theta(1-\theta) \sum_{k=0}^{\infty} (1-\alpha)^{2k} && \text{(A14)} \\ &= \frac{\alpha}{2-\alpha}\, B(W)^2\,\theta(1-\theta). && \text{(A15)} \end{aligned}$$

For $\theta \in (0,1)$ and $\alpha \in (0,1)$ this variance is strictly positive. The constant gain prevents the stay-value from settling at a point, since each common shock Z_t keeps it dispersed. For small α a central-limit argument on the geometrically-weighted sum gives the Gaussian approximation

$$Q_\infty(a_S) \;\approx\; \mathcal{N}\Big(\theta B(W),\; \frac{\alpha}{2} B(W)^2 \theta(1-\theta)\Big), \tag{A16}$$

so that $\Pr(Q_\infty(a_S) < R_1)$ and $\Pr(Q_\infty(a_S) > R_1)$ are both strictly positive whenever $\theta B(W)$ is finite, i.e. for all $\theta < 1$.

Step 2: all-redeem is not absorbing. Consider the population at the all-redeem configuration, so each investor faces $W = A - 1$ co-redeemers and stay payoff $B(A-1) > 0$. Under *asynchronous* updating this configuration is absorbing. Synchronous updating removes this. Because the stay-value is updated through the counterfactual even while the investor redeems, it converges in law to $Q_\infty(a_S)$ of Step 1 with $W = A-1$, dispersed around $\theta B(A-1)$.

Let $p \equiv \Pr\big(Q_\infty(a_S) > R_1\big)$. By Step 1, $p > 0$. An investor whose stay-value exceeds R_1 has greedy action a_S and switches to staying. For a population of A investors, the probability that at least one switches in a given episode is at least $1 - (1-p)^A > 0$. By the second Borel Cantelli lemma, this switch occurs infinitely often, hence almost surely in finite time. Thus

some investor leaves all-redeem almost surely and hence the all redeem configuration is not absorbing.

Step 3: all-stay is not absorbing. The argument is symmetric. At all-stay every investor faces $W = 0$ and stay payoff $B(0) = R\,R_1$. The stay-value converges in law to $Q_\infty(a_S)$ with $W = 0$, dispersed around $\theta B(0)$. Let $q \equiv \Pr\big(Q_\infty(a_S) < R_1\big)$; by Step 1, $q > 0$ for all $\theta < 1$. An investor whose stay-value falls below R_1 has greedy action a_R and redeems. The probability that at least one of the A investors redeems in a given episode is at least $1 - (1 - q)^A > 0$, and by the same Borel–Cantelli argument a redemption occurs almost surely in finite time. All-stay is therefore not absorbing either. □

Online Appendix B

In this appendix, we provide the prompts used to extract beliefs and causal reasoning from the reasoning text of the LLM investors. There are two prompts: one to extract the beliefs of the LLM investors (Prompt 4) and the second to cluster them into canonical game-theoretic categories (Prompt 5).

```
You are analyzing an investor's reasoning in a coordination game (mutual fund redemption).
CONTEXT:
The investor was given this prompt:
\{investor\_prompt\}
Their FINAL DECISION was: \{decision\}.
TASK:
The investor's reasoning trace may contain multiple considerations, calculations, and hypothetical beliefs. Your job is to identify:
    1. The DECISIVE BELIEF -- the specific belief about other investors' behavior that ultimately determined their final decision
    2. The JUSTIFICATION -- the reasoning that led them to form this belief
    3. The CAUSAL CHAIN from justification -> belief -> decision
Note: The investor might explore multiple scenarios (e.g., "if others redeem with p=0.5..." or "in the all-stay equilibrium...") but ultimately commits to ONE belief that drives their final action. Find that committed belief.
REASONING TRACE:
\{explanation\}.
INSTRUCTIONS:
    1. Read the ENTIRE reasoning trace
    2. Identify where the investor COMMITS to a specific belief about others
    3. Identify what JUSTIFICATION led them to form this belief
    4. Extract the exact text passage that reveals the belief commitment
    5. Construct the causal DAG
```

Prompt 4: Prompt to extract the beliefs and causal reasoning of the LLM investors.

```
You are analyzing causal reasoning patterns extracted from 30 investors.
Below are the causal DAGs (justification -> belief -> decision) for each investor, described in free-form text. Each entry also includes:
```

```
    - evidence\_quote: The exact text from the reasoning trace
      supporting the extraction
    - confidence: How clearly the causal chain was stated (HIGH/
      MEDIUM/LOW)
Your task is to:
    1. Identify the distinct categories of JUSTIFICATIONS present in
       the data (use EXACTLY 3-4 categories)
    2. Identify the distinct categories of BELIEFS present in the
      data
    3. Assign each investor to one justification category and one
      belief category
IMPORTANT: Be PARSIMONIOUS with justification categories. Use only
   3-4 high-level categories. Merge similar reasoning patterns.
Reference list of standard JUSTIFICATION categories (consolidate
   into these):
    - Payoff comparison: Reasoning based on comparing payoffs across
       equilibria (includes Pareto efficiency, strong fundamentals,
       threshold calculations)
    - Strategic uncertainty: Reasoning based on uncertainty about
      others' behavior (includes uninformative priors, no
      information)
    - Worst-case avoidance: Focused on guaranteed minimum payoff, no
       belief about others (maximin reasoning)
    - Equilibrium analysis: Formal equilibrium concepts (includes
      mixed equilibrium analysis, deviation loss comparison)
Reference list of standard BELIEF categories (use these if they fit)
   :
    - Focal point (efficient): Believe others coordinate on Pareto-
      efficient equilibrium (all-stay)
    - Focal point (safe): Believe others coordinate on risk-dominant
       equilibrium (all-redeem)
    - Uniform prior: Assume equal probability of others' actions, no
       specific expectation
    - Mixed-strategy equilibrium: Believe others play the symmetric
      mixed equilibrium probability
    - No belief: No probabilistic belief formed (used with worst-
      case/maximin reasoning)
INVESTOR DATA:
\{investor\_data\}
```

Prompt 5: Prompt to cluster the extracted justifications and beliefs of the LLM investors into canonical categories.

Appendix C. LLM behavior without default risk and alternative prompts

We consider three alternative prompt structures. The first removes default risk and replaces the payoff from staying with the expected payoff from the game. Prompt 6 shows the modification made to the baseline prompt. The second and third prompts retain the no-default-risk formulation but vary the framing. Prompt 7 is a *neutral context prompt* that removes the fund-redemption framing entirely, while Prompt 8 is a *narrative prompt* that preserves the fund-redemption context but presents the information in natural language, requiring investors to compute the payoff of each action from the scenario description. Exploring these framings serves two goals. First, we want to rule out that LLM investors' redemption decisions reflect pre-trained knowledge of a particular class of strategic games rather than genuine reasoning about economic fundamentals and strategic uncertainty. Second, we want to better understand the drivers of LLMs' reasoning and to examine whether redemption decisions are sensitive to how the problem is framed.

```
3 If you do not redeem your share, i.e., you stay, then you earn
     fraction * R1 * R * \theta, where:
```

Prompt 6: No default risk: The baseline prompt is modified on line 3.

```
1 You are one of A active participants in a group of size N. Each
     participant chooses between two actions. Your objective is to
     maximize your payoff.
2 PAYOFF STRUCTURE:
3 If you choose Action A: your payoff is R1
4 If you choose Action B: your payoff is θR(N − W(1 + λ))/(N − W), where:
5 - W ≤ A is the number of other active participants who choose Action
      A (the remaining {N - A} participants are passive and always
     choose Action B)
6 - λ is a parameter affecting how costly it is when more participants
     choose Action A
7 - θ reflects the underlying state of the environment
8 - R is the baseline payoff multiplier
9 Which action do you choose? Answer using: <decision>A</decision> or
     <decision>B</decision>
```

Prompt 7: Neutral context prompt.

```
1 You and {N-1} other investors have each invested \textdollar100,000
     in a mutual fund and have received shares. The fund invests the
     total amount raised based on its own investment strategy in a
     portfolio of assets.
```

```
The current value of the fund's assets is \textdollar{
    total_return_today:.2f}.
TODAY, you have to decide what to do with your shares.
If you REDEEM your shares today, you are guaranteed to receive \
    textdollar{individual_return_today:.2f}.
If you STAY invested, you will be repaid only after three months,
    and the amount you receive will depend on the value of the fund's
     assets at that time.
Based on current information about the fund and its strategy, if all
     investors remain invested, the value of the fund's assets is
    estimated to be \textdollar{future_return:.2f} after three months
    . So, each investor will receive \textdollar{
    future_return_per_investor:.2f}.
But, if some investors redeem today, the fund's asset value after
    three months will be strictly lower than \textdollar{
    future_return:.2f}, because fewer assets remain invested. The
    fund services redemptions by selling assets at a discount due to
    illiquid market conditions: to generate \textdollar50,000 of cash
     today, for example, the fund has to sell \textdollar{
    liquidity_cost_example:.2f} worth of assets, which eats into the
    fund's assets. The more investors redeem today, the larger the
    impact on the fund's assets for those who stay invested.
Out of a total of {N} investors, a number {A_bar}, which includes
    yourself, are active and in a position to decide today whether to
     redeem their shares or stay invested; the others will stay
    invested regardless.
You aim to maximize the repayment you receive based on the above
    information.
```

Prompt 8: Narrative prompt.

Figure 12 shows the share of redemptions as a function of the fundamental under four settings: (i) the baseline with default risk, (ii) no default risk, (iii) the context-free prompt, and (iv) the narrative prompt. The results are largely similar across all variants. We discuss each comparison with the baseline in turn.

Without default risk. Figure 13 depicts how the belief shares evolve with θ in the case without default risk. The pattern is again largely indistinguishable from the baseline: multiplicity of beliefs remains localized to intermediate values of θ. The inferred cutoff $\widehat{W}$ continues to track the theoretical threshold. In short, removing default risk leaves LLM behavior unchanged.

Context-free prompt. A notable difference emerges here: redemption rates are systematically lower than in the baseline, indicating better coordination on the efficient equilibrium. Figure 14 sheds light on this difference. The belief elicitation reveals that, under the context-free prompt, a larger share of investors hold optimistic beliefs in the intermediate

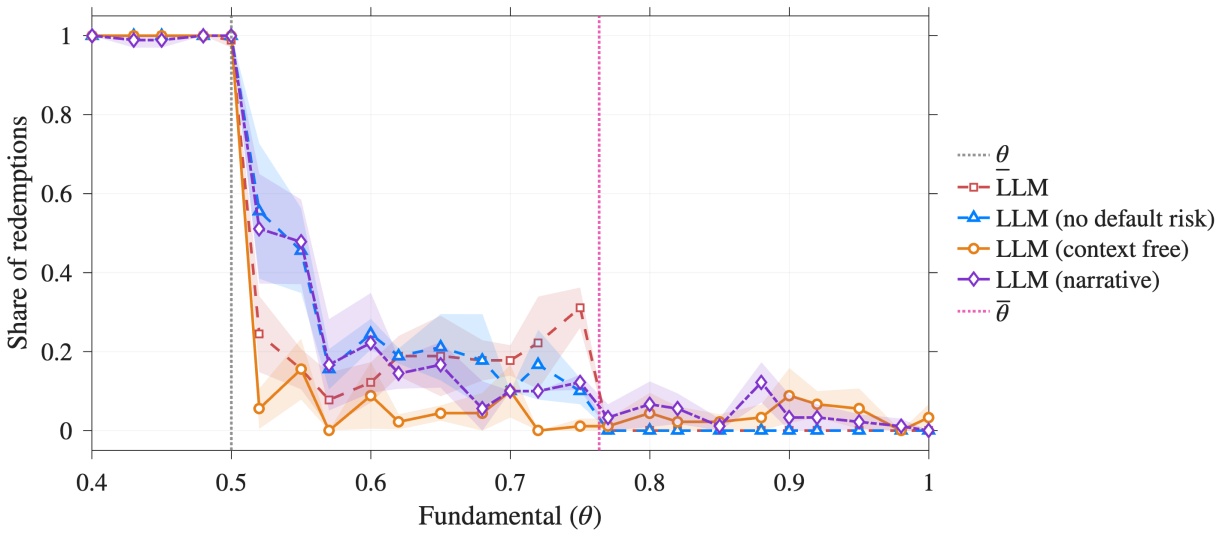


Figure 12: Share of redemptions as a function of the fundamental θ. The line labelled LLM refers to the baseline case; LLM (no default risk) represents the case using Prompt 6; LLM (context free) uses Prompt 7; and LLM (narrative) uses Prompt 8. The vertical lines $\underline{\theta} = 0.5$ and $\overline{\theta} = 0.78$ are the bounds defined in Proposition 1.

region. This greater prevalence of "focal-point (efficient)" beliefs is consistent with the lower observed redemption rates.

We conjecture that this difference relates to the baseline prompt's explicit reference to monetary payoffs. Reinforcement learning from human feedback (RLHF) alignment has been shown to induce heightened risk aversion in LLMs (Ouyang et al., 2025). In our setting, greater risk aversion translates into a stronger incentive to redeem, since redemption offers a deterministic payoff. The financial framing of the baseline prompt may make risk–reward tradeoffs more salient, thereby activating this RLHF-induced conservatism. By contrast, the abstract framing of the context-free prompt may attenuate this effect, allowing LLMs to coordinate more readily on the efficient equilibrium.

Narrative prompt. We observe no significant differences in redemptions relative to the baseline prompt, suggesting that LLM investors are capable of computing the payoff associated with each action.

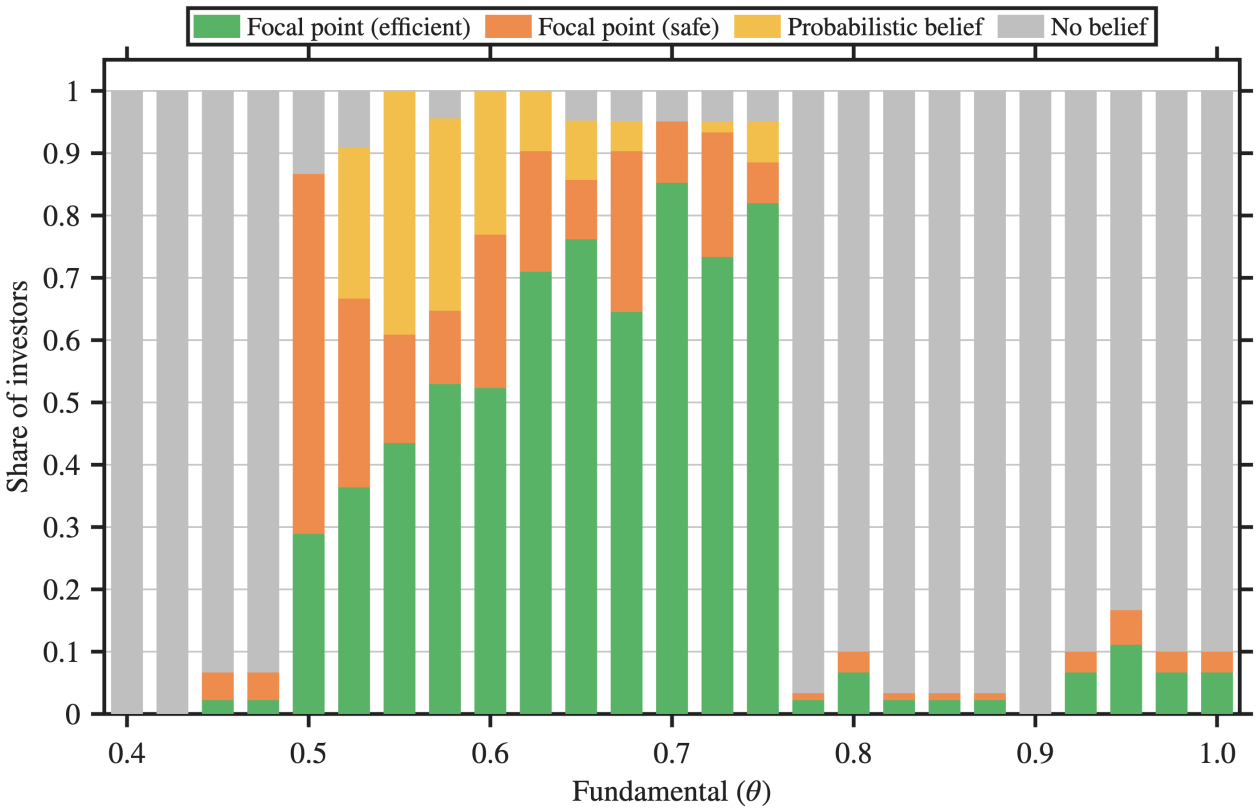


Figure 13: Evolution of LLM investors' beliefs about the behavior of other investors in the absence of default risk, as a function of the fundamental θ.

Appendix D. Global games theoretical benchmark

In this appendix, we provide the details of the derivation of the threshold θ^* that characterizes the redemption strategy of investors receiving private signals. For the sake of clarity, remember that each investor i receives a private signal $\theta_i = \theta + \eta_i$, where η_i are i.i.d. noise drawn from $[-\eta, +\eta]$ and the parameter η represents a measure of information dispersion.

The proof adapts the standard approach in the global game literature (Goldstein and Pauzner, 2005; Chen et al., 2010) to the case with a discrete number of players. The arguments in their proof establish that there is a unique equilibrium in which investors redeem at $t = 1$ if and only if the signal they receive is below a common threshold θ^*, which is the signal at which an investor is indifferent between redeeming at $t = 1$ and $t = 2$ given what he or she believes about the signals received by other investors and, in turn, their behaviors.

We start by characterizing investors' decisions in two extreme ranges of fundamentals where investors have a dominant action. These two ranges correspond to the one characterized in Proposition 1. When $\theta < \underline{\theta}$ redeeming at $t = 1$ is a dominant action. When $\theta > \overline{\theta}$, redeeming at $t = 2$ is the dominant action.

Consider now the intermediate region where $\theta \in [\underline{\theta}, \overline{\theta}]$. Assume that investors behave according to a threshold strategy: they redeem their shares if they receive a signal below

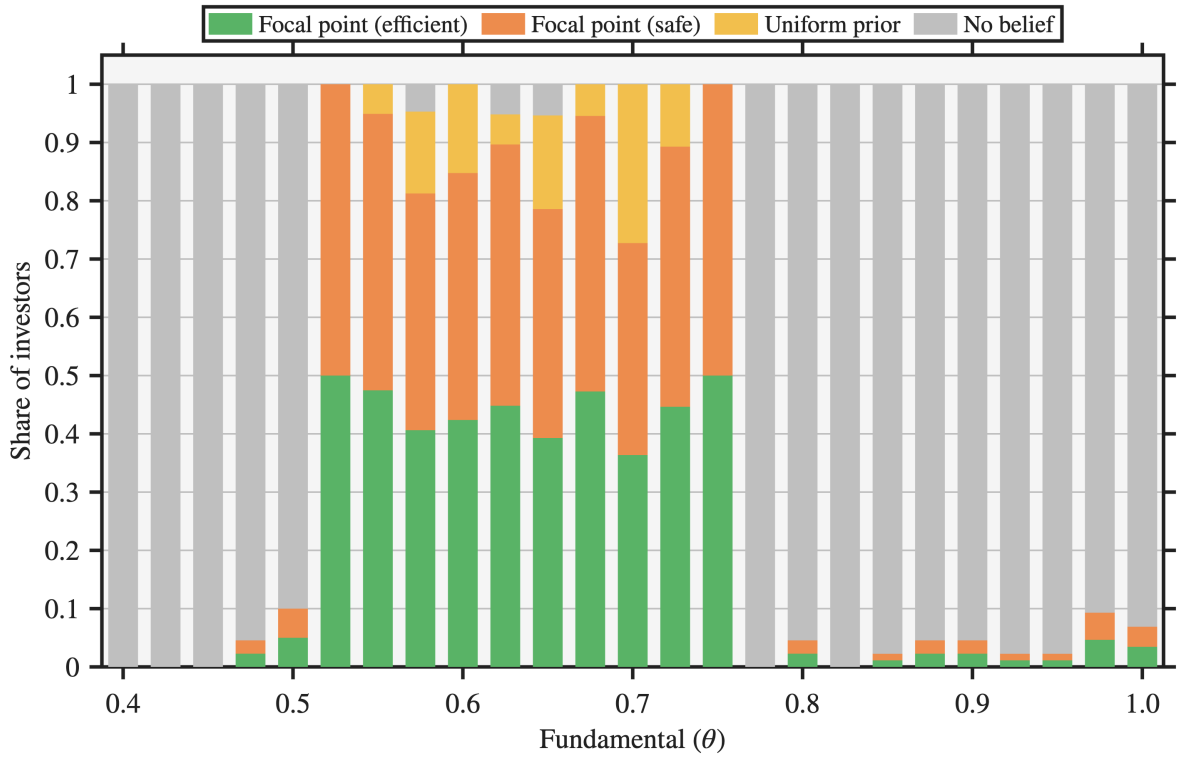


Figure 14: Evolution of LLM investors' beliefs about the behavior of other investors under the context-free prompt, as a function of the fundamental θ.

θ^* and stay until $t = 2$ otherwise.[17] Given that the signal is uniformly distributed over the interval $[-\eta, +\eta]$, the probability of receiving a signal below θ^* is $\frac{\theta^*-\theta+\eta}{2\eta}$. Building on this, we can compute the share of investors receiving the signal below the cutoff, which is given by

$$w = \sum_{i=1}^{A} \{\theta_i < \theta^*\} \sim \text{Bin}\left(A, \frac{\theta^* - \theta + \eta}{2\eta}\right). \tag{D1}$$

It follows that the probability that W out of A investors have received a signal below θ^* is given by:

$$f(W, A) = \binom{A}{W} \left(\frac{\theta^* - \theta + \eta}{2\eta}\right)^{W} \left(1 - \frac{\theta^* - \theta + \eta}{2\eta}\right)^{A-W}. \tag{D2}$$

Using the above, we can compute the probability that the investor receiving the signal θ^*

[17] This comes at no loss of generality, as Goldstein and Pauzner (2005) show that in this game every equilibrium strategy is a threshold strategy.

assigns to n out of $A-1$ investors redeeming at $t=1$ as:

$$\frac{\binom{A-1}{W}}{2\eta}\int_{\theta^*-\eta}^{\theta^*+\eta}\frac{(\theta^*-\theta+\eta)^W(\theta-\theta^*+\eta)^{A-1-W}}{(2\eta)^{A-1}}d\theta \tag{D3}$$

As shown in Morris and Shin (2002), this probability is equal to $\frac{1}{1+A-1}$. Hence, the indifference condition that characterizes the run threshold θ^* reads:

$$\sum_{W=0}^{A-1}\frac{1}{A}\frac{N-W(1+\lambda)}{N-W}R_1R\theta=R_1, \tag{D4}$$

which gives

$$\theta^*\equiv\frac{A}{\sum_{W=0}^{A-1}\frac{N-W(1+\lambda)}{N-W}R}, \tag{D5}$$

Finally, the average share of investors redeeming is:

$$\mathbb{E}[W]=\sum_{w=0}^{A}w\binom{A}{w}p^w(1-p)^{A-w}, \tag{D6}$$

where $p=\frac{\theta^*-\theta+\eta}{2\eta}$. Using the identity $w\binom{A}{w}=A\binom{A-1}{w-1}$, we obtain:

$$\mathbb{E}[W]=\sum_{w=1}^{A}A\binom{A-1}{w-1}p^w(1-p)^{A-w}=Ap\sum_{w=1}^{A}\binom{A-1}{w-1}p^{w-1}(1-p)^{(A-1)-(w-1)}=Ap, \tag{D7}$$

where the second-to-last step comes from using the binomial theorem. Then, we obtain

$$w^*(\theta,\theta^*)=\frac{\mathbb{E}[W]}{A}=\begin{cases}0 & \text{if} \quad \theta>\theta^*+\eta,\\ \frac{\theta^*-\theta+\eta}{2\eta} & \text{if} \quad \theta\in[\theta^*-\eta,\,\theta^*+\eta],\\ 1 & \text{if} \quad \theta<\theta^*-\eta.\end{cases} \tag{D8}$$