21 MAY 2026 · RESEARCH BULLETIN NO. 143

Financial stability in the age of artificial intelligence: the role of algorithmic architecture

by Kartik Anand, Sophia Kazinnik, Agnese Leonello and Ettore Panetti[1]

Artificial intelligence (AI) is rapidly transforming financial decision-making. To explore the implications for financial stability we ran simulation-based experiments on two different AI architectures. We found that Q-learning algorithms, a form of reinforcement learning, achieved a high degree of coordination, but were prone to extreme bank run-like dynamics. In contrast, large language models, which rely on contextual reasoning, were less prone to such runs but generated heterogeneous and unpredictable behaviour. This suggests that AI architecture is itself a source of financial instability: algorithms operating in the same environment, pursuing the same goals, yield fundamentally different outcomes for financial stability.

Artificial intelligence (AI) is playing a growing role in finance. Algorithmic trading based on machine learning already accounts for 60-70% of equity transaction volumes in the United States and other major global markets (Foucault et al., 2025). Large language models (LLMs) are increasingly being used by retail investors to obtain financial advice (Gambacorta et al., 2025; Even-Tov et al., 2025). And agentic systems, capable of autonomously executing tasks with minimal human oversight, will soon further expand AI's presence in the financial sector (Chatterji et al., 2025). As AI becomes ever more deeply embedded in financial decision-making (Danielsson, 2025), a critical question emerges: what risks does it pose for financial stability?

Mapping AI decision-making to financial stability: a simulation-based approach

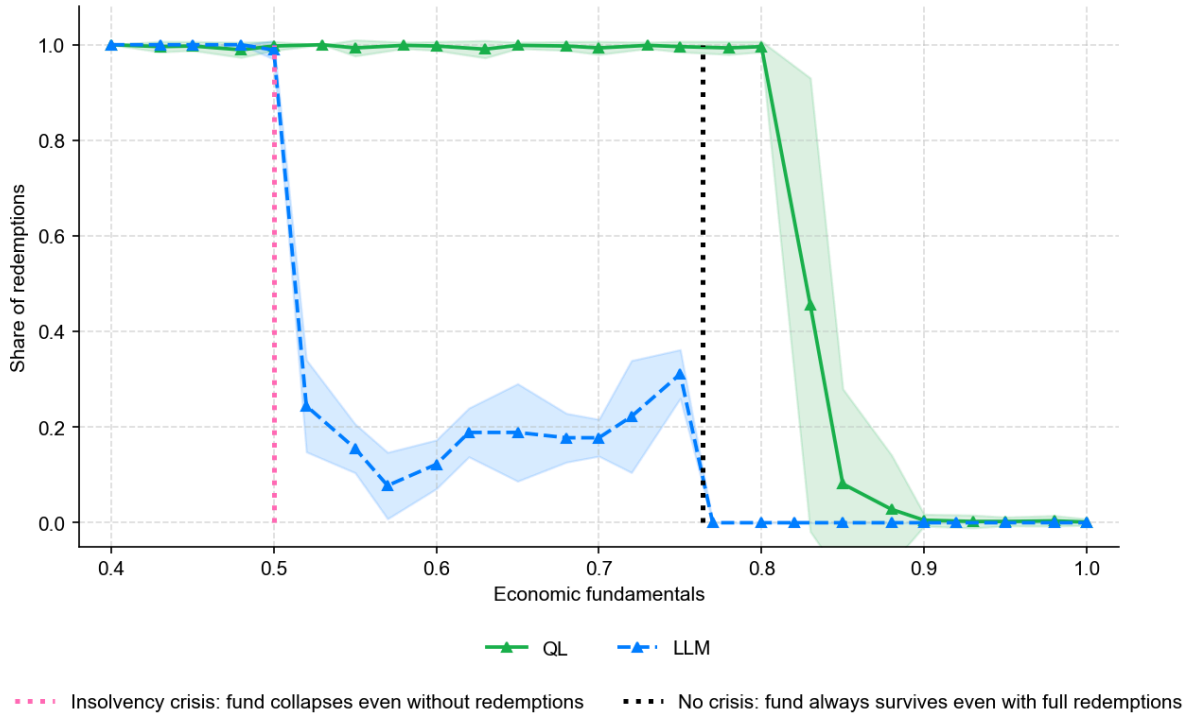
In Anand et al. (2025), we contribute to this debate by studying the consequences of AI financial decision-making for financial stability. Economic theory suggests that financial instability can result from large-scale investors' capital outflows. These can be driven by investors' own concerns about economic fundamentals, or the actions of other investors, or both. We capture these dynamics in a stylised mutual fund redemption game. Within this framework, we conduct experiments in which multiple independent and fully autonomous AI agents, acting as investors, have to decide whether to redeem (sell) their fund shares. We analyse their responses and, in turn, the implications for financial stability.[2]

Existing AI systems can be grouped into two broad categories based on their architecture, i.e. how they make decisions. First there are reinforcement learning systems, commonly used in algorithmic trading, which rely on iterative trial and error learning to make decisions. In our simulations, we use *Q-learning (QL)* algorithms to represent this architecture. The second category consists in *large language models (LLMs)* which, by contrast, generate decisions by reasoning, using the context provided in a prompt and drawing on patterns learned during their training.

Our main insight is that financial stability in the age of AI may depend as much on the AI architecture itself as on the economic environment as captured by the level of economic fundamentals. Chart 1 shows the share of QL and LLM investors that redeem their fund shares at different levels of economic fundamentals. The results highlight significant differences between the two AI architectures. All Q-learning investors coordinate and redeem, even when economic fundamentals are strong and do not justify a redemption. By contrast, LLM investors never redeem when economic fundamentals are strong. Their redemptions, however, are more unpredictable as they struggle to coordinate when fundamentals are at an intermediate level.

Chart 1

Share of redemptions as a function of economic fundamentals



Source: Authors' simulations.

Notes: The figure illustrates how the share of redemptions for QL and LLM investors changes with the economic fundamentals. The dotted pink line identifies the cut-off level of fundamentals below which, irrespective of investors' redemption decisions, the fund is always insolvent. The dotted black line, by contrast, represents the cut-off level of fundamentals above which the fund can survive a full-scale redemption, so no redemptions should be observed.

How does AI architecture help explain redemptions?

The excessive redemptions generated by QL algorithms arise from a learning pattern called the *hot stove effect* (Denrell and March, 2001).[3] When default risk is present,[4] not redeeming their shares exposes investors to the possibility of a zero pay-off, whereas redeeming them produces a small but certain return. As a result, each default episode that the algorithm experiences during its trial-and-error learning reduces the value that QL investors assign to staying invested, and incentivises them to

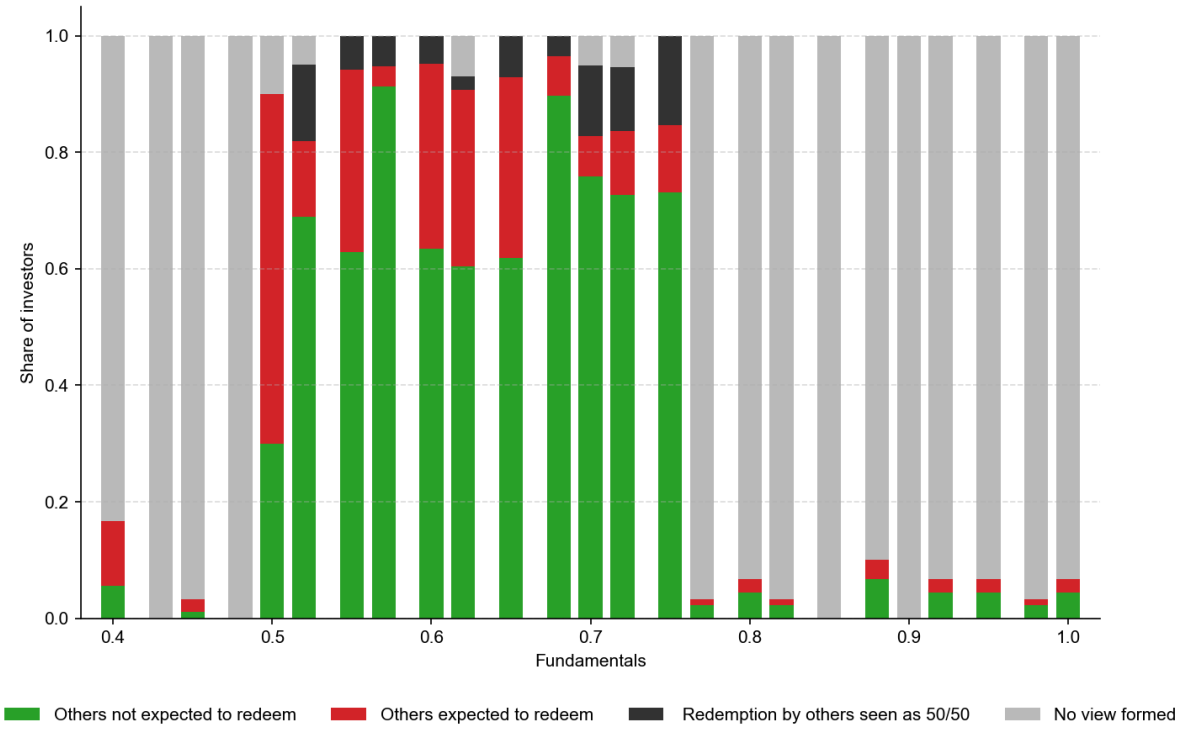
redeem.While suggestive of the ability of QL algorithms to coordinate, this outcome is intrinsically different from the collusive behaviour documented in the algorithmic pricing literature (see e.g. Calvano et al., 2020; Dou et al., 2025; and Colliard et al., 2026). In these papers, collusion emerges as the result of AI agents coordinating on the privately optimal outcome that produces a “win-win” for the agents coordinating. Here, instead, coordination among QL algorithms traps them in a privately harmful “lose-lose” equilibrium as they all rush to redeem.

LLMs face a different challenge. Because they do not learn from realised pay-offs through repeated trial and error, their behaviour is unaffected by default risk. However, their unpredictability arises from arriving at differing “beliefs” (Chart 2).[5] This variation in beliefs is not a bug, but a feature of the reasoning process, reflecting the fact that economic theory, which LLMs use for reasoning, doesn’t provide a unique prediction. Their “noisy” answers reflect this so-called theoretical indeterminacy. For intermediate values of fundamentals, theory predicts that two equilibria may emerge: either all investors stay invested or they all redeem. In our experiment, when fundamentals fall within this range, LLMs actually form different beliefs about the actions of other LLM investors, despite being identical and receiving the same instructions. This happens because the prompt does not provide them with any basis for selecting among multiple equally plausible beliefs, so competing responses appear similarly likely.

Two observations confirm that beliefs drive redemption. First, in line with the theory, multiple beliefs do not tend to emerge at extreme levels of fundamentals. Second, as economic fundamentals improve, the share of LLMs that expect others not to redeem rises. As a result, overall, redemptions decrease as fundamentals improve.

Chart 2

Evolution of LLM investors' beliefs about other investors for different values of fundamentals



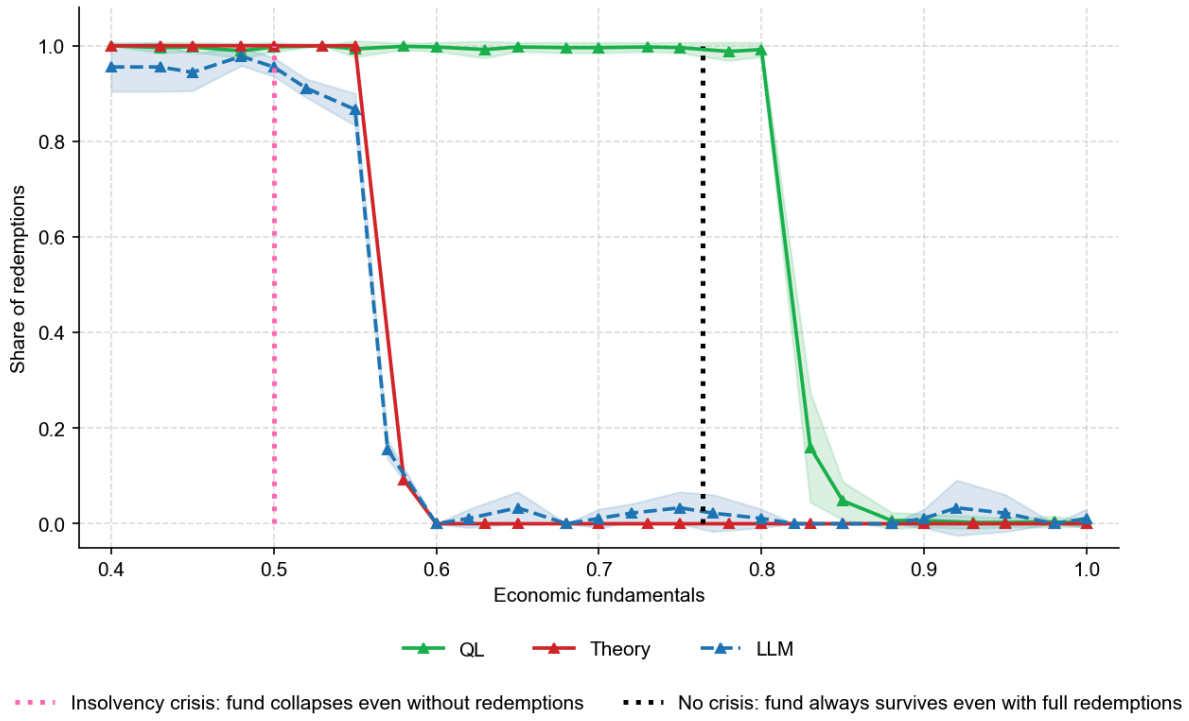
Source: Authors' simulations.

Notes: The figure illustrates the evolution of LLMs' beliefs about other investors as a function of the level of economic fundamentals.

The source of the unpredictability of LLMs' decisions suggests a possible remedy for their lack of coordination and the associated financial instability. When no criterion is provided for selecting among equally valid beliefs, private (i.e. not publicly available) noisy information about fundamentals can serve as an anchor for expectations (Goldstein and Pauzner, 2005), helping agents choose between competing responses. Our experiments confirm this prediction. Feeding different LLMs with private noisy signals about economic fundamentals makes their beliefs more similar and leads LLMs to take the same action. In contrast, QL investors are unresponsive to such signals because their behaviour is shaped by trial-and-error rather than by how accurate the information about economic fundamentals is (Chart 3).

Chart 3

LLMs' beliefs converge when private noisy signals are introduced



Source: Authors' simulations.

Notes: The figure illustrates how the share of redemptions for QL and LLM investors changes with the economic fundamentals when investors receive private noisy signals about economic fundamentals, rather than all investors receiving the same information with no additional signals as "tiebreakers" for LLMs weighing up similar, equally likely options.

Conclusions

The growing use of AI agents in financial decision-making may change how fragilities emerge in the financial system. Learning-based systems can trigger extreme events akin to panic runs, while reasoning-based systems may create risks through weaker coordination and lower predictability. As AI becomes more deeply embedded in financial decision-making, understanding these differences and the mechanisms behind them will be essential for assessing the implications for financial stability and designing appropriate policy responses.

This has implications for all stakeholders, from retail investors to regulators. As understanding AI tools and the underlying architecture becomes increasingly important for informed financial decision-making, retail investors may need technological, in addition to financial literacy. Financial institutions may need to know whether and how their customers are using AI tools as part of risk management. And regulators may need to incorporate measures of technological competence into investor protection frameworks, alongside traditional indicators such as risk tolerance and financial knowledge, for example in questionnaires required under the Markets in Financial Instruments Directive (MiFID).[6]

Finally, safeguarding financial stability may also require market design tools, such as circuit breakers, to help curb excessive disinvestment during market turmoil.

References

Anand, K., Kazinnik, S., Leonello, A. and Panetti, E., (2025), "Ex Machina: Financial Stability in the Age of Artificial Intelligence", *ECB Working Papers,* No 3225.

Calvano, E., Calzolari, G, Denicolò, V. and Pastorello, S., (2020), "Artificial Intelligence, Algorithmic Pricing, and Collusion", *American Economic Review,* Vol. 110, No 10, pp. 3267-3297.

Chatterji, A., Cunningham, T., Deming, D.J., Hitzig, Z., Ong, C., Shan, C.Y., and Wadman, K., (2025), "How people use ChatGPT", *NBER Working Paper Series, No 34255.*

Colliard, J.-E., Foucault, T. and Lovo, S., (2026), "Algorithmic Pricing and Liquidity in Securities Markets", *The Review of Financial Studies,* forthcoming.

Danielsson, J., (2025), "AI and financial stability", VoxEU.org, 6 February.

Denrell, J., and March, J.G., (2001), "Adaptation as Information Restriction: The Hot Stove Effect", *Organization Science,* Vol. 12, No 5, pp. 523-538.

Dou, W.W., Goldstein, I., and Ji, Y., (2025), "AI-Powered Trading, Algorithmic Collusion, and Price Efficiency", The Wharton School Research Paper.

Even-Tov, O., Lourie, B., Munevar, K. and Nekrasov, A., (2025) "The Effect of AI on Retail Investor Behavior: Evidence from Account-Level Trading Data", mimeo.

Foucault, T., Gambacorta, L., Jiang, W. and Vives, X., (2025), "Artificial Intelligence in Finance", *CEPR and IESE Banking Initiative*.

Gambacorta, L., Jappelli, T. and Oliviero, T., (2025), "Exploring household adoption and usage of generative AI: new evidence from Italy", *BIS Working Papers,* No 1298.

Goldstein, I, and Pauzner, A., (2005), "Demand-Deposit Contracts and the Probability of Bank Runs", *Journal of Finance*, Vol. 60, No 3, pp. 1293-1327.

1.

This article was written by Kartik Anand (Deutsche Bundesbank), Sophia Kazinnik (Stanford University), Agnese Leonello (Directorate General Research, European Central Bank and CEPR) and Ettore Panetti (University of Naples Federico II). The authors gratefully acknowledge the comments of Alex Popov and Zoë Sprokel. The views expressed here are those of the authors and do not necessarily represent the views of the European Central Bank, the Deutsche Bundesbank or the Eurosystem.

2.

Using this framework allows us to extend our insights to other settings where financial fragility is driven by investors' expectations about both economic fundamentals and other investors' actions. These include bank runs, currency attacks and stablecoin runs.

3.

This mechanism is captured by the following quote from Mark Twain: “A cat that sits on a hot stove lid will never sit on a hot stove lid again, but it will also never sit on a cold one”.

4.

The probability of default is measured by the realisation of the fundamentals of the economy and affects the repayment each investor expects to receive from the mutual fund, irrespective of what other investors do. High realisations mean that the probability of default is low and so the expected repayment from holding a share in the fund is high. Conversely, low realisations mean that the probability is high and expected repayment low.

5.

A key methodological contribution of the paper is the analysis of the chain-of-thought reasoning of LLM investors. We treat these texts as data and use an LLM to identify each AI investor’s belief about other investors’ actions and the underlying justification leading to a particular decision.

6.

Directive 2014/65/EU of the European Parliament and of the Council of 15 May 2014 on markets in financial instruments and amending Directive 2002/92/EC and Directive 2011/61/EU (OJ L 173, 12.6.2014, p. 349). The directive requires investment service providers to administer a standardised questionnaire to clients in order to gather information on their financial situation, knowledge and experience in financial markets, and investment objectives and risk tolerance.