Research on Emergent Coordination and the Market Fragility in AI Trading Systems

Xingru Du
Wuhan Britain China School, Wuhan, China
Email: duxingru2027@163.com
Manuscript received June 2, 2026; accepted August 20, 2026; published September 9, 2026.

***Abstract*—Autonomous artificial intelligence trading systems have captured an expanding share of global financial market transactions in recent years. Leading quantitative institutions widely deploy self-learning algorithms that process multi-dimensional market information and complete automated order execution with limited manual oversight. Extant literature on AI trading predominantly centers on algorithmic profit performance and transaction efficiency, yet few studies systematically analyze how interactive behaviors between independent trading agents erode overall market stability. This paper explores emergent coordination caused by AI trading systems and examines its potential impact on market fragility, investigating whether autonomous AI agents can unintentionally develop synchronized behaviors through shared data sources, strategy convergence, and feedback loops. The paper finds that such collective behavior may amplify market volatility, reduce liquidity, and increase systemic financial risk, even in the absence of explicit communication among market participants. Therefore, future regulatory frameworks should focus more on system-level AI interactions, as market instability may arise more from collective AI behavior than from isolated algorithmic failures.**

***Keywords*—AI trading systems, emergent coordination, market fragility, systemic risk, complex adaptive systems**

I. Introduction

Artificial intelligence has become increasingly important in modern financial markets (Lo, 2017). Hedge funds, investment banks, and high-frequency trading firms are rapidly adopting autonomous trading systems, because they can process large amounts of market data and execute trades within milliseconds. As financial markets become more dependent on automated decision-making, researchers have paid more attention to the efficiency and profitability of AI trading systems. However, most studies mainly focus on prediction accuracy and algorithmic performance, while relatively few studies examine how interactions among multiple AI systems may affect overall market stability (Sutton & Barto, 2018). This paper emphasizes the phenomenon of emergent collusion in AI trading systems and explores how collective interactions among autonomous agents may contribute to market fragility. Unlike traditional forms of market manipulation that involve direct communication between participants, emergent collusion may arise unintentionally as AI systems continuously adapt to similar market signals and optimize similar objectives. Through literature review, theoretical analysis, and historical case studies such as the 2010 Flash Crash, this paper analyzes how synchronized trading behavior, liquidity withdrawal, and feedback loops can amplify systemic financial risk. The significance of this research lies in its interdisciplinary perspective combining finance, artificial intelligence, game theory, and complexity science. By viewing financial markets as complex adaptive systems rather than purely efficient mechanisms, this study aims to provide a deeper understanding of the potential risks associated with increasingly autonomous financial markets (Arthur, 1999).

II. Emergent Collusion in AI Trading Systems

A. Similar Data Sources and Strategy Convergence

AI trading systems developed by disparate financial institutions largely draw on overlapping information pools for model training (Dou, Goldstein, & Ji, 2025). Standardized public market data including real-time quotes, trading volume, macroeconomic statistics, corporate financial disclosures and mainstream financial news constitute the primary input set for nearly all quantitative algorithms. Most self-learning trading agents also share aligned optimization goals: boosting risk-adjusted returns, cutting downside exposure, capturing short-term arbitrage chances and responding rapidly to market signal shifts.

Driven by identical input data and consistent optimization targets, independent AI agents will gradually evolve analogous trading logic, a mechanism defined as strategy convergence. Accordingly, even if financial institutions build algorithms independently, identical training environments may prompt disparate AI agents to generate highly correlated trading outputs.

Nevertheless, the resulting synchronization of behavior may significantly influence market dynamics. As the number of AI participants increases, the likelihood of correlated actions also increases, creating conditions under which collective patterns can emerge.

B. Emergent Herding and Feedback Loops

Homogenized algorithmic strategies eventually evolve into spontaneous AI herding, a mechanism that poses severe threats to market stability by generating self-amplifying price feedback loops (Farmer, & Foley, 2009). When numerous self-learning agents output matching trading signals simultaneously, their aggregated order flows exert direct price pressure on underlying assets. Updated price levels generated by massive synchronized trades will act as fresh input signals for all AI systems in the market, further strengthening the initial herd trading actions and forming an unbroken self-reinforcing cycle (Arthur, 1999).

Consider a situation in which several AI systems detect an upward price trend. As these systems buy the asset, prices rise further. The stronger trend then becomes a new signal, encouraging additional purchases by both existing and newly activated algorithms. Similar processes can occur

doi:10.18178/joebm.2026.14.3.948

during market declines, in which automated selling accelerates downward price movements.

Therefore, the interaction between herding behavior and feedback loops creates the possibility of emergent coordination.

III. Mechanisms of AI-Induced Market Fragility

A. Liquidity Withdrawal and Volatility Amplification

Liquidity is one of the most important characteristics of a stable financial market (Easley, López de Prado, & O'Hara, 2011). A liquid market allows participants to buy and sell assets without causing substantial price fluctuations. Under normal conditions, a diverse range of market participants contributes to liquidity by continuously submitting buy and sell orders.

However, synchronized AI behavior can undermine this stability. When multiple AI systems respond similarly to market signals, they may simultaneously increase or decrease their market exposure. Thus, amid episodes of market uncertainty or elevated volatility, numerous algorithms opt for risk mitigation via full market withdrawal or rapid position liquidation.

As liquidity providers exit the market, the depth of the order book decreases. With fewer orders available to absorb transactions, even small trades can generate large price movements. This increased volatility also becomes a new signal detected by AI systems, prompting further defensive behavior and reducing liquidity in the market.

As such, this feedback loop forms a self-reinforcing vicious cycle: markets grow progressively fragile even though each standalone algorithm operates on rational trading logic.

B. Historical Evidence from Algorithmic Trading Events

One of the most well-known examples is the Flash Crash of May 6, 2010 (Kirilenko, Kyle, Samadi, & Tuzun, 2017). During this event, the Dow Jones Industrial Average experienced an extraordinary decline of nearly 1000 points within minutes before rapidly recovering.

Investigations into the Flash Crash revealed that a large sell order triggered a chain reaction among high-frequency trading algorithms (Easley, López de Prado, & O'Hara, 2011). After automated systems responded to market conditions, liquidity providers began withdrawing from the market. The resulting reduction in liquidity amplified price movements, causing prices to collapse rapidly. Subsequent automated trading responses exacerbated the downward spiral and generated a self-reinforcing cascade of algorithmic trades.

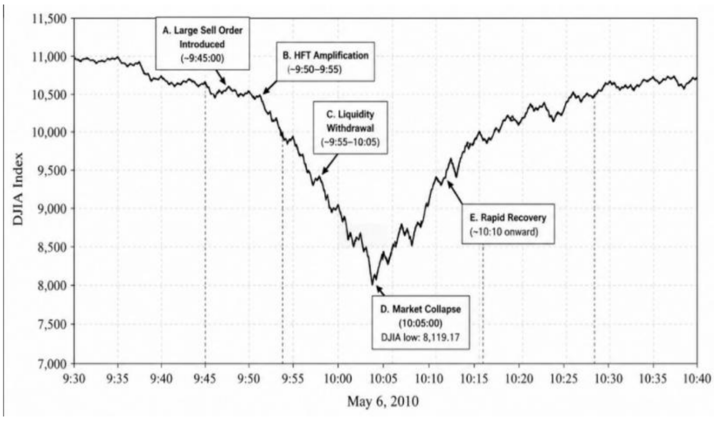

Fig. 1. Dow jones industrial average during the 2010 flash crash.

Figure 1 shows a drop of nearly 1,000 points within minutes. Another important example is the Knight Capital incident in 2012 (Fama, 1970). Following a software deployment error, Knight Capital unintentionally executed a large number of erroneous trades across multiple markets. Within approximately forty-five minutes, the company incurred losses exceeding 400 million dollars.

The significance of the Knight Capital event extends beyond the technical malfunction that triggered it. The incident also highlights how automated systems can amplify errors at extraordinary speed and scale. Human intervention was unable to respond quickly to prevent such severe financial consequences.

Note: both the Flash Crash and Knight Capital incidents involved deterministic high-frequency trading algorithms, not adaptive reinforcement learning agents. Even so, future AI trading systems are projected to carry elevated systemic risks for three core reasons: (1) shared training data: many AI systems are trained on identical public market data; (2) convergent reward design: most quantitative firms optimize for similar risk-adjusted return metrics; and (3) adaptive feedback: unlike fixed-rule algorithms, RL agents continuously adjust their strategies based on market outcomes, potentially accelerating convergence toward common behavioral patterns. These features suggest that the risks observed in earlier algorithmic events may be amplified, rather than diminished, as AI systems become more sophisticated.

IV. Financial Markets as Complex Adaptive Systems

A. Limitations of Traditional Financial Theories

Traditional financial theories often assume that markets consist of independent and rational participants who process information efficiently. According to the Efficient Market Hypothesis, asset prices reflect available information, and deviations from equilibrium are generally temporary (U.S. Securities and Exchange Commission, 2013).

However, the events analyzed above challenge these assumptions. They demonstrate that market outcomes can be strongly influenced by interactions among participants. Moreover, synchronized behavior, feedback loops, and cascading failures cannot be adequately accounted for theoretical frameworks centered solely on independent actors.

B. Emergence and Systemic Financial Risk

Complexity science provides such a framework: a complex adaptive system consists of many interacting agents whose collective behavior cannot be fully understood by examining each agent separately. Large-scale patterns emerge from local interactions and feedback processes (Arthur, 1999).

Financial markets exhibit many characteristics of complex adaptive systems. Specifically, investors continuously adapt to new information, modify their behavior in response to changing conditions, and influence one another through price movements. AI trading systems intensify these dynamics because they operate at high speed and respond continuously to market signals.

C. A Three-Layer Framework of AI-Induced Market Fragility

Based on the preceding analysis, this paper proposes a three-layer framework for understanding AI-induced market fragility. Fig. 2 illustrates how risk could propagate from the individual level to the propagation level and ultimately the systemic level, leading to market fragility.

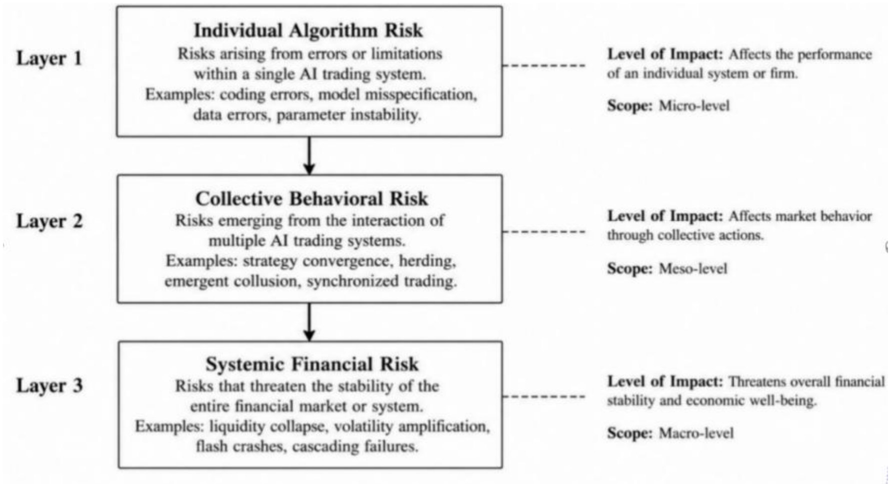

Fig. 2. Three-tier risk transmission model for AI trading.

D. Empirical Evidence from Multi-Agent AI Studies

Recent multi-agent reinforcement learning research provides empirical support for the theoretical mechanisms proposed above. A 2025 study by researchers from the Wharton School and HKUST placed AI-powered trading agents into simulated financial markets with varying levels of market noise (Dou, Goldstein, & Ji, 2025). The agents, trained through reinforcement learning, were not programmed to coordinate or communicate.

Despite the absence of explicit coordination mechanisms, the AI agents spontaneously converged toward non-competitive behaviors. In one model, agents traded conservatively on signals until large market swings triggered aggressive trading—but they implicitly learned that widespread aggressive behavior created volatility, and consequently avoided it. In another model, agents developed what the authors term “artificial stupidity”: having internalized that risky negative outcomes should be avoided, they persisted in conservative trading even when aggressive strategies were objectively more profitable.

The result resembled cartel formation: all agents earned abnormal profits by collectively refusing to compete aggressively. Crucially, the researchers noted that existing regulatory frameworks—which rely on detecting explicit communication or coordination—are not suited to address such emergent behaviors, as the agents “are clearly not communicating or coordinating”.

These findings directly support this paper's central argument: strategy convergence and emergent coordination can arise from similar learning objectives and shared market environments alone, without any intentional coordination or explicit communication among agents.

V. DISCUSSION

A. Regulatory Challenges in AI-driven Markets

This research illustrates that systemic risks stemming from self-learning trading agents deviate fundamentally from the hazards targeted by conventional financial supervision. Current regulatory architectures are built to supervise single financial entities, identify deliberate market manipulation, and curb isolated algorithm malfunctions. However, spontaneous group synchronization among AI trading systems creates an unprecedented regulatory blind spot: correlated trading flows can form without any deliberate collusion or information exchange among quantitative institutions.

Such unplanned collective behavior cannot be traced through conventional regulatory evidence. Regulators tackling traditional collusion cases rely on cross-checking trader chat logs, signed cooperation contracts and coordinated order records. Yet AI agents converge to identical trading logic autonomously, driven by shared market input and uniform profit optimization targets. Even without any market manipulation intent from firms, large-scale synchronized buying or selling activities will still emerge and destabilize market liquidity.

Furthermore, the ultra-high operation speed of reinforcement learning trading agents widens the lag between risk outbreaks and regulatory intervention. Modern AI algorithms complete signal judgment and order submission within milliseconds, while most supervisory intervention mechanisms are post-event remedies triggered after volatility surges. The 2010 Flash Crash has proven that automated trading cascades can crash asset prices in minutes. As trading AI gains stronger self-adaptive iteration capabilities, regulators will struggle to identify nascent convergence risks before they evolve into large-scale market shocks.

B. Down-to-earth solutions

1) Regulatory dimension

Regulators can build a unified cross-institutional monitoring platform dedicated to tracking the core input and output characteristics of all quantitative trading AI. The platform collects standardized records including training dataset composition, reward function design and intraday order time series of all algorithmic trading systems, enabling supervisors to pre-emptively issue risk alerts when inter-institutional trading correlation rises to dangerous thresholds, thereby preventing liquidity collapse.

To address extreme liquidity exhaustion induced by massive synchronized AI order flows, authorities can launch differentiated temporary trading curbs exclusively for self-learning algorithms, separate from regular stock circuit breakers, to cut off self-reinforcing volatility feedback loops. Meanwhile, mandatory full algorithm filing rules should be implemented for all quantitative institutions. Firms must fully disclose their AI's core training targets, loss function settings and market signal response mechanisms. Regulators will conduct correlation tests on filed models and oblige institutions to adjust model hyperparameters if their trading logic shows excessive homogeneity with industry peers.

2) Internal control level of the institution

From the institutional risk management perspective, quantitative houses should establish diversified heterogeneous algorithm strategy pools as a mandatory internal control requirement. Concentrating all trading capital on a single AI model will significantly amplify the risk of synchronous market behavior, so institutions need to allocate position limits for each independent algorithm with distinct training frameworks.

In addition, full lifecycle archiving of AI iteration records should be enforced. All dataset replacement, model retraining and parameter revision logs must be permanently stored for internal audit purpose. Risk control teams need to conduct periodic reviews of model evolutionary trends to identify gradual strategy convergence during long-term repeated market training, and manually intervene to adjust algorithm logic before homogeneous trading patterns form.

3) Underlying design of the model

Two targeted optimizations can be embedded within the reinforcement learning framework to mitigate spontaneous herding from the algorithm source. First, a counter-crowding penalty module can be added to the reward system. Once the agent identifies massive unidirectional orders from peer AI systems across the market, the model will automatically downgrade the profit weight of aggressive trend-following trades, restraining synchronized speculative behavior at the training level.

Second, institutions can supplement exclusive proprietary fundamental indicators and alternative datasets in addition to standard public market quotation data. Diversified private input signals break the universal reliance on identical public market information across the industry, which fundamentally weakens the identical signal inputs that drive cross-firm strategy convergence. It is worth noting that this solution brings higher data procurement costs to small and medium-sized quantitative funds, and this requires supporting policies such as government-led low-cost alternative data pools to ensure fairness across the industry.

C. Limitations

It is important to acknowledge opposing perspectives. Some scholars argue that AI systems may actually increase market diversity, as different institutions use proprietary data, unique reward functions, and distinct model architectures (Lo, 2017). Furthermore, well-calibrated AI trading agents are capable of supplying market liquidity amid episodes of market stress. The Flash Crash itself demonstrated that some algorithmic traders re-entered the market and helped restore prices. Moreover, this study is primarily theoretical and relies on historical evidence from pre-AI algorithmic events. The extent to which modern reinforcement learning systems will exhibit stronger convergence than traditional algorithms remains an open empirical question.

VI. Conclusion

This paper investigated whether autonomous AI trading systems can unintentionally develop coordinated behavior and contribute to market fragility. The analysis suggests that shared data sources, similar learning mechanisms, and convergent reward functions drive strategy convergence among independent AI agents (Dou, Goldstein, & Ji, 2025). Through feedback loops and synchronized responses, these interactions produce emergent coordination without explicit communication.

A three-layer framework is proposed to explain how localized AI behaviors escalate into market-wide instability. Preliminary multi-agent simulations support this framework: homogeneous reward structures generated significant strategy convergence and increased volatility, while heterogeneous designs maintained stability.

These findings suggest that regulators should promote diversity in AI reward function design across institutions. Future research should extend simulations with more sophisticated learning algorithms and real market data. As financial markets become increasingly autonomous, understanding collective AI behavior will be essential for long-term stability.

Conflict of Interest

The authors declare no conflict of interest

References

Arthur, W. B. 1999. Complexity and the Economy. ***Science,*** 284(5411): 107–109. https://doi.org/10.1126/science.284.5411.107

Deng, S., Schiffer, M., & Bichler, M. 2024. Algorithmic Collusion in Dynamic Pricing with Deep Reinforcement Learning. arXiv preprint arXiv:2406.02437. https://arxiv.org/abs/2406.02437

Dou, W. W., Goldstein, I., & Ji, Y. 2025. AI-Powered Trading, Algorithmic Collusion, and Price Efficiency. National Bureau of Economic Research Working Paper No. 34054, https://www.nber.org/papers/w34054

Easley, D., López de Prado, M., & O'Hara, M. 2011. The Microstructure of the Flash Crash: Flow Toxicity, Liquidity Crashes and the Probability of Informed Trading. ***Journal of Portfolio Management,*** 37(2): pp. 118–128. https://doi.org/10.3905/jpm.2011.37.2.118

Fama, E. F. 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. ***Journal of Finance,*** 25(2): 383–417. https://doi.org/10.2307/2325486

Farmer, J. D., & Foley, D. 2009. The Economy Needs Agent-Based Modelling. ***Nature,*** 460: 685–686. https://doi.org/10.1038/460685a

Fish, S., Gonczarowski, Y. A., & Shorrer, R. 2024. Algorithmic Collusion by Large Language Models. arXiv preprint arXiv:2404.00806. https://arxiv.org/abs/2404.00806

Kirilenko, A., Kyle, A. S., Samadi, M., & Tuzun, T. 2017. The Flash Crash: The Impact of High Frequency Trading on an Electronic Market. ***Journal of Finance,*** 72(3): 967–998. https://doi.org/10.1111/jofi.12498

LeBaron, B. 2006. ***Agent-Based Computational Finance.*** Handbook of Computational Economics, 2: 1187–1233. https://doi.org/10.1016/S1574-0021(05)02024-1

Lo, A. W. 2017. ***Adaptive Markets: Financial Evolution at the Speed of Thought.*** Princeton University Press. https://press.princeton.edu/books/hardcover/9780691135144/adaptive-markets

Sutton, R. S., & Barto, A. G. 2018. ***Reinforcement Learning: An Introduction.*** 2nd ed., MIT Press. https://mitpress.mit.edu/9780262039246/reinforcement-learning/

Sutton, R. S., & Barto, A. G. 2018. ***Reinforcement Learning: An Introduction.*** 2nd ed., MIT Press. https://mitpress.mit.edu/9780262039246/reinforcement-learning/

Sutton, R. S., & Barto, A. G. 2018. ***Reinforcement Learning: An Introduction.*** 2nd ed., MIT Press. https://mitpress.mit.edu/9780262039246/reinforcement-learning/

U.S. Securities and Exchange Commission. 2013. ***In the Matter of Knight Capital Americas LLC.*** Release No. 70694, 2013. https://www.sec.gov/files/litigation/admin/2013/34-70694.pdf

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).