Proceedings of ICFTBA 2025 Symposium: Global Trends in Green Financial Innovation and Technology
DOI: 10.54254/2754-1169/2025.BJ28129

Research on Herding in AI Trading

Shuo Han

Orange Lutheran High School, Orange, USA
hanshuo20070907@163.com

Abstract. Financial markets have undergone continuous transformation, with artificial intelligence (AI) marking a pivotal stage in their evolution. While AI-driven trading enhances efficiency by processing vast data at high speed, it also introduces systemic vulnerabilities through herding behavior. Unlike traditional human herding driven by psychology, algorithmic herding stems from model convergence, information amplification, recursive feedback loops, cross-market spillovers, and digital sentiment cascades. Historical episodes—including the 2010 Flash Crash, the 2015 Chinese stock market collapse, the 2021 GameStop surge, and the 2022 cryptocurrency crash—illustrate how minor shocks can escalate into large-scale market disruptions. These cases reveal that AI not only codifies new forms of irrationality but also accelerates contagion across markets. The implications extend beyond volatility, threatening liquidity, investor trust, and systemic stability. To mitigate risks, strategies such as technological diversification, anti-herding algorithm design, circuit breakers, regulatory technology, and international coordination are proposed. Ultimately, herding in AI-driven markets cannot be eliminated but must be managed through intentional design and governance. This paper contributes to the interdisciplinary understanding of AI herding by integrating insights from behavioral finance, complexity science, and regulatory studies, aiming to balance efficiency with stability in an era of machine-dominated finance.

Keywords: Artificial Intelligence Trading, Algorithmic Herding, Financial Stability.

1. Introduction

Since ancient times, the financial markets have never remained inactive. They have evolved from the informal markets within city centers and coffee houses to some of the most sophisticated electronic spaces in which billions of assets can change hands within a small part of a second. Each technological development from the telegraph to the tape from the telephone through the electronic screen altered the velocity and nature of market activity. The adoption of the use of artificial intelligence is the latest and perhaps the future most important step in this development.

Artificial intelligence-based traders don't have just one use. At the most basic level, it's a program of machine learning that identifies patterns in historical data. At the highest levels of complexity, neural networks may be unpacking raw text, reinforcement learning players optimizing strategy in simulated markets, and deep learning architectures predicting changes in volatility. Regardless of the application, the classic characteristic of AI trading is that the system is capable of processing staggering amounts of information with scorching velocity and then responding to that information in real time. These qualities promise efficiency but also accompany subtle dangers.

The spread of algorithmic trading over the past two decades was unprecedented. The handful of hedge funds and investment banks deploying automated strategies in the initial decades of the 2000s mushroomed exponentially in the ensuing decades. High-speed trading emerged as the queen of U.S. equities in the decade of the 2010s, commanding over half the daily volume. Algorithmic infrastructures in Europe and in Asia also became the top, reshaping the dynamics of markets. In cryptocurrencies, with no long-established institutional framework, AI-based bots saw immediate popularity as the center of exchanges. The combination of global connectivity, vast computational capabilities, and available data sets has democratized access so that gargantuan asset managers and small houses of trade employ algorithmic decision-making.

On the surface, that revolution has gone well. Liquidity is wider, bid-ask spreads are narrower, and prices adjust more quickly to information than in the past. Investors can buy and sell at any time, anywhere, with the least possible friction. Such outward benefits, though, hide vulnerabilities that become critical under pressure. What if there are thousands of algorithms that have all learned from the same data sets and written with the same objectives and behave similarly? What if the same speed that is promised to engender efficiency only fuels instability?

The answer is the concept of herding. Herding is the market participant's behavior of copying one another moves rather than exercising independent thinking. At the human level, this is responsible for explaining what causes crowds to stampede into hot markets or stampede in unison out of decaying assets. What occurs in the digital age is that the concept of herding reaches new levels. If various algorithms are all pointing in the same directions, then the movements verify each other, producing market movements far more vicious than the grounds would justify. Minor turbulence can mushroom into full-blown stampedes.

The past offers refreshing reminders of this possibility. The Flash Crash of May 2010 demonstrated how high-frequency traders who at the beginning were considered liquidity providers themselves as a group could become liquidity takers when volatility spiked. The Dow Jones Industrial Average fell in minutes near one thousand points, voiding and then recreating hundreds of billions in market value. Chinese stocks saw in 2015 a spectacular collapse as margin-motivated speculation combined with algorithmic selling obliterated trillions in capitalization. The recent tale of the GameStop saga in 2021 and the crash of large cryptocurrencies in 2022 demonstrated how digital coordination among retail investors and machine trading would be amplified in unpredictable manners.

These episodes highlight a paradox. Artificial intelligence was supposed to decrease irrationality through the elimination of human emotion like fear and greed. However, in practice, the AI programs end up codifying novel forms of collective irrationality. Rather than emotional contagion, there is algorithmic contagion. Rather than fear being spread via rumor, there is spread via code. The end result is an environment in which the efficiency of machines is accompanied by the brittleness of crowds.

The research in AI herding is more than intellectual curiosity. It has direct policy relevance in financial stability, investor protection, and regulation. Central banks and market regulators are wary that algorithmic stampedes will lead systemically important crises like traditional nineteenth- or twentieth-century runs on the banks. Institutional investors are confronted with the challenge of how the use of similar models in their portfolios increases portfolio risk. Retail traders are no exception as their trades blend with algorithms that dominate the setting of prices. The understanding of this herding is central in the design of the protection that maintains the automation advantage while mitigating the automation risk.

The paper seeks to investigate the phenomenon in depth. The paper begins with the tremendous pool of behavioral finance's herding scholarship, in which researchers have long asked why individuals mimic one another under ambiguity. The paper then cites the specific mechanics by which the following is intensified by AI, namely algorithmic convergence, information amplification, and feedback loops. Case applications—in equities, emerging markets, meme stocks, and cryptocurrencies—clarify how the theory translates. The paper concludes with an array of possible mitigation approaches, from technological diversification to regulation innovations, and concludes with future directions in AI in international finance.

Through exploring these questions, the paper seeks both intellectual insight and practical debate. Policymakers have long-term decisions to make over how much they regulate markets shaped by machines. Investors are presented with the dilemma of whether they should embrace or resist strategies that offer reward but are liable to be volatile. Society as a whole is confronted with the issue of whether artificial intelligence, if unchecked, would be a source of system-wide financial risk. Much is at stake. As the historical record informs us, markets create wealth when they function well but causes of crisis if they get out of hand. The challenge in the era of AI is how to balance efficiency with stability.

2. Literature review

The finance literature on herding is founded upon multiple disciplines: economics, psychology, sociology, and most recently, computer science. Herding is at its core decision-making under ambiguity. Investors, when in doubt as to what is the correct decision, will observe what others are doing. What is rich about the topic is that the imitation may be rational or unjustified, stabilizing or destabilizing, depending upon the context. The review will trace the development of the theory of herding, examine the empirical evidence across markets, and examine how the new resurgence of AI in finance redefined the debate.

2.1. Foundations in behavioral finance

The modern herd behavior research begins from Banerjee, who presented the least complicated model with sequentially interacting agents who observe predecessors' choices [1]. Assuming people have private information doesn't rule out the public action weight being so large that they will be uninformed of their own signal and will be inclined to imitate the majority. The result is that markets will move in some direction due to both the pioneers forming the standard and because fundamentals prescribe so.

Bikhchandani, Hirshleifer, and Welch generalized this argument in their informational cascade model [2]. They demonstrated that after the establishment of the cascade, the latter is always self-reinforcing: future players rationally disregard private information in favor of believing the crowd. Such cascades can proceed long after the moment when fundamental changes take place, revealing bubbles and crushes.

Psychologists have long reported the same kinds of phenomena in non-financial contexts. Solomon Asch's conformity experiments in the 1950s established that people tend to take the clearly wrong answer if others in the group take that answer first. In markets, the cost of being wrong by oneself often exceeds the cost of being wrong as part of the herd. The insight links finance with broader patterns in human conformism.

Devenow and Welch gave an important refinement in that they argued that there could be herding in a career context [3]. The fund manager who is different from consensus is likely to be criticized if he performs poorly, but the one who conforms to consensus shares blame with peers. Such reputational incentive favors mimicry even when analysis recommends something else. The reason is not being correct but not being the odd man out.

2.2. Evidence from financial markets

Empirical evidence confirms that there is no abstraction called herding. Lakonishok, Shleifer, and Vishny surveyed U.S. pension funds and found systematic instances of managers buying or selling in the same direction [4]. Their discovery was that institutional herding far more often exacerbates mispricing rather than corrects mispricing.

Chang, Cheng, and Khorana compared US and Asian markets, under which they concluded that the force of the herd was stronger in the emerging markets [5]. The reason was partly the weaker information structure and partly cultural differences that prefer uniformity. Asian investors depended more on public information and followed the others rather than putting their trust in their own analysis.

Sias provided more evidence from the behavioral finance perspective through the examination of institutional investors' trading patterns [6]. He demonstrated that while herding is observed during times of panic, it is also noticed during typical portfolio re-balancing. Simultaneous position readjustments by funds, even for innocuous motives, engender correlated trade waves that induce price movements.

Other studies have expanded the evidence. Foreign exchange market studies have demonstrated the existence of herding among foreign exchange traders in volatile times. The bond markets, similarly, display trade clustering as the investors respond to macroeconomic releases. The phenomenon transcends the asset classes, geography, as well as the time frames.

2.3. Critiques and counterarguments

Not all scholars agree on the dangers of herding. Some argue that imitation can be stabilizing, not destabilizing. If early movers have superior information, then following them can help disseminate knowledge through the market. In this view, herding is a mechanism of information aggregation rather than distortion.

Others suggest that what appears as herding may simply reflect common responses to fundamentals. If all investors sell after an interest rate hike, this may not be imitation but parallel rational reaction. Distinguishing true herding from correlated fundamentals is a challenge for empirical research.

Moreover, some economists argue that AI might reduce herding rather than increase it. By removing human emotion, algorithms could act more independently and be less swayed by social cues. Machines, in theory, should not care about reputational risk or peer conformity. From this perspective, AI could make markets more rational.

Yet critics of this optimistic view point out that algorithms are designed by humans and trained on historical data. If human biases are encoded into datasets, they will persist in machine behavior. Furthermore, the commercial pressure to adopt profitable strategies leads firms to design similar models, undermining independence. In practice, AI has not eliminated herding but altered its form.

2.4. AI and the transformation of herding

The rise of algorithmic trading since the 2000s has brought herding into new focus. Kirilenko et al. examined the 2010 Flash Crash and concluded that high-frequency traders, initially providing liquidity, quickly became net sellers as volatility rose [7]. Their algorithms, reacting to similar signals, amplified the downturn.

Cont and Bouchaud framed financial markets as complex adaptive systems, where interactions among agents produce emergent phenomena [8]. In such systems, small shocks can trigger large cascades. AI accelerates this dynamic because it compresses reaction times. A piece of news that might once have taken hours to spread can now move markets in seconds.

Easley and O'Hara emphasized that liquidity provision is crucial [9]. If machines react to new information faster than liquidity providers can adjust, prices may move violently before stabilizing. AI herding thus interacts with structural features of markets, not just psychology.

More recent work has studied cryptocurrencies, where algorithmic bots dominate trading. Researchers find that herding in these markets is particularly severe due to high leverage, thin liquidity, and absence of circuit breakers. Automated liquidation cascades on exchanges illustrate herding in its purest form.

2.5. Interdisciplinary perspectives

Herding in AI trading is also examined from perspectives beyond finance. Complexity science views markets as networks of interacting nodes, where local decisions create global patterns. The concept of criticality—systems poised on the edge of sudden change—maps neatly onto algorithmic markets.

Computer scientists study multi-agent systems, where autonomous algorithms interact in shared environments. Insights from this field show how coordination problems and feedback loops emerge naturally even without explicit imitation. In trading, the shared environment is the order book, and the feedback is price movement.

Sociologists highlight the role of trust and legitimacy. Investors adopt certain technologies not only for performance but also because peers do. The diffusion of algorithmic trading itself followed a herding pattern, as institutions feared being left behind if they did not adopt similar systems.

2.6. Research gaps

While there is vast scholarship, much is yet absent. Few papers study the two-way movement of human and mechanical herding that coexist in practice. Little is produced on the subject of the lesser-known forms of the asset classes such as commodities and high-quality corporate debt. Policy responses, while highly theorized, are rarely thoroughly tested. Finally, the ethical dimension of AI herding—is there fairness, responsibility, and resilience in the system?—is just in its infancy in terms of being researched.

3. Mechanisms of herding in AI trading

Understanding how herding arises in AI trading requires attention to the technical structures of algorithms, the informational environment of modern markets, and the incentives that shape institutional behavior. Unlike human herding, which stems from psychological pressures, algorithmic herding emerges from design choices, optimization goals, and the architecture of trading systems. The mechanisms are multiple, overlapping, and reinforcing.

3.1. Algorithmic convergence and model homogeneity

The most direct and immediate of the three is algorithmic convergence. The majority of institutional trading models are built from similar sets of data and optimized against similar metrics. The majority are based on historical prices, measures of volatility, order book depth, and measures of sentiment. The majority use publicly available frameworks and toolkits such as TensorFlow and PyTorch, which encourage the use of similar architectures.

If objectives of optimisation are aligned—maximizing current profits while decreasing variances—the many models all churn out in the end the same result given the same input. A breakthrough through resistance in prices, for example, is coded as buy signal in many systems. As the programs cascade in at the same moment, prices take off. When, however, support is broken, the programs all run frantically to sell, accentuating losses.

Such convergence is no accident. Competitive pressures force firms to try strategies others have determined profitable. A hedge fund that shuns volatility-based signals will lag behind. Fear of lagging causes uniformity. Algorithmic convergence effectively engrains herding in code: machines are instructed to think similarly.

3.2. Information transmission and signal amplification

A second system is the intensification of information signals. News-sensitive markets always existed, but the AI reduces response times profoundly. Natural language processing programs read headlines, earnings announcements, filing documents with the regulator, and even tweets and translate unstructured text into numerical scores of sentiment. The trading models respond based on such scores in mere milliseconds, well ahead of the time that human analysts can assess context.

Amplification occurs when many systems share the same threshold. Pretend dozens of funds use sentiment models that will trigger sales when negativity crosses some predetermined threshold. Some negative article or governmental comment drives the sentiment over that threshold, and thousands of machines sell in unison. The prices fall decisively, which in and of itself is fed into momentum-based models that interpret the decline as confirmation. The cycle continues.

Stop-loss orders worsen the effect. If algorithms are putting protective stops at congested levels of prices, some small decline invokes one cluster of sales, then another, and another, in what amounts to a cascade. What would have remained within the realm of containable correction is now a precipitous decline. The speed and force of AI make amplification far more perilous than in purely human markets.

3.3. Feedback loops and recursive dynamics

The most destabilizing of all possible mechanisms is the feedback loop. Feedback loops in human markets were there but evolved over the long haul. Under AI, they play out virtually immediately. Machines are trading and they move prices; prices move and the machines take the changes as new information. Selling begets more selling, buying begets more buying.

These feedback loops are self-reinforcing because the systems being automated themselves constantly adjust their parameters. Risk management modules, for example, will delever when there is high volatility. If many systems see this rule, an increase in volatility triggers wholesale deleveraging, which increases the level of volatility some further, beckoning even more deleveraging. The system was at work during the 2010 Flash Crash, when in two seconds liquidity dried up as programs delevered universally.

Feedback loops may also extend across asset classes. Equity losses may induce risk-off trades in bonds or currencies. The presence of derivatives creates added layers of complication: options hedging may commit dealers to buy or sell underlying assets in ways that intensify movements. Feedback loops, when they get started, may continue until some outside intervention—in the form of circuit breakers or central announcements—brings order.

3.4. Cross-market arbitrage and herding spillovers

A fourth mechanism is the role of cross-market arbitrage. Many AI systems operate across multiple asset classes simultaneously, seeking correlations and mispricings. While arbitrage in theory improves efficiency, it also transmits shocks. A move in one market can trigger algorithmic responses in another, spreading volatility.

For example, if equity prices fall, algorithms may short related futures contracts or shift capital into bonds. Currency markets respond as funds move internationally. The result is herding spillovers: algorithms that were not directly exposed to the initial shock nonetheless join the movement because of programmed correlations.

During the European debt crisis, for instance, declines in Greek bonds spilled rapidly into other sovereign markets as algorithms sold correlated assets. In the 2022 collapse of cryptocurrencies, losses in Bitcoin triggered automated selling in Ethereum and altcoins, even in cases where fundamentals differed. Cross-market herding transforms local disturbances into global ones.

3.5. News, social media, and digital sentiment cascades

The third mechanism is the digital information flow effect. Unlike the history when markets relied upon official news wires, contemporary AI systems consume vast amounts of information from social media. Twitter tweets, Reddit posts, and web fora become alerts after being processed through sentiment programs. The end result is that there is an alternative channel of herding now: virally spreading online group stories can be propagated through machines.

The 2021 GameStop saga is classic. WallStreetBets retail investors coordinated the buying of shares and call options. Algorithms that looked for volume and momentum surged ahead, seeing the new activity as the signal. What had started as an organically driven movement was soon becoming an AI-fueled mania that drove prices far ahead of fundamentals.

The system blurs the lines between mechanical and human herding. Human narratives give the spark; the algorithms provide the fuel. Once the sentiment in the online community shifts, the machines emphasize the effect, solidifying the tale with real price action. Digital cascades become market cascades.

3.6. Synthesis

Collectively, these mechanics reveal how markets become the structure in which AI converts the psychological inclination of herding into an institutional attribute. Convergence causes machines to think similarly. Amplification causes small alerts to get amplified. Feedback loops cause movements to self-reinforce. Cross-market arbitrage causes local shocks to propagate globally. Social media integration causes human narratives to get coded directly into the action of machines.

The large picture is one of brittleness. Speed-and-efficiency-optimized markets grow brittle against collective errors. The same qualities that render AI attractive—speed, consistency at large scales—are the same qualities that worsen herd behavior. The examination of such mechanisms is vital in the construction of protections that preserve efficiency but defer instability.

4. Case studies and graphical evidence

Case studies provide insight into how herding happens in practice. They disclose that algorithmic trading is no independent force but interacts with human behavior, institutional incentives, and market structures. Examination of specific episodes identifies patterns of convergence, amplification, as well as feedback loops, and how local events spread across markets.

4.1. The 2010 flash crash

The American equity markets saw one of the biggest intraday losses in history on May 6, 2010. The Dow Jones Industrial Average fell in little more than ten minutes by nearly one thousand points, wiping out hundreds of billions in market capitalization before recovering the losses largely at the same speed. The initial triggers seemed unexplainable. There was no big macroeconomic release, nor any signal of a big shock from the external world.

Additional research revealed that a single institutional sell order in E-mini S&P 500 futures, valuing around $4 billion, initiated the selling. Under normal conditions, such an order would've absorbed over the course of hours. On this day, high-speed trading algorithms multiplied the effect. Instead of being liquidity providers, they began purchasing and selling contracts back and forth from each other, creating an illusion of liquidity but receding when the volatility spiked. Once the prices began falling, other algorithms recognized the movement as the cue to sell and entered.

The episode demonstrated all the characteristics of algorithmic herding. Convergence meant numerous algorithms reacting in the same way to the prices falling. Amplification meant the initial order having far greater effects than expected. Feedback loops meant selling causing more selling. Cross-market spillovers followed, with losses in futures being equal to equity sell-offs. The flash crash demonstrated that herding could be triggered not by panicking among human investors but by code executing its logic too well.

4.2. The 2015 Chinese market crash

China's stock markets had a bumpy ride in 2015. Retail investors spurred on by inexpensive credit and margin financing have flocked into shares, sending valuations shooting into the stratosphere. The Shanghai Composite Index shot up more than 150 percent in just twelve months. The atmosphere was cheerful, and millions of new brokerage accounts were opened every month.

But in June, the bubble burst. Prices began to fall, and the sell-off accelerated as margin calls took effect. Algorithmic systems add to the turmoil. Many domestic funds employ simple momentum-driven models. As the market turned downward, the algorithms triggered waves of sell orders. Herding becomes apparent on two levels: retail investors panic and machines amplify the decline.

By July, trillions of dollars in market value had evaporated. Authorities intervened through trading suspensions, restrictions on sales and outright buying by state funds. Yet confidence has been badly shaken. The Chinese case shows how herding in A.I. trading can combine with human enthusiasm to create a particularly damaging cycle. The machines themselves did not cause the crash, but they accelerated the process and deepened the losses.

4.3. GameStop and meme stocks in 2021

The GameStop episode in January 2021 highlighted a new form of herding, where online coordination among retail investors intersected with algorithmic responses. Members of Reddit's WallStreetBets forum encouraged each other to buy shares and call options in GameStop, a struggling video game retailer. The goal was to force a short squeeze against hedge funds with large bearish positions.

The campaign succeeded spectacularly. GameStop's stock price rose from under $20 to over $400 in a matter of weeks. While retail enthusiasm was the catalyst, algorithmic trading amplified the frenzy. Quantitative funds monitoring volume and momentum joined the rally, interpreting the unusual activity as a positive signal. Option market dynamics created further feedback: as call options were purchased, dealers hedged by buying underlying shares, driving prices even higher.

The episode revealed that AI herding need not start with machines. Human narratives can ignite a movement that algorithms then magnify. Once price action validated the Reddit narrative, machines reinforced the trend, making the feedback loop more powerful. The GameStop case illustrates the hybrid nature of modern markets, where social media cascades and algorithmic cascades are intertwined.

4.4. The collapse of FTX and the 2022 crypto crash

Cryptocurrency markets provide perhaps the clearest view of algorithmic herding because of their reliance on bots and automated liquidation systems. In 2022, the collapse of the FTX exchange triggered a massive sell-off across the sector. Once doubts about FTX's solvency spread, investors rushed to withdraw funds. Prices of Bitcoin, Ethereum, and other tokens began to slide.

Automated systems magnified the panic. Many traders had taken leveraged positions, using their holdings as collateral for further borrowing. As prices fell, margin calls triggered automated liquidations. Algorithms across multiple exchanges sold assets simultaneously, driving prices down further. Liquidity thinned as market makers pulled back. Within days, billions of dollars in value evaporated.

The crypto crash illustrated every herding mechanism. Convergence was visible in the similarity of trading bots' strategies. Amplification came from liquidation cascades. Feedback loops developed as falling prices triggered more sales. Cross-market spillovers spread from Bitcoin to altcoins and related derivatives. Social media added fuel, as rumors of insolvency spread rapidly on Twitter, pushing sentiment lower.

4.5. Graphical evidence and common patterns

Graphical data from these episodes reveal striking patterns. Spikes in trading volume coincide with sharp price movements, suggesting collective action rather than independent decision-making. V-shaped intraday collapses, such as in the Flash Crash, show how liquidity can vanish and reappear once herding subsides. Extended declines, such as in China's 2015 crash, illustrate how herding sustains downward momentum over weeks.

Despite differences in geography and asset class, the episodes share common features. They begin with a shock—sometimes a single order, sometimes a rumor or policy change. Algorithms react simultaneously, overwhelming human capacity to intervene. Liquidity vanishes, volatility explodes, and losses compound. Only when external forces intervene—be it trading halts, government action, or market exhaustion—do prices stabilize.

4.6. Lessons from the cases

The case studies offer several lessons. First, AI herding is not hypothetical but observable. Second, machines rarely act alone; they interact with human behavior, amplifying emotions like fear and greed. Third, herding is not confined to equities: it spans futures, options, cryptocurrencies, and beyond. Fourth, the absence of safeguards, as in crypto markets, makes outcomes more severe.

Finally, the episodes show that herding is a structural feature of AI trading. It arises not from programming errors but from the logic of optimization and speed. Unless countermeasures are in place, similar episodes are likely to recur.

5. From herding to market stampede

The danger of herding in AI-driven markets is not confined to short-term volatility. When left unchecked, collective algorithmic behavior can evolve into full-scale market stampedes that threaten systemic stability. Stampedes differ from ordinary swings in both intensity and consequence: they unfold faster, involve more participants, and spill into more domains. Three pathways—liquidity crises, cross-market contagion, and erosion of investor trust—illustrate how herding escalates into systemic risk.

5.1. Liquidity crises

Liquidity is the lifeblood of financial markets. In normal times, high-frequency trading is often praised for enhancing liquidity by continuously posting bids and offers [10]. Yet this liquidity is fragile. It is provided under the assumption that algorithms can exit positions quickly without undue risk. When volatility spikes, this assumption collapses.

During episodes of stress, algorithms that normally supply liquidity may all withdraw simultaneously. Instead of absorbing sell orders, they cancel quotes, leaving markets thin. The Flash Crash of 2010 demonstrated this vividly. Liquidity that seemed abundant evaporated within seconds, leaving trades executed at absurd prices—blue-chip stocks briefly changing hands for a penny or for $100,000.

The risk is structural. Because many algorithms share risk management triggers, once volatility crosses thresholds, liquidity providers vanish together. The result is a vacuum where small trades cause outsized price swings. For investors, the paradox is stark: the very actors who stabilize markets in calm conditions disappear when stability is most needed.

5.2. Cross-market contagion

The second pathway is contagion across markets. Modern trading systems rarely operate in isolation. Funds hedge equity exposure with futures, balance bond risk with currency positions, or deploy strategies that scan multiple asset classes simultaneously. AI algorithms, designed to exploit correlations, link markets ever more tightly.

When herding begins in one domain, algorithms spread the stress to others. A sell-off in equities prompts sales of correlated futures and options. Bond yields may rise as funds liquidate safe assets to cover losses. Currencies swing as capital shifts across borders. In globalized markets, contagion is swift.

The 1997 Asian financial crisis and the 2008 global crisis demonstrated contagion in human-driven markets. AI accelerates the mechanism. When machines act simultaneously across asset classes, shocks propagate in milliseconds. What begins as a local disturbance becomes a global tremor. Crypto markets in 2022 illustrated this: the collapse of a single token, Terra-Luna, cascaded into losses across Bitcoin, Ethereum, stablecoins, and derivatives, fueled by algorithmic arbitrage and liquidation bots.

5.3. Erosion of investor trust

Perhaps the most insidious risk is the erosion of trust. Financial markets function only if participants believe prices reflect underlying value and that trading is reasonably fair. When herding causes extreme dislocations, confidence suffers.

Retail investors, in particular, may feel markets are “rigged” if algorithms move prices in ways disconnected from fundamentals. Sudden crashes undermine the idea that rational analysis can guide investment. Institutions, too, may hesitate to commit capital if they fear that liquidity will vanish at critical moments. This reluctance can reduce long-term investment and increase reliance on government backstops.

Trust is also damaged when technical errors or feedback loops create bizarre outcomes, such as blue-chip shares trading for pennies. Each episode reinforces the perception that markets are fragile and unpredictable. Over time, this perception can alter behavior, reducing participation and liquidity in a self-reinforcing cycle.

5.4. Systemic implications

Together, liquidity crises, contagion, and loss of trust create systemic vulnerabilities. Markets are no longer just mechanisms for price discovery; they are infrastructures on which economies depend. Pension funds, insurance companies, corporations, and governments rely on stable markets to allocate capital. If herding undermines this stability, the effects extend beyond trading floors to the real economy.

Central banks recognize the danger. Reports from the U.S. Federal Reserve, the European Central Bank, and the Bank of England have all flagged algorithmic herding as a potential source of systemic risk. Yet the challenge is complex. Unlike traditional bank runs, which can be slowed by closing branches, algorithmic runs unfold in milliseconds. Containing them requires new tools and coordinated oversight.

6. Mitigation and policy responses

If herding in AI-driven markets is inevitable to some degree, the key challenge is how to mitigate its destabilizing effects without eroding the efficiency benefits that automation provides. Effective responses must operate at three levels: the technological design of algorithms, the institutional architecture of markets, and the regulatory frameworks that govern global finance.

6.1. Technological diversification and algorithmic design

At the level of technology, diversification is the most obvious remedy. If all models are trained on similar datasets and optimized for similar objectives, convergence is unavoidable. Encouraging firms to diversify inputs—by incorporating alternative data sources such as satellite imagery, supply-chain indicators, or long-term macroeconomic variables—can reduce homogeneity. Similarly, exploring a broader range of model architectures, rather than relying on standard deep learning frameworks, can create heterogeneity in outputs.

Another approach is to embed anti-herding features directly into algorithms. These might include throttling mechanisms that automatically scale down trading activity when market volume spikes abnormally, or introducing randomized delays to break synchronization. Some researchers propose reinforcement learning systems that penalize strategies which correlate too closely with peers, nudging algorithms toward independence.

Transparency also matters. Firms often treat algorithms as black boxes, reluctant to disclose design details. Yet without some degree of transparency, systemic risks remain invisible. Shared standards for reporting algorithmic strategies—without revealing proprietary secrets—could help regulators monitor convergence.

6.2. Market-level safeguards

Exchanges themselves play a central role. Two tools have gained prominence: circuit breakers and speed bumps. Circuit breakers halt trading when prices move beyond predefined thresholds, allowing time for human reassessment. They were expanded after the Flash Crash and have since proven effective in moderating panic. Speed bumps, by contrast, introduce tiny delays in order processing, disrupting the ability of algorithms to act simultaneously. By staggering execution, they reduce synchronization without affecting overall liquidity.

Position limits and leverage caps are additional safeguards. When traders cannot accumulate excessively leveraged positions, liquidation cascades become less severe. Margin requirements that adjust dynamically with volatility can also prevent herding from escalating into stampedes.

Market infrastructure can also be redesigned for resilience. Dark pools, for example, allow large trades to be executed without signaling intentions to the broader market, reducing the chance of triggering herding. Likewise, consolidated audit trails enable regulators to track algorithmic flows in real time, providing data to intervene when necessary.

6.3. Regulatory technology and international cooperation

At the policy level, regulators face the challenge of governing markets that move faster than human oversight. Traditional tools—such as after-the-fact investigations—are insufficient. Instead, regulators must adopt their own AI systems to monitor flows in real time. RegTech solutions include machine learning models that detect abnormal clustering of trades, neural networks that identify suspicious order cancellations, and anomaly detection systems that flag rapid liquidity withdrawals.

Yet regulation cannot remain purely national. Global capital flows mean that a stampede in one jurisdiction can spill into others within seconds. The European Union’s MiFID II rules, the U.S. SEC’s market structure reforms, and Singapore’s MAS initiatives all attempt to address algorithmic trading, but coordination is limited. Without harmonization, firms can exploit gaps by routing trades through less regulated markets. International organizations such as the Financial Stability Board and IOSCO must therefore play a stronger role in setting common standards.

Cryptocurrency markets add another layer of complexity. Operating across borders with limited oversight, they exemplify the risks of fragmented regulation. The collapse of FTX in 2022 underscored how the absence of global rules leaves investors vulnerable. Coordinated frameworks that extend beyond traditional markets are essential to contain AI-driven herding in the digital asset space.

7. Conclusion

The evidence is clear: herding is not an incidental byproduct of AI trading but a structural feature of markets dominated by algorithms. Convergence, amplification, feedback loops, cross-market spillovers, and digital sentiment cascades combine to create environments where small shocks can trigger massive dislocations. Case studies from equities, emerging markets, meme stocks, and cryptocurrencies show how theory translates into practice, often with devastating consequences.

Yet the story is not one of despair. AI is also a tool that can enhance stability if designed wisely. The same data-processing power that accelerates herding can be redirected toward monitoring and mitigating it. Anti-herding algorithms, dynamic safeguards, and real-time oversight systems can counterbalance the risks. The challenge is one of intentional design and coordinated governance.

Looking forward, several directions merit attention. First, future research should integrate behavioral finance insights into AI models. Algorithms that account for crowd psychology may better anticipate and counteract cascades. Second, real-time cross-market monitoring should be developed as a standard feature of global financial infrastructure. With proper data-sharing, regulators could detect emerging stampedes before they spiral. Third, the rise of quantum computing may again reshape market dynamics. By enabling even faster processing, quantum systems could exacerbate herding—or, if carefully managed, could provide new tools for stabilization.

Ethical considerations also deserve greater weight. If AI systems amplify instability, who bears responsibility? Designers, firms, or regulators? Clearer frameworks of accountability are needed. Moreover, fairness must be addressed: markets dominated by machines risk marginalizing human investors, eroding confidence in their legitimacy. A balance must be struck between efficiency and inclusiveness.

Ultimately, the question is not whether herding can be eliminated—it cannot—but whether it can be managed. Just as past generations of regulators learned to handle bank runs, this generation must learn to handle algorithmic runs. The task is urgent. Markets are engines of prosperity when they function well but sources of crisis when they malfunction. Ensuring that AI enhances rather than undermines stability is therefore not just a technical challenge but a societal imperative.

References

[1] Banerjee, A. V. (1992). A simple model of herd behavior. Quarterly Journal of Economics, 107(3), 797–817.

[2] Bikhchandani, S., Hirshleifer, D., & Welch, I. (1998). Learning from the behavior of others: Conformity, fads, and informational cascades. Journal of Economic Perspectives, 12(3), 151–170.

[3] Devenow, A., & Welch, I. (1996). Rational herding in financial economics. European Economic Review, 40(3–5), 603–615.

[4] Lakonishok, J., Shleifer, A., & Vishny, R. W. (1992). The impact of institutional trading on stock prices. Journal of Financial Economics, 32(1), 23–43.

[5] Chang, E. C., Cheng, J. W., & Khorana, A. (2000). An examination of herd behavior in equity markets: An international perspective. Journal of Banking & Finance, 24(10), 1651–1679.

[6] Sias, R. W. (2004). Institutional herding. Review of Financial Studies, 17(1), 165–206.

[7] Kirilenko, A. A., Kyle, A. S., Samadi, M., & Tuzun, T. (2011). The flash crash: The impact of high frequency trading on an electronic market. Journal of Finance, 72(3), 967–998.

[8] Cont, R., & Bouchaud, J. P. (2000). Herd behavior and aggregate fluctuations in financial markets. Macroeconomic Dynamics, 4(2), 170–196.

[9] Easley, D., & O'Hara, M. (2010). Liquidity and valuation in an uncertain world. Journal of Financial Economics, 97(1), 1–11.

[10] Zhang, F. (2010). High-frequency trading, stock volatility, and price discovery. SSRN Working Paper.