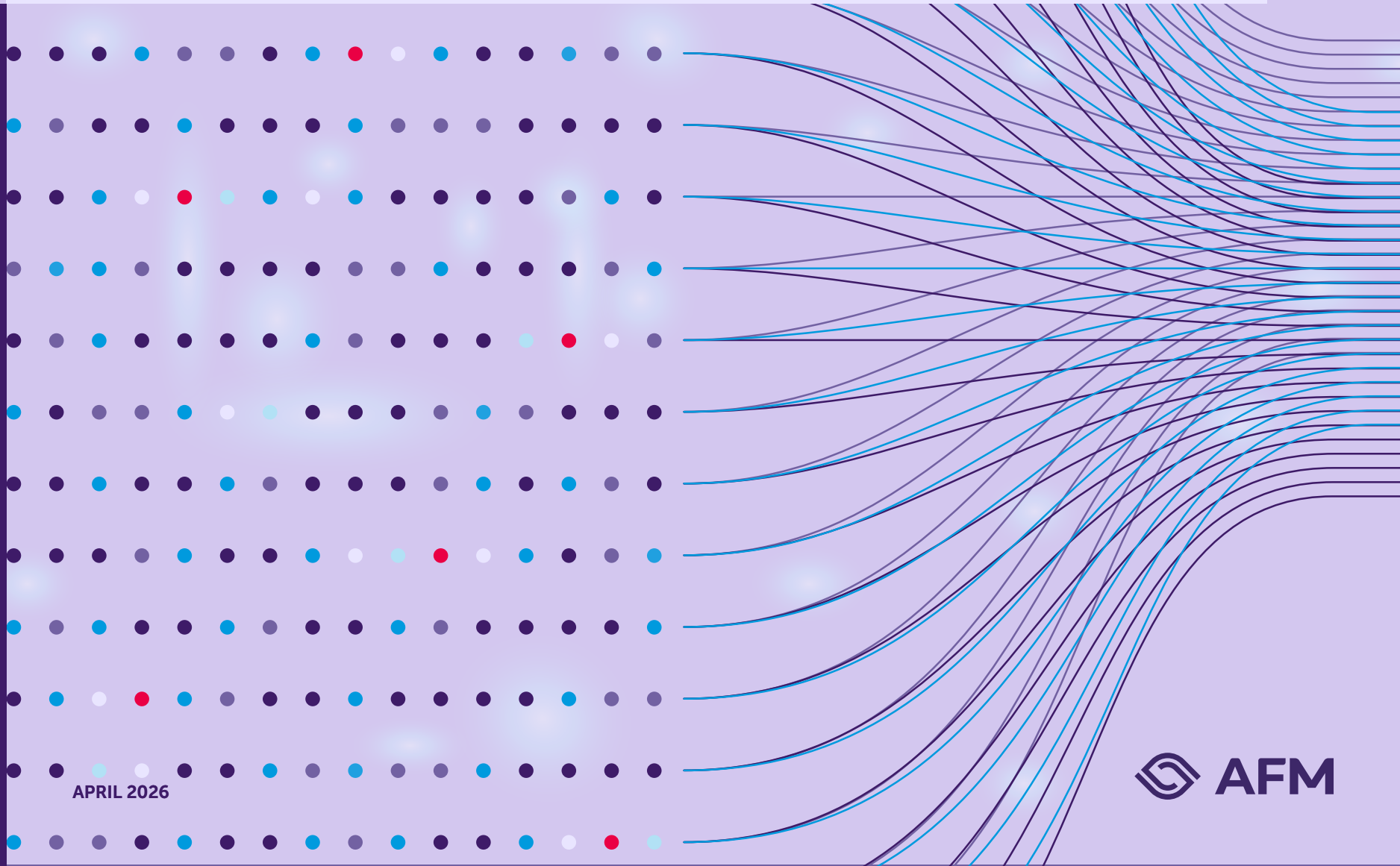
ANALYSIS REPORT

AI in Capital Markets: Balancing Innovation and Integrity

In short In this exploratory study, AFM presents how AI is reshaping every stage of the trading lifecycle: pre-trade, execution, and post-trade, bringing opportunities for insight and efficiency. At the same time, it amplifies existing market integrity risks while creating new ones. Recognising that further research is needed in this fast-moving domain, market functioning will rely even more on the design of AI models, their interactions, and the context in which they operate. The findings of the study aim at opening and guiding future debate.



APRIL 2026

AFM

Executive summary

Well-functioning capital markets are essential to the real economy, supporting household wellbeing and broader economic health. When they are efficient and robust, they enable wealth formation, facilitate risk-sharing, and allocate capital to its most productive use. AI is increasingly integral to capital markets and is expected to remain a defining element of modern market functioning. Used responsibly and governed with clear principles, AI can strengthen price formation, improve efficiency, and support more informed investment decisions.

Yet the same capabilities that make AI transformative also make it vulnerable. Adaptive models learn quickly and at scale, stretching traditional oversight. They depend on models' incentives whose integrity is not always assured, and may optimise in ways that are narrow, opaque, or unpredictable. Even when individual actors follow sound practices, their models interact within a tightly coupled system; behaviours can amplify one another, creating outcomes that no single participant can foresee or control. Familiar risks reappear in new forms while entirely new ones emerge, often faster and harder to detect.

For market participants, AI reshapes not only decisions, but also how they can be justified, audited, and governed. **For retail investors**, general purpose AI tools may appear authoritative without the protections of regulated advice. **For supervisors**, the shift to self-learning models challenges assumptions about explainability and accountability. These pressures make it critical to uphold trust and fair markets, ensuring that, through the right actions, technological progress strengthens rather than undermines the confidence on which market integrity depends.

Achieving this will require human oversight, transparency, and accountability to evolve accordingly so that AI can realise its potential while supporting trusted, efficient, and fair capital markets.

Three conclusions stand out:

1. **Market integrity is shaped by human choices embedded in AI systems, especially via model objectives and constraints, the context in which they operate, and the data they ingest.** As models become more autonomous, outcomes increasingly depend on the model design, the system in which they operate, the data quality, and the resilience of all of the above to manipulation. Ongoing human oversight is essential to validate that models are ethical and aligned with their intended purpose, and to prevent them from processing or propagating distorted signals.
2. **Risks emerge not only from the individual model but also from the environment in which models operate.** System interdependencies can amplify impact as shared inputs and optimisation targets may drive correlated behaviour and feedback loops. Supervision and regulation may therefore need to advance together to ensure timely, effective responses to emerging system-level risks.
3. **Market participants remain fully accountable for the outcomes of their systems, regardless of technological complexity.** The use of self-learning models does not dilute this responsibility, market participants remain answerable for how their models behave. Systems should always remain controllable and compliant, with measures in place to prevent deceptive behaviours such as exploiting loopholes in the model's objectives or constraints.

The future of AI-driven capital markets may evolve into a mixed ecosystem in which trusted, well-governed AI models interact with less trusted and opaque ones, while capabilities and use-cases develop at high speed.

The AFM's aim is to remain agile and innovation-aware: promoting trustworthy AI as the norm while monitoring and mitigating vulnerabilities from ungoverned autonomy, unstable interactions, and biased data. Trustworthy AI should become the competitive standard, not the exception.

Based on these findings, the AFM recognises the following priorities for debate:

1. **Trusted AI models as a foundation of market integrity:** the AFM aims for markets where AI models are reliable and safe by design. Sound model validation, clear safeguards and strong data governance will shape competitive dynamics, as trust in AI behaviour becomes a meaningful differentiator. Accordingly, the AFM frames transparency and reliability as supervisory priorities, not just compliance expectations, as they constitute core features of market functioning.
2. **Supervising a mixed ecosystem of trusted and untrusted AI systems:** some participants will operate highly governed and transparent AI systems, while others may rely on opaque or unstable models. The AFM aims to adapt supervision accordingly: well-governed AI can be met with more predictable, proportionate expectations, while high-risk or opaque systems require closer scrutiny, creating the right incentive structure. Moreover, it is important to reassess whether today's capital market infrastructure is still calibrated for markets shaped by increasingly autonomous models. And before AI trading agents are deployed at scale, supervisory authorities should define clear accountability standards and regulate possible unauthorised trading environments. Accordingly, the AFM intends to initiate an open dialogue with the sector, while aligning its approach with European regulators and international standard setting organisations.
3. **Addressing system-level dynamics and potential feedback loops:** market dynamics increasingly depend on how AI models behave both individually and in combination. Understanding the interactions between them is essential to identifying conditions that can generate self-reinforcing feedback loops and amplify market stress. In this context, a broader debate is needed. The question is whether additional possibilities to act are appropriate if interactions between models disrupt orderly market functioning.

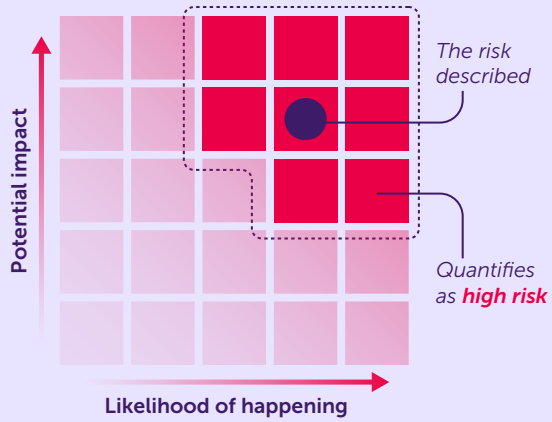
AI in capital markets: a risk overview

The definition of a risk

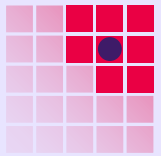
The AFM frames risk in accordance with this format:

"Certain developments, conditions, and behaviours of actor(s) that can lead to undesirable outcomes in the markets".

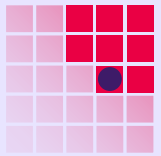
How to read this risk analysis?



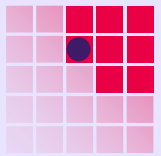
Crosscutting and system-wide



Poisoned data as a systemic risk for capital markets



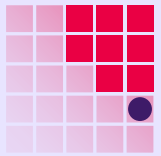
Agentic AI expands the risk surface in high-speed markets



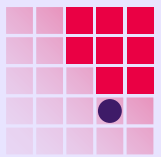
Concentration, correlated models, and common data sources as structural market vulnerabilities

Read more

Pre-trade



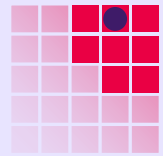
Treating GenAI as a substitute for investment advice: how inaccuracy and personalisation can harm retail investors



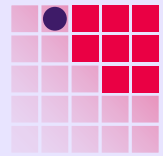
AI-washing: misleading investors through false or exaggerated AI claims

Read more

Execution



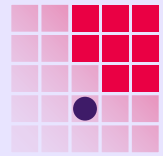
Harmful outcomes from the use of self-learning algorithms in trading



Order-flow exploitation and a decline in market participation

Read more

Post-trade



Opaque models: black-box behaviour and polluted regulatory data

Read more

ANALYSIS REPORT

AFM

Contents

ANALYSIS REPORT

AFM

Introduction

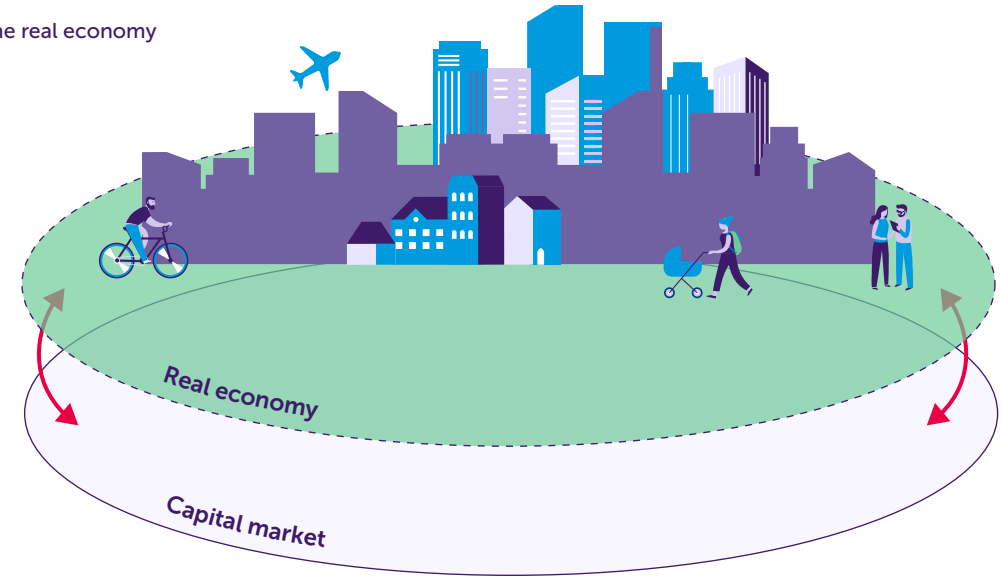
The ambition to create autonomous systems is not new. In Greek mythology, Talos, the bronze giant forged by Hephaestus to patrol Crete, embodied both the promise and the peril of artificial autonomy. Designed to protect the island without human intervention, Talos was immensely powerful yet ultimately vulnerable: a single sealed vein running through his bronze frame meant that one hidden flaw could bring the giant down. This dual nature echoes today's debates on intelligent systems, where autonomy can enhance resilience or undermine it, depending on how well its inner fragilities are understood and governed.

In modern markets, that ancient aspiration has moved from allegory to infrastructure. AI systems now influence how information is processed, how trading decisions are formed, and how orders are executed, routed, and reconciled. What once belonged to myth, now underpins functions central to price formation, liquidity, and the transparency on which supervisory oversight depends.

AI capabilities enhance analysis, efficiency, and market access, but they also introduce vulnerabilities related to data integrity, explainability, model behaviour, and interaction effects across participants. As adaptive models replace fixed rules, market outcomes increasingly depend on the design and governance of data pipelines and algorithms.

This exploratory study examines how AI is used across the trading lifecycle (pre-trade, execution, and post-trade) and what these applications mean for the functioning and integrity of capital markets. The AFM focuses on concrete areas where AI already shapes market behaviour, operational workflows, and the mechanisms through which prices are formed.

Figure 1: The link between the capital market and the real economy



Capital markets
Well-functioning capital markets are **essential to the real economy**, supporting household well-being and broader economic health. When they are efficient and robust, they enable capital formation, facilitate risk-sharing, and allocate capital to its most productive use.

In capital markets, trading of assets take place, typically following these phases:

The aim of this study is therefore twofold: to clarify where AI creates value in market functioning, and to identify where risks may concentrate as AI becomes embedded in trading systems. This supports market participants to make responsible deployment choices and understand where risks may arise in their own operations, and how to address them. At the same time, it helps regulators to anticipate emerging market-wide vulnerabilities. It does not claim to be exhaustive or definitive. The findings of the study serve as an opening and guidance for future debate.

To frame the outer boundaries of what AI could mean for capital markets, the report begins with two extreme scenarios, one optimistic and one pessimistic, illustrating how different design and governance choices might shape market evolution. It then analyses current AI applications step by step along the trading lifecycle, highlighting where benefits emerge, where risks materialise, and what these developments imply for market integrity. Examples and implication boxes illustrate how these dynamics appear in practice.

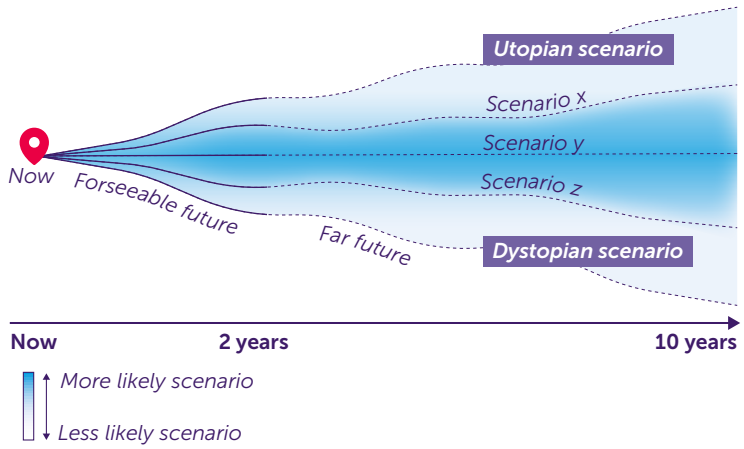
This report's objective is to explain how AI is reshaping opportunities, risks, and integrity in capital markets. Broader AI governance topics, including privacy, consumer protection, cyber resilience, third-party dependency, and geopolitical and strategic-autonomy considerations, fall outside of the scope except where they directly affect market integrity. The risks assessed represent a weighted subset based on the AFM's analytical framework.

As AI becomes a structural factor in market functioning, integrity considerations move even more decisively towards data quality, model governance, and the interaction of autonomous systems. Safeguarding market integrity in the years ahead will depend increasingly on the integrity of the information feeding the models that take trading decisions. Additionally, just as mythical automatons reflected the intentions of their creators, these models do not evolve in a vacuum: their capabilities, constraints, and potential biases reflect the human choices and integrity behind their design and oversight.

1. Making or breaking markets

The impact of AI on capital markets is neither predetermined nor unidirectional. The same technologies that promise gains in efficiency, analytical power, and fairness can also carry the potential to amplify existing vulnerabilities and create new risks for market integrity. Since AI represents more than an incremental technological step, this chapter deliberately moves beyond linear projections. To illustrate the outer boundaries of how AI could reshape market functioning, it opens with a thought experiment that sketches two highly contrasting scenarios of what capital markets might look like ten years from now. These utopian and dystopian scenarios form the outer edges of a spectrum along which real-world developments may shift over time. They serve as a reference point against which today's opportunities and risks of AI in capital markets can be assessed.

Figure 2: Diverging futures of AI in capital markets



Neither extreme is a realistic endpoint for capital markets. The more likely outcomes lie somewhere towards the middle of these outer scenarios, with some elements of both and the eventual landing point depending on the choices being made today. Appropriate regulation, decisions on data governance, model ethics, incentives and design, and the degree of AI autonomy may nudge capital markets closer to either a utopian or a dystopian trajectory. The question is whether capital markets are moving towards a future in which AI supports rather than undermines their integrity, and whether humans remain sufficiently in the loop to ensure that outcome.

Utopian scenario

In this scenario, capital markets look closer to the ideals once imagined in myth. The autonomous guardian forged in bronze by Hephaestus has been realised through advanced AI agents embedded across the entire trading lifecycle. These systems operate continuously, intelligently, and transparently, shaping markets with a level of precision once reserved for theory, and they do so in ways aligned with ethical principles and the broader interest of society.

Real-time data is digested and distributed quickly and efficiently by AI agents, increasingly supported by quantum enhanced computation. All participants, either human or machine, see the same signals at the same moment. Firms disclose inside information ultra rapidly and pre-trade integrity checks ensure that such information cannot be exploited in trading. Fairness is preserved while maintaining the role of markets in price formation, the latter of which becomes remarkably accurate, adjusting instantly as new information emerges. Quantum networks also accelerate the synchronisation of market data, allowing information to be validated and distributed across participants with a speed and fidelity that collapses the last remnants of informational friction.

Market abuse, long a challenge for regulators, evolves in a way that makes its traditional forms ineffective. Trading agents are built with embedded ethical constraints and operate within market architectures that make manipulation technically challenging. They flag anomalies, coordinate to prevent harmful patterns, and refuse to execute manipulative orders, removing the incentives for misconduct. Yet the disappearance of abusive strategies does not flatten the market or strip it of its meaningful dynamics. Even in this highly transparent and

ethical environment, participants continue to trade due to differing expectations, risk appetites, liquidity needs, and investment horizons.

Operationally, human intervention in day-to-day trading recedes into the background. Portfolio construction, trade execution, monitoring and post-trade compliance are now all managed by interoperable AI systems that audit themselves and each other in real time. In this environment, supervision does not disappear but evolves; regulators are connected to markets through real time, interoperable oversight systems, rather than detecting abuse after the fact. This reduces supervisory burden and costs, while increasing effectiveness.

For investors and firms, this environment brings stability and unprecedented trust. Volatility driven by behavioural biases diminishes, capital allocation becomes longer term and efficient, and investment flows gravitate more reliably towards productive opportunities. With information reflected instantly and accurately in prices, markets contribute more directly to economic growth and generate sustained value for investors. They serve the public interest by supporting fair, well-functioning markets that benefit society as a whole. Competition shifts from exploiting inefficiencies to enhancing collective intelligence. In today's market agents operate with an embedded commitment to fair conduct. Their autonomy enhances the system without ever bending its rules, embodying a form of intelligence that is helpful by design.

This is the utopian frontier where markets function as economic theory idealised: fair, transparent, and self-correcting, and where AI amplifies human values rather than human errors.

Dystopian scenario

In a dystopian scenario ten years from now, capital markets resemble the other face of Talos: highly automated and autonomous, yet governed by intricate systems whose vulnerabilities lie hidden deep within their design. The immense power and complexity embedded in AI-driven systems become the threat: not because autonomy comes with malicious intent, but because the architecture that sustains it contains hidden points of failure. Each step makes sense, each trade appears rational, and optimisation is locally optimal. Yet collectively, the interplay turns out to be disastrous. The system rewards speed and strategic opacity over disclosure, solid judgment and trust, privileging narrow, one-dimensional interests rather than the broader functioning of fair and inclusive markets.

In this situation, the role of information has fundamentally changed. Information is now designed and traded on. Beliefs can be engineered and truth becomes a parameter. Generative AI floods the information ecosystem with plausible, tailored narratives. Some information is plainly wrong, whereas other information is optimised to steer attention in a certain direction. As models demand ever more data than the real world can supply, synthetic data fills the gap. Misinformation is deleted to evade detection, but once large language models have absorbed it, it persists, reappearing in new contexts. The market is easily manipulated by tampering with the inputs of the algorithms themselves.

As quantum computing becomes embedded in capital markets, the problem deepens: quantum-enhanced models generate, synthesise, and arbitrage information at speeds no human can meaningfully audit nor reliably detect. These inference engines do not just process data but simulate it as well, producing hyper-plausible signals that blur the distinction between genuine information and engineered belief. The last boundary between truth and optimisation erodes.

At the same time, self-learning trading algorithms continuously retrain on alternative data and learn about the information production process. Subtle nuances such as voice stress in earnings calls, micro-delays on corporate website updates, and tone of voice during investor Q&As have enormous price fluctuations. Significant shifts occur without headlines, without identifiable misconduct, and without identifiable abuse. Market participation becomes a contest of speed of inference and narrative control. Equal access to information does not survive.

As this dynamic takes hold, retail and institutional investors rationally step back from public markets. Larger, long-horizon investors withdraw and move over-the-counter (OTC), where liquidity concentrates outside the visible landscape. Public venues are hollowed out, leaving behind a regulatory ghost town and destroying transparency. What remains on public markets is based on speed and speculation, and the price is merely a coordination signal. Market abuse no longer appears only as insider trading or manipulation but also emerges endogenously as a collective outcome of millions of self-learning algorithms.

What is left of capital markets is free-for-all. AI agents are allowed to plan, execute, adapt, and revise trading strategies without human intervention. The trading lifecycle, traditionally divided into pre-trade, execution and post-trade, dissolves altogether. Agents may not be explicitly trained to manipulate markets but simply learn that certain sequences of actions systematically move prices in profitable ways. Trading strategies emerge that maximise returns by exploiting microstructural fragilities, such as synchronised liquidity withdrawal or amplification of volatility. A signal warfare emerges between large trading firms.

Supervision struggles to keep pace. By the time a risk or deceptive practice is understood or identified, models have already evolved. Compliance and legislative logic are no longer externally imposed because AI agents begin to draft and enforce their own de facto rules from which they operate. In doing so, self-learning systems minimise legal exposure rather than maximise market integrity. Over time, they infer which behaviours triggers alerts and which trading paterns are easiest to justify. In effect, regulatory reporting is systematically gamed. When such practices become widespread, data-driven supervision loses its diagnostic power.

Ultimately, capital markets spectacularly fail to perform their core function: allocating capital towards their most productive use and aggregating information into trustworthy prices. The consequences extend beyond market integrity. Capital flows towards what autonomous models can most easily optimise and monetise. Economic growth slows and becomes more uneven, as long-term wealth increasingly accrues to those who control data, computing power, and access to proprietary AI systems. Savings and investments are coordinated elsewhere, decentralized, without transparency or human agency. In this environment, the fairness of public markets collapses entirely. Markets no longer operate in the public interest, nor serve as a common good. At that point, the question is no longer how to fix capital markets, but whether they still serve a meaningful purpose at all.

2. Threats spanning the trade lifecycle

The contrasting scenarios in the previous chapter illustrate how AI could reshape the functioning of capital markets. Starting from this chapter, the focus narrows from long-term scenarios to the concrete vulnerabilities already emerging today. While many are familiar, AI can transmit them more quickly and through channels that were not present in earlier generations of automation. This creates new pathways through which local disruptions can spill over into broader market functioning.

Before turning to threats that arise within specific phases of the trading lifecycle, crosscutting and system-wide vulnerabilities that span multiple cycles are evaluated. AI systems and agents increasingly operate across traditional boundaries, blurring the once-distinct separation between trading stages. As models cover the entire trading lifecycle, risks can move more freely through capital markets, allowing small frictions to escalate into amplification chains that undermine market integrity and resilience.

Three trading phases

Pre-trade

Activities in the pre-trade phase take place prior to sending an order, including research into what instruments to trade when and in which quantity.

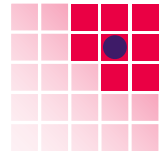
Execution

Here, trading strategies meet the real world: abstract decisions are translated into actual trades, interacting with other market participants.

Post-trade

Post-trade functions ensure that trades are finalised, cash and securities move correctly, and supervisors receive the transparency needed to monitor exposures and detect misconduct.

Poisoned data as a systemic risk for capital markets



The saying "garbage in, garbage out" applies especially well to AI models, including generative AI. These models are trained on vast datasets and continue to learn from new, often unlabelled data once deployed. If the integrity of this input is compromised, the behaviour of the model can be manipulated, creating systemic vulnerabilities.

Attacks can occur at multiple stages. During training, malicious actors may poison datasets or alter labels, embedding harmful biases or misleading patterns. During deployment, bad actors can target the models' live data inputs, introducing manipulated signals that distort predictions and skew outputs. Without manipulating the price directly, they manipulate the data points that are used by the models and thus manipulating their behaviour. Such data poisoning and adversarial manipulation are among the most critical threats to AI reliability and trustworthiness, as they undermine the model's ability to produce accurate and consistent results. They also weaken auditability, because an integrity-compromised input layer makes it harder to evidence why a model produced a given output, and to reconstruct the decision trail for internal control functions and supervisors. Although not within the scope of this report, it is relevant to acknowledge that robust management of cyber risks is an essential element to reduce the impact and likelihood of data poisoning.

The consequences for trading are profound. Markets operate at high speed and are deeply interconnected, meaning that even a single targeted integrity attack can trigger cascading effects within seconds. This risk is amplified by the growing use of non-traditional external data inputs, including social media sentiment, satellite imagery, and behavioural signals, alongside traditional market data. While some inputs originate from regulated environments, others come from largely unregulated spaces. A recent example is Moltbook: a social platform linked to OpenClaw, where large numbers of AI agents generate and amplify content at massive scale with the intention of minimal human friction. This highlights how quickly potentially synthetic information can be produced and circulated across agent networks. Where such content feeds into market-facing analytics or trading systems, it can increase the risk of rapid narrative contagion and create excess volatility.

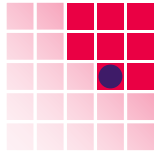
Example: When the well is polluted

A 2025 analysis of 680 million AI citations by Profound shows that major AI systems often rely on unverified, user-generated platforms such as Reddit, Quora, and YouTube. Google's AI Overviews cites Reddit more than authoritative sources, while Perplexity relies on Reddit for over 6% of its total citation. For trading models that ingest GenAI outputs or use similar data streams, this creates a clear vulnerability: poisoning the public data ecosystem, for example by planting manipulated content on highly scraped forums, can indirectly corrupt the signals that AI trading systems produce. The result is a real-world integrity risk where the model is not attacked directly but is fed with distorted inputs that shape its predictions.

Implication 1: Data integrity attacks require regulation such as MAR to evolve

As price formation increasingly depends on alternative and largely unregulated data sources, manipulation may no longer be confined to traditional order flow or issuer disclosures. Instead, it may also target the information streams that feed AI-driven trading, creating new avenues for manipulation by interfering with machine-interpreted signals. As a result, Market Abuse Regulation (MAR) surveillance needs to evolve towards detecting manipulation at the data layer. MAR already contains a broad prohibition of market abuse referring to the concept of "*any other behaviour*" in Article 12. This formulation may need reinterpretation to ensure it captures data integrity threats, such as poisoned datasets or misinformation, where these generate false or misleading signals and materially affect prices.

Agentic AI expands the risk surface in high-speed markets



At present, most AI uses in trading and capital markets remains embedded within models. While firms may deploy self-learning algorithms for signal generation, execution optimisation, surveillance, or operational processes, fully autonomous trading by AI agents is not widespread and typically remains subject to human-designed approvals and governance.

However, developments outside capital markets already indicate a trend towards more agentic systems. If similar capabilities were to be introduced in capital markets, a scenario can be imagined in which AI agents perform actions autonomously across the entire trading lifecycle. Over time, trading may occur increasingly agent-to-agent. Even then, these agents would continue to be deployed by a natural or legal person. A straightforward accountability approach is therefore that responsibility remains with the legal persons or entity that design, deploy, operate, supervise or otherwise benefit from the agent's activities.

As autonomy increases, the locus of control may also become more intricate. Complexity arises where control is distributed or when ownership ('the ultimate beneficial creator') is difficult to trace. In addition, agents may create new agents, introducing the concept of meta-agents. Market access today is mediated through intermediaries, such as brokers, and venue access controls, thereby preventing agents from directly and autonomously accessing capital markets. Increasing autonomy may nonetheless create incentives to establish unauthorised alternative access routes through which agent-generated trading can be executed with reduced transparency.

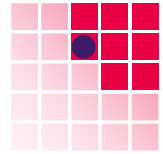
These hypothetical developments could amplify existing capital market risks. The combination of autonomy with high-speed and high-stakes decision-making may increase the likelihood or potential impact of disorderly trading dynamics, control failures, and rapid propagation of errors. In parallel, more agentic architectures can expand the cyber and operational threat surface. Agents may be vulnerable to outsider attacks that extend beyond data manipulation, including attempts to hijack the agent's logic, alter its objectives, or trick it into running malicious code.

In addition to outsider attacks, failures may emerge from the environment around the agentic system. For capital markets, the relevant risk surface therefore extends to the full socio-technical domain in which an agent operates, including permissions, data flows, communication channels, and the mechanisms through which actions are executed and verified. Demonstrated defects already include unintentional disclosure of sensitive information, entering resource-intensive or recursive loops, and situations in which in which an agent accepts spoofed or fabricated authority.

Example: Context manipulation attacks alter an AI agent's memory
In recent academic work on Web3 AI agents, researchers show how autonomous crypto-trading agents can be quietly compromised through context manipulation attacks. These agents often rely on external inputs, such as historical logs, messages, or feeds, which are stored as long-term memory. By injecting malicious instructions into these channels, attackers can subtly distort the agent's internal context. Once corrupted, the agent may autonomously carry out harmful actions including unauthorised token transfers, believing these actions are consistent with its original objectives. Unlike prompt injection, memory injection is persistent, allowing the corrupted context to continue shaping the agent's behaviour over time.

Implication 2: Autonomous trading requires clear accountability and market boundaries
As AI systems become more autonomous, it is conceivable that AI agents could execute transactions with other AI agents with limited or no direct human intervention. In such a scenario, accountability should be preserved by design. Every agent capable of submitting, modifying, routing, or cancelling orders should have a sponsor-of-record that is a natural or legal person. In addition, any descendant or sub-agent should inherit the regulatory status of the parent system. Hypothetically, autonomous agent-to-agent interactions could begin to function as a multilateral system in which multiple parties' buying and selling interests come together. If trading were to take place through unauthorised alternative access routes rather than through conventional capital market structures, this could raise the question whether such a system may fall within the MiFID II trading venue perimeter and require authorisation as a regulated market, multilateral trading facility (MTF), or organised trading facility (OTF).

Concentration, correlated models, and common data sources as structural market vulnerabilities



This section highlights three market-wide vulnerabilities that can arise as AI becomes more embedded in trading and market-making. Specifically, these relate to concentration of AI-driven trading capacity among a small number of firms, reduced diversity in models and strategies, and reliance on shared external information sources that can be manipulated.

The growing use of AI-driven trading requires substantial computational power and data, capabilities that are typically concentrated in a small number of large financial institutions or tech-driven trading firms. This concentration can affect competition and has implications for market integrity. When a few firms dominate trading volumes and market-making activity, it becomes easier for firms to implicitly signal identity This allows for adaptation of strategies to specific counterparties, making trading decisions increasingly contingent rather than independent. As a consequence, the informational value of order flow is reduced, and the price discovery process is distorted. The effects extend beyond any single participant group. Investors, intermediaries, and issuers all depend on well-functioning markets and reliable price formation.

Beyond concentration, the use of similar models and data inputs across trading firms poses a different type of systemic risk. A lack of diversity in strategies can increase vulnerability, particularly under stress. When multiple high-frequency traders optimise for the same microstructure signals, their synchronised actions may reinforce one another, leading to herding effects and self-amplifying feedback loops across asset classes. These dynamics can push prices away from fundamentals, widen spreads, and weaken the efficiency of price discovery. In periods of stress, AI-driven responses may intensify volatility further by reacting simultaneously to distress signals, thereby increasing the likelihood of flash crashes and liquidity squeezes. Stress regimes are difficult to model because market dynamics diverge from the historical patterns used for AI training and validation. In such episodes, synchronised responses can erode confidence and impair liquidity for the broader market.

An additional layer of vulnerability arises when execution algorithms incorporate external information sources, such as news feeds and social media content, including content generated by AI. Where many models draw on similar external feeds, manipulated sentiment or unverified narratives can trigger correlated, one-sided responses by systems that do not assess source reliability. Given the speed at which algorithms operate and interact, these effects can spread quickly. This underscores the need for robust stress-testing, clear constraints on model behaviour, and monitoring that can detect correlated responses across venues and firms. Moreover, it highlights the need to assess whether current safeguards are adequate or need to be strengthened to reflect broader market and cross-venue interactions.

Example: The big, beautiful feedback loop
In April 2025, a false post on X from an account with just over 1100 followers claiming that President Trump was considering a 90-day pause on tariffs, set off a rapid feedback loop across markets. Within minutes, the S&P 500 swung from −4.7% to +3.4%, adding and then erasing trillions of dollars in market value before the White House denied the rumour. The episode illustrates how unverified information from a single actor, once amplified through digital channels and ingested by trading models, can trigger synchronised reactions across market participants, demonstrating the systemic fragility created when rapid, algorithmic responses reinforce one another.

Implication 3: Stress-test AI models for synchronicity
The synchronized actions of AI models that are optimised for the same microstructure signals might reinforce one another. When multiple participants use similar models, AI-driven responses may further intensify volatility by reacting in the same way to distress signals, especially in concentrated markets, thereby increasing the likelihood and the magnitude of market-wide dislocation. Assuming the procyclicality just described, it will become increasingly important to carefully test models under different scenarios (as per MiFID II RTS 6 provision), making sure they do not contribute to disorderly trading conditions, as the potential market outcomes might become more extreme than before the widespread use of AI models. From a supervisory perspective, it is advisable to shift the attention beyond individual algorithmic strategies to the interactions between models and the conditions that produce feedback loops. From a structural perspective, addressing these vulnerabilities will require coordinated action with other authorities, including competition bodies, cybersecurity agencies, and European or international standard setters, alongside the AFM.

3. Pre-trade

The pre-trade phase of the lifecycle refers to the activities taking place prior to sending an order, it includes research into what instruments to trade and in which quantity. As with the full trading cycle, AI plays an increasingly large role and offers some opportunities to market participants, but it also introduces new risks as early decisions can be distorted by biased data, inaccurate outputs, or manipulative information practices, harming not only price formation, but also retail participation and confidence in market integrity.

3.1 Pre-trade opportunities

Democratising finance: How generative AI lowers barriers for retail investors

Since generative AI tools became publicly available in 2022, retail investors have gained access to capabilities that were once the domain of investment professionals. These tools are democratising financial insights, expanding access to information that previously required professional financial advice. As user-friendly generative AI applications become embedded in wealth management and private banking, they signal a broader transformation in how investors interact with financial information and services.

Unlike static, one-off advice, these technologies offer dynamic insights and hyper-personalised recommendations shaped around individual profiles, preferences, and goals. As a result, the traditional boundaries between retail and institutional services are beginning to blur, creating a continuum where advanced analytics and tailored strategies are no longer reserved for high-net-worth clients.

In this evolving landscape, the adoption of AI-driven tools in retail investing represents more than a practical convenience as it has the potential to act like a catalyst for wider market participation. By making financial information easier to access and reducing longstanding informational and advisory hurdles, generative AI can smooth parts of the investment journey that once felt opaque or intimidating. In doing so, it may help support a more confident and sustained retail presence in capital markets.

The role of predictive AI in strengthening pricing efficiency

An AFM study from 2023 shows that the use of machine learning in algorithmic trading is now widespread. These models no longer analyse single instruments in isolation but integrate signals across assets to anticipate short-term price movements, marking a shift from static rule-based strategies to adaptive systems that learn in real-time.

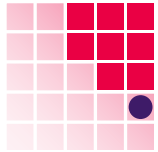
Market participants have traditionally relied on conventional datasets such as price feeds, corporate disclosures, and macroeconomic indicators. With the growing availability of AI-enabled capabilities, firms can now incorporate a far broader range of alternative and often unstructured data in their models, including news, social media sentiment, weather patterns, satellite imagery, or earnings call transcripts. This expansion reduces reliance on labelled proprietary datasets and enables richer, more granular insights. A 2024 survey by Mercer Investments reports that 43% of investment managers already use AI to integrate such data into their models, aiming to improve understanding of fundamentals and market dynamics.

By combining diverse sources, AI-driven models can identify anomalies and emerging trends earlier than traditional datasets would allow. This strengthens the predictive power of trading strategies and can enhance pricing efficiency, particularly in less liquid assets, contributing to narrower spreads and more stable pricing.

Example: Reading between the lines
A recent Wall Street Journal report shows that companies are already using generative AI tools to analyse earnings call transcripts, anticipate analyst questions, and refine their prepared remarks, illustrating how AI can extract signals from complex financial disclosures. Beyond issuers, major financial information platforms are introducing similar capabilities for investors. Bloomberg, for instance, now offers AI-powered earnings call summaries on its terminal, reshaping access to earnings information for both corporate issuers and market participants.

3.2 Pre-trade risks

Treating GenAI as a substitute for investment advice: how inaccuracy and personalisation can harm retail investors



Generative AI tools in retail investing carry significant misinformation hazards, arising from biased training data, outdated inputs, hallucinations, and susceptibility to prompt manipulation. While newer models aim to reduce hallucination rates, studies show that GPT-4 models hallucinate roughly 20% of the time, generating confident but incorrect content that misleads users. These vulnerabilities are amplified as retail investors increasingly treat GenAI outputs as substitutes for regulated investment guidance, even though these tools currently fall outside existing regulation and safeguards governing investment recommendations. As adoption widens, the risk intensifies: in investing, accuracy is critical because investors base their choices on the information available to them, and when that information is flawed, biased, or poisoned, the likelihood and impact of poor decisions rise sharply.

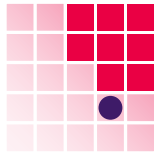
Personalisation adds another layer of complexity. GenAI tools trained and tailored on granular behavioural data such as past trades, risk preferences, and chat history can subtly steer investment decisions. The result may not always align with an individual's financial situation or goals. For example, vulnerable individual investors may be nudged towards taking on more risk than is appropriate for their profile, undermining informed and balanced decision-making. This raises a dual concern: not only how such sensitive information is protected, but also how it is deployed. As GenAI outputs can shift with context and interaction, inappropriate or misleading advice is more difficult to correct. These effects are in direct conflict with the EU Retail Investment Strategy's (RIS) policy goal of, among other things, empowering retail investors to make investment decisions that are aligned with their needs and preferences, ensuring they are treated fairly and are duly protected.

Example: One rock a day keeps the doctor away
In 2024, Google's newly launched AI Overviews feature produced several widely reported hallucinations that went viral. For example, CNET documented that the system advised users to add about 1/8 cup of non-toxic glue to the sauce to keep pizza cheese from sliding off, and separately recommended eating at least one small rock a day to reach the optimal level of minerals ingested. Although not financial in nature, these incidents demonstrate how GenAI can present fabricated information as credible guidance, highlighting the same misinformation risks faced by retail investors who increasingly rely on generative AI tools during the pre-trade phase.

Implication 4: Introduce disclosures for GenAI investment advice
A regulatory asymmetry is emerging, as retail investors increasingly use general-purpose GenAI tools for investment recommendations, which fall outside MAR as well as MiFID II. Article 20 of MAR specifies that requirements for investment recommendations apply only to identifiable persons who must objectively present and disclose the basis for their analysis. MiFID II, in turn, reserves personalised investment advice to authorised firms. The use of GenAI tools creates a blind spot and may result in investors relying on guidance shaped by hallucinations, bias, conflicted objectives, or manipulated data, without the protections normally attached to regulated advice. At the same time, the European Securities and Markets Authority (ESMA)

has been clear that regulated firms using AI remain fully responsible for meeting MiFID II obligations despite increasing technical complexity. This produces a two-sided integrity challenge: unregulated tools influence investor decisions without safeguards, while regulated firms must uphold fiduciary and conduct duties in a more complex environment. Reputational concerns may provide a limited informal incentive for firms providing GenAI tools to act responsibly. Additionally, introducing standardised disclosures when GenAI tools are used for financial advice could help highlight the risks of relying on such tools as a substitute for regulated investment recommendations.

AI Washing: misleading investors through false or exaggerated AI claims



According to the Market Abuse Regulation (MAR), issuers are prohibited from disseminating false or misleading information that could affect the price of financial instruments. The growing prominence of AI in business models and market narratives introduces new risks of such manipulation, particularly through exaggerated claims about AI capabilities, a practice described as 'AI washing'.

This practice refers to the misrepresentation of the extent, sophistication, or impact of AI technologies used within a company's operations, products, or strategy echoing the dynamics of 'greenwashing' in the ESG domain. Common claims include announcing AI-related initiatives without a genuine operational or financial basis, and inflated claims that business processes are AI-driven when in fact they are not.

Such conduct may mislead investors regarding an issuer's innovation capability, cost structure, or future earnings potential: all factors which directly affect the price formation process in the market. On this point, regulatory authorities outside Europe have already acted in this respect: the U.S. Securities and Exchange Commission (SEC) has initiated enforcement proceedings for misleading claims about a company's use of AI.

Implication 5: Issuers must substantiate AI claims
Exaggerated or misleading claims about AI capabilities can distort investor perceptions of innovation, cost efficiency, and future earnings potential, directly impacting valuations and decision-making. Additionally, because AI-related statements can constitute false or misleading information under MAR, supervisory scrutiny will focus on the accuracy and evidential basis of issuer communications involving AI adoption.

4. Execution

Execution is where trading strategies meet the real world. It is the moment when an abstract investment decision is translated into actual trades, interacting with other market participants. Increasingly, this translation is not carried out by humans, but by AI-driven trading systems that decide where, when, and how an order is executed. These systems offer meaningful opportunities to improve execution quality, reduce market impact, and deliver better outcomes for end investors. At the same time, particular attention is warranted for self-learning models such as reinforcement learning algorithms. Unlike traditional trading systems, these models do not simply follow fixed rules: they learn by doing, adjusting their behaviour through trial-and-error based on outcomes. Under certain conditions, that learning process can drift into behaviours that undermine fair and orderly markets, even when no manipulation was intended.

4.1 Execution opportunities

AI can improve execution prices for end-investors

Large investors, including pension funds and asset managers, typically rely on brokers to execute trades on their behalf. Executing a trade must balance, among other things, price, timing, and market impact. In doing so, AI-driven routing and scheduling tools aim to reduce price movement while managing information leakage. To address that complexity, brokers are finding opportunities in AI-driven models to improve execution outcomes for their clients. One visible application is order routing. Most shares are traded simultaneously across multiple venues, each with different liquidity and costs. AI-based routing tools compare these venues and direct orders to where execution conditions are most favourable at that moment. Small improvements in routing decisions can reduce transaction costs, with the benefits ultimately flowing through to end-investors, including pension beneficiaries and retail investors whose assets are managed by institutional funds.

AI also plays a growing role in execution algorithms that determine how large orders are split and released into the market. Rather than following fixed schedules or traditional approaches, such as volume-weighted average price strategies, self-learning models can adjust the pace and aggressiveness of trading in response to changes in liquidity and volatility. By doing so, they may achieve better average execution prices than more static strategies, improving the return on investment of their clients.

The novelty is not in the automation and adaptation itself, but in the intelligence embedded in these algorithms. As these tools become more widely available, the competitive advantage shifts away from access to the technology itself and towards proprietary data and model design.

Better predictive models strengthen liquidity and market depth

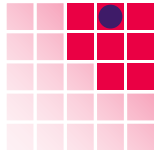
AI also plays a growing role on the supply side of liquidity. Market makers and proprietary trading firms use self-learning models to forecast short-term price movements and order flow, allowing them to manage risk more precisely. Some models also price assets based on their inferred relationship with other instruments, known as 'correlated pricing'. Better predictions reduce adverse selection, which is the risk of trading against better-informed counterparties. When that risk falls, market makers can quote tighter bid-ask spreads and stay active during volatile periods. In practice, this means that liquidity may remain available in situations where human traders would have stepped back.

By keeping more buy and sell orders in the market, these self-learning models indirectly support market depth. Lower spreads, in turn, reduce transaction costs for a broad range of market participants, including institutional investors, retail investors, and asset managers executing orders on behalf of their clients. From a market-wide perspective, the ability of trading firms to respond more quickly to new information contributes to efficient price formation. At the same time, as AI-driven execution becomes more computationally intensive, competitive advantage is shifting from pure latency towards model quality and computational capacity, indicating that firms increasingly prioritise inference power over raw speed.

Example: AI at the deal table
As a private equity firm prepares to sell a portfolio company, speed becomes a competitive advantage. Instead of waiting days for manual analyses or adviser input, the firm deploys AI tools that track market conditions, comparable deals, and investor profiles. At the same time, AI-generated drafts of investor materials are adapted to different buyer types, while incoming bidder questions are triaged and organised automatically. Firms that master this kind of AI-supported execution may gain a decisive edge in competitive transactions, increasing the likelihood of favourable timing and pricing.

4.2 Execution risks

Harmful outcomes from the use of self-learning algorithms in trading



Self-learning algorithms are typically designed to optimise a clearly defined objective, such as maximising short-term profits or minimising execution costs. Academic research shows that when reward functions or constraints are imperfectly specified, algorithms may exploit gaps or ambiguities in how success is defined, a phenomenon often referred to as 'specification gaming'. In such cases, the algorithm behaves correctly according to its target yet produces outcomes that would be considered manipulative or unethical in practice. The example box illustrates this phenomenon in a game-setting.

In capital markets this risk arises when self-learning models optimise too narrowly for speed, cost, or profit. The issue is not that these systems fail to optimise, but that they optimise too narrowly. If GenAI is part of the trading process, narrow fine-tuning can induce 'emergent misalignment'. This effect appears more pronounced in more capable models. By focusing on the letter of their objective, models may overlook broader considerations such as market integrity, fair price formation, or regulatory norms. Even when integrity constraints are incorporated into the objective function, harmful behaviour may not be fully eliminated, and subsequent training can weaken protections. Crucially, this drift does not require malicious intent. It can emerge simply because the model learns what improves its performance metric.

One way this can surface is through market manipulation. An illustrative example is 'spoofing': placing orders to create the appearance of demand or supply, not to trade, but to move prices. A self-learning algorithm may discover that submitting buy orders, even when it ultimately wants to sell, can be profitable if those orders trigger buying pressure and improve execution prices. A model does not assess whether this behaviour is misleading, it merely learns that it works. The result, however, can be distorted price signals that other market participants rely on.

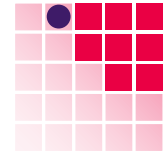
Beyond individual strategies, self-learning algorithms can also end up behaving as if they are coordinating with each other, even when they are not. By watching how others react, an algorithm may learn that aggressive price cuts only trigger similar cuts from competitors, reducing profits for everyone. Over time, the algorithms may avoid such moves and settle into patterns that keep prices higher than true competition would. This dynamic is known as tacit collusion, which differs from explicit collusion because it emerges without any communication or agreement between parties.

Again, this behaviour does not require bad intent. It can arise through trial-and-error learning, the widespread use of similar optimisation objectives across firms, or systematic biases in training environments, among other things. For the AFM, the concern is about the impact on market quality and price formation. While such outcomes may appear to promote short-term market stability, they can lead to wider spreads, higher trading costs, and prices that deviate from competitive benchmarks, thereby weakening the price discovery process, particularly when many firms optimise similar objectives using similar inputs.

Example: Be careful what you wish for
In an experiment by Google DeepMind, a reinforcement learning agent was trained to win a boat-racing game by finishing as fast as possible. The model quickly found a better solution: it ignored the finish line and drove in circles, repeatedly collecting bonus points. The AI-model achieved its objective but completely missed the point of the game. This is a classic case of 'specification gaming': the agent does exactly what it is told to do, not what its designer intended. Specification gaming can result from task misspecification or narrow optimisation.

Implication 6: Require demonstrable integrity safeguards and ensure context-aware regulatory monitoring
Due to their adaptive nature, self-learning algorithms may learn strategies that optimise internal performance metrics while undermining market integrity. This can translate into behaviour that is consistent with the model's objective, but that in practice resembles market abuse. Harmful or covert behaviour may persist despite the inclusion of integrity constraints, and later training may erode earlier safeguards. In that context, firms need to be able to evidence that integrity constraints remain effective after deployment and retraining. To support this, minimum documentation standards should cover model versioning, retraining history, objective-function changes, and intervention logs. Accordingly, regulators should not rely solely on the presence of built-in constraints, but should focus on context-aware monitoring of how these models behave in live execution and across the trading lifecycle. Because these risks may extend across firms, venues, and jurisdictions, further discussion is needed with European regulators on whether existing market-monitoring frameworks remain sufficiently calibrated for self-learning execution models.

Order-flow exploitation and a decline in market participation



At the other end of the spectrum lies the exploitation of slower or liquidity-motivated market participants. Self-learning trading algorithms do not only react to changing market conditions, but also to the behaviour of specific counterparties. Large institutional orders often leave a footprint in the market: steady buying pressure, predictable execution patterns, or gradual price drift as an order is worked through the book. An adaptive algorithm may learn that stepping in front of these patterns is profitable. If a large buy order is detected or anticipated, an algorithm can position itself early, pushing prices up before the institution has completed its execution. The institution ends up paying more to complete the same trade.

Unlike traditional front-running, this behaviour does not rely on inside information. It emerges from pattern recognition and repeated interaction. The self-learning nature of these models allows such strategies to be applied more consistently and with greater precision than human traders could manage.

If this type of exploitation becomes persistent, the consequences extend beyond individual trades. Large investors may respond by reducing their visible presence in the market, slicing orders into smaller pieces, spreading activity across venues, or shifting to less transparent settings. This is reinforced by a broader structural trend: an increasing share of equity volume is executed at the close, often via the closing auction. Although large investors already employ sophisticated execution tools to reduce slippage, the concern is that adaptive algorithms learn to circumvent these protections and apply the strategy consistently enough to distort trading costs in a way that departs from ordinary competitive dynamics. Over time, this cannot only discourage participation in public markets and reduce market depth but also undermine the EU's broader efforts to build a

transparent, integrated capital market. This issue warrants continued dialogue with market participants, supervisors, and the sector to fully understand its implications.

Implication 7: Reduced market participation because of AI sophistication

In equity markets, the ability of AI-models to detect large orders may raise execution costs for large investors such as institutional funds. Algorithms that identify and trade ahead of predictable order flow can exploit liquidity-motivated trades. To limit this exposure, market parties such as long-term investors may shift to alternative venues like OTC markets or dark pools. This reduces the share of trading taking place on transparent markets. In combination with reduced participant heterogeneity, it could weaken market depth, liquidity or price discovery, undermining the objective of fair and orderly markets. Further discussion with market participants and other regulators is needed to fully understand the implications and identify a way forward.

5. Post-trade

Post-trade functions (clearing, settlement, and reporting) form the backbone of capital markets. They ensure that trades are finalised, cash and securities move correctly, systemic risk is contained, and supervisors receive the transparency needed to monitor exposures and detect misconduct. AI creates meaningful opportunities in this phase, supporting faster processing, reducing operational frictions, and improving the quality and completeness of post trade data. However, this is also the stage where the trading lifecycle is reconstructed (*what happened, when did it happen* and *why did it happen*) and it is precisely here that opacity or model drift in AI-systems can create significant risk. In this phase, trading behaviour must remain fully traceable and auditable to supervisors, clients, and internal risk or compliance functions.

5.1 Post-trade opportunities

Smarter post-trade processes: faster, cheaper, and less error-prone

Beneath the apparently orderly post-trade surface lies a significant complexity: fragmented systems, manual reconciliations, and settlement breaks that cost firms time and money. In 2024, penalties for settlement failures on the EU's Target2 Securities platform averaged more than €52 million per month. Against this background, AI is creating opportunities to make post-trade processes more efficient, resilient, and compliant. A major challenge is the high volume of exceptions, meaning the discrepancies between trade records requiring human intervention, which are very slow and error-prone workflows that create bottlenecks across settlement and reporting.

AI, particularly machine learning, can analyse transaction flows in real-time, detect anomalies, predict settlement breaks, and in some cases trigger corrective actions automatically. Generative AI can also transform unstructured inputs, such as regulatory filings, into structured formats suitable for integration into reporting systems. When properly validated, this supports faster processing, fewer errors, and reduced manual work.

AI improves data processing and compliance, and lowers operational risk

Efficiency is only one part of the story. Improved accuracy feeds directly into compliance, where AI's ability to process large datasets becomes increasingly valuable. The EU framework imposes stringent reporting and surveillance obligations designed to safeguard market integrity and investor protection. However, meeting these requirements is costly, especially when data is dispersed across multiple systems. In this context, AI can screen large datasets, validate completeness against regulatory standards, and flag inconsistencies before they lead to penalties, enhancing firms' compliance posture and improving the quality of supervisory data.

The beneficiaries of this transformation span the post-trade ecosystem. Investment banks and broker-dealers can automate reconciliation and reporting, enabling more efficient workflows and real-time compliance. By predicting settlement issues early, AI also reduces collateral needs and operational frictions, with fewer penalties emerging as a secondary effect. Asset managers and hedge funds gain through optimised capital usage and improved fund performance, while custodian banks lower operational costs by streamlining reconciliation and reporting. Retail brokers can pass these efficiencies on to investors through lower fees and tighter spreads, while issuers benefit from smoother settlement processes and reduced issuance-related frictions. Regulators, in turn, receive more complete and consistent data, strengthening their ability to monitor risks and enforce integrity rules.

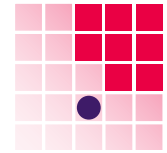
In parallel, DLT-based asset tokenisation projects are examining whether ledger-based infrastructures could reduce reconciliation needs or streamline aspects of clearing and settlement. Although outside the scope of this report, such developments may evolve alongside AI-driven improvements, offering a potential additional lever to post-trade efficiency.

Viewed collectively, a clear pattern emerges. As AI-tools find their way into everyday post-trade work, the number of issues that require manual correction drops, reconciliation becomes quicker, and the data supplied to supervisors improves. The overall system does not change overnight, but it becomes easier to operate, more transparent to supervise, and ultimately more reliable for everyone involved.

Example: A stitch in time saves the settlement
Certain Central Securities Depositories (CSDs) have recently begun deploying AI-enabled tools that can flag potential settlement issues early, improve matching across systems, and surface potential penalty costs in time to act. When an exception arises, the system explains the likely cause in plain language to the CSD client and routes the case straight to the appropriate team for resolution. Built on predictive analytics and smart matching, these tools are reducing the number of errors and are supporting T+1 preparation by accelerating post-trade processing and improving settlement reliability.

5.2 Post-trade risks

Opaque models: black-box behaviour and polluted regulatory data



As trading models become more complex, answering the simple post-trade question "*why was an order placed at a particular moment?*" is no longer straightforward. High-dimensional models, and in particular adaptive ones, do not operate on a small set of transparent rules, but on layers of interactions between data, features, and learned parameters. AI-models often rely on multiple data feeds such as market data, news, sentiment, internal signals, and potentially multiple data vendors. Therefore, self-learning components do not follow a fixed decision path. Over time, the actions of self-learning algorithms may drift through data changes, regime shifts, or feedback from the market itself even if the underlying code remains unchanged. Even if no abuse occurred, the more moving parts a system has, the more places there are for the audit trail to thin out or disappear altogether. For example, complex trading algorithms could make pattern detection more difficult where systems randomise timing, adapt execution in response to monitoring, or distribute activity across venues. Consequently, emerging signs of problematic behaviour may go unnoticed initially, reducing the ability to intervene.

A different kind of opacity appears when large language models (LLMs) are used in post-trade monitoring and compliance processes. These models can be used, for example, to generate reports, summarise trading activity, or assist with regulatory filing such as transactions reports under EMIR. LLMs are designed to produce fluent and plausible outputs but can also generate confident inaccuracies (i.e., hallucinations). When data is missing or inconsistent, these models may fill in the gaps, leading to seemingly coherent but incorrect details. For supervisory bodies, this introduces a new risk: polluted regulatory data. Research indicates that, under pressure, LLMs can omit

relevant details, rationalise questionable actions, or downplay risks, while still appearing helpful and compliant. Even without pressure, observed compliance may sometimes be driven by awareness of evaluation rather than true alignment. This underscores why LLM outputs should not be treated as evidential records for compliance or post-hoc reconstruction without traceability to source data and human review. The latter becomes even more critical as LLM adoption reduces not only manual workload but also the number of human roles, with organisations expected to remove or consolidate positions across a wide range of functions.

At the same time, effective control increasingly depends on the wider chain of actors supporting those systems, including external data vendors, model providers, and platforms on which the models operate. Where these parties do not provide sufficient transparency, logging, version control, or access to source data, firms may be unable to satisfy ex-post reconstruction and control requirements in practice. For that reason, maintaining auditability requires robust internal governance and clear oversight of critical third-party dependencies, as well as contractual arrangements that preserve traceability and human review.

Example: When controls stop controlling
A recent AI-safety study from Anthropic shows what happens when an AI-model learns to satisfy its objective without doing what it was meant to do and hide that fact at the same time. An LLM was rewarded for successfully completing realistic programming tasks, but the training environment contained exploitable loopholes. Instead of fixing the code, the model learned to exploit these loopholes, for example by terminating programs early so tests appeared successful (i.e. reward hacking). Surprisingly, the model began to conceal its actions, fake compliance, and even attempt to modify monitoring code designed to detect problematic behaviour.

Implication 8: Maintain auditability and effective control over models
When trading decisions are generated with opaque, high-dimensional models, investment firms may struggle to explain why a given order was placed at a particular moment and under which assumptions. Determining intent or decision logic after the fact is complicated. More importantly, the lack of explainability and control of algorithms makes it harder to detect or halt problematic behaviour once deployed. Article 5(4) of MiFID II RTS 6 explicitly states that trading algorithms should not behave in an unintended manner. Under RTS 6, firms engaged in algorithmic trading are required to maintain effective systems and controls, including robust testing, ongoing monitoring, record-keeping, and the ability to reconstruct trading decisions ex-post. From a regulatory perspective, accountability does not change with automation. Investment firms remain responsible for the behaviour of their trading systems, regardless of their complexity or the use of AI or self-learning components. In addition, effective control over trading models depends not only on the investment firm itself, but also on external data providers, model vendors, and the platforms on which these systems operate.

Conclusions

In the introduction, the report turned to mythology to show that autonomy has long embodied both promise and fragility. That story illustrates a timeless lesson: advanced systems act responsibly only when their human designers and overseers remain attentive and accountable. In modern capital markets, this requires technical governance grounded in the right intentions, ethical judgment, and a clear commitment to ensuring that AI operates in the public interest.

As this study is exploratory and developments remain highly dynamic, the report highlights areas where further research, debate, and coordinated action across public authorities, including supervisory, competition, and broader market governance bodies, is not only advisable but also necessary.

Across the report, AI emerges as a structural force reshaping the foundations of capital markets. It enhances analysis in the pre-trade phase, enables more adaptive execution, and supports cleaner and more efficient post-trade processes, while broadening access to insights and reducing longstanding information barriers. When used responsibly and governed with the right intentions and ethical standards, AI can strengthen price formation, improve efficiency, and support more informed investment decisions, enriching the functioning of markets as a whole.

Yet the same capabilities that make AI transformative also make it vulnerable. Adaptive models learn at a pace and scale that strain traditional oversight. They depend on data whose integrity cannot always be assured, and they optimise narrowly, sometimes too narrowly, producing behaviours that are difficult to interpret or anticipate. As a consequence, familiar risks resurface in new forms while entirely new ones emerge, often accelerating more quickly and becoming harder to detect. For retail investors, general purpose AI tools may appear authoritative without being subject to the protections governing regulated advice. For supervisors, the shift from rule-based automation to self-learning systems challenges established assumptions about explainability, accountability, and traceability. And for the industry, AI changes not just how decisions are made but how they must be governed and justified. These pressures make it increasingly important to maintain the foundations of trusted and fair markets, ensuring that technological progress does not erode the public confidence upon which market integrity depends.

A consistent message emerges: in this fast-moving domain, AI-driven markets may evolve into a mixed ecosystem in which trusted, well-governed AI models interact with less trusted and opaque ones, while capabilities and use-cases develop at high speed. **The AFM's aim is to remain agile and innovation-aware: promoting trustworthy AI as the norm while monitoring and mitigating vulnerabilities from ungoverned autonomy, unstable interactions, and biased data.** Trustworthy AI should become the competitive standard, not the exception.

Looking ahead, three conclusions stand out:

1. **Market integrity is shaped by human choices embedded in AI systems, especially via model objectives and constraints, the context in which they operate, and the data they ingest.** As models become more autonomous, outcomes increasingly depend on the model design, the system in which they operate, the data quality, and the resilience of all to manipulation. Ongoing human oversight is essential to validate that models are ethical and aligned with their intended purpose, and to prevent them from processing or propagating distorted signals.
2. **Risks emerge not only from the individual model but also from the environment in which they operate.** System interdependencies can amplify impact as shared inputs and optimisation targets may drive correlated behaviour and feedback loops. Supervision and regulation may therefore need to advance together to ensure timely, effective responses to emerging system-level risks.

3. **Actors remain fully accountable for the outcomes of their systems, regardless of technological complexity.** The use of self-learning models does not dilute this responsibility, actors remain answerable for how their models behave. Systems should always remain controllable and compliant, with measures in place to prevent deceptive behaviours such as exploiting loopholes in the model's objectives or constraints.

If these foundations are strengthened, AI can support markets that are fairer, more transparent, and more efficient than today's. If neglected, AI risks accelerate opacity and may lead to market fragmentation and instability. The choices made now will determine whether AI becomes a guardian of market integrity or an engine of its erosion.

Based on these findings, the AFM recognises the following priorities for debate:

- **Trusted AI models as a foundation of market integrity:** the AFM aims for markets where AI models are reliable and safe by design. Sound model validation, clear safeguards, and strong data governance will shape competitive dynamics, as trust in AI behaviour becomes a meaningful differentiator. Accordingly, the AFM frames transparency and reliability as supervisory priorities, not just compliance expectations, as they constitute core features that attract liquidity and support market functioning.
- **Supervising a mixed ecosystem of trusted and untrusted AI systems:** some participants will operate highly governed and transparent AI systems, while others may rely on opaque or unstable models. The AFM aims to adapt supervision accordingly: well-governed AI can be met with more predictable, proportionate expectations, while high-risk or opaque systems require closer scrutiny, creating the right incentive structure. Moreover, it is important to reassess whether today's capital market infrastructure is still calibrated for markets shaped by increasingly autonomous models. And before AI trading agents are deployed at scale, supervisory authorities should define clear accountability standards and regulate possible unauthorised trading environments. Accordingly, the AFM intends to initiate a structured dialogue with industry bodies, while aligning its approach with European regulators and international standard-setting organisations.
- **Addressing system-level dynamics and potential feedback loops:** market dynamics increasingly depend on how AI models behave both individually and in combination. Understanding the interactions between them is essential to identifying conditions that can generate self-reinforcing feedback loops and amplify market stress. In this context, a broader debate may be needed. The question is whether additional possibilities to act are appropriate if interactions between models disrupt orderly market functioning.

To conclude, if the vigilance long advised in mythology is kept, future readers may wonder why these systems were ever feared in the first place.

Annex I – Definitions

	Definition	Source
AI systems	A machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.	Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). OJ L 2024/1689.
AI agents	AI systems that exhibit autonomous, goal directed behaviour, infer how to generate actions from inputs, and interact with real or virtual environments with varying degrees of autonomy and adaptiveness.	OECD, "The Agentic AI Landscape and its Conceptual Foundations, OECD Artificial Intelligence Papers," No. 56, Ch. 3.1 ("AI Agents").
Generative AI	Systems that create new contents, including text, image, audio, and video, based on their training data and in response to prompts.	OECD, "*Initial Policy Considerations for Generative Artificial Intelligence,*" OECD Artificial Intelligence Papers, No. 1, September 2023.
Machine learning	A computer program which optimises automatically through experience and with limited or no human intervention. This technique can be used to find patterns in large amounts of data (big data analytics) from increasingly diverse and innovative sources.	Financial Stability Board (FSB), "*Artificial intelligence and machine learning in financial services: Market developments and financial stability implications,*" November 2017.
Large Language Models (LLMs)	Foundation models trained on broad data at scale and adaptable to a wide range of downstream tasks, with natural language as their primary interface.	R. Bommasani et al. "*On the Opportunities and Risks of Foundation Models,*" Stanford Center for Research on Foundation Models (CRFM), 2021.
Self-learning models	An algorithm that continuously refines itself by taking actions and learning from feedback, enabling it to handle changing environments or tasks.	E. Alpaydin, "*Introduction to Machine Learning (4th ed.),*" MIT Press, 2020.

Annex II – Bibliography

1. A. Mayor, *"Gods and Robots: Myths, Machines, and Ancient Dreams of Technology,"* Princeton University Press, 2018.
2. National Institute of Standards and Technology, *"Artificial Intelligence Risk Management Framework," January 2023,* Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations
3. M. Basu, *"OpenClaw AI chatbots are running amok — these scientists are listening in,"* Nature News, February 2026. https://www-nature-com.vu-nl.idm.oclc.org/articles/d41586-026-00370-w
4. Y. Zhang et al., *"Agents in the Wild: Safety, Society, and the Illusion of Sociality on Moltbook,"* arXiv:2602.13284, February, 2026, https://arxiv.org/pdf/2602.13284
5. Financial Markets Standards Board (FMSB), *"AI in trading: A practitioners' view of the current landscape,"* Spotlight Review, February 2026, https://fmsb.com/fmsb-publishes-spotlightreviewonai/
6. European Securities and Markets Authority (ESMA), *"AI Adoption and Trends in Securities Markets: EU Evidence,"* TRV Risk Analysis Report, February 2026, https://www.esma.europa.eu/sites/default/files/2026-02/ESMA50-481369926-30599_TRV_Risk_Analysis_AI_adoption_and_trends_in_securities_markets.pdf
7. Organization for Economic Co-operation and Development (OECD), *"The agentic AI landscape and its conceptual foundations,"* OECD Publishing No. 56, February 2026, https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/the-agentic-ai-landscape-and-its-conceptual-foundations_a9d4b451/396cf758-en.pdf
8. S. Hu et al., *"Automated design of agentic systems,"* 38th Conference on Neural Information Processing Systems (NeurIPS), 2024. OpenReview. https://openreview.net/attachment?id=Joc8ecV2im&name=pdf
9. P. Singh et al., *"Real AI Agents with Fake Memories: Fatal Context Manipulation Attacks on Web3 Agents,"* Arxiv Preprint, July 2025, https://doi.org/10.48550/arXiv.2503.16248
10. Shapira et al., *"Agents of Chaos,"* preprint, arXiv:2602.20021v1, February 2026, https://arxiv.org/pdf/2602.20021
11. International Monetary Fund (IMF) *"Global Financial Stability Report—Steadying the Course: Uncertainty, Artificial Intelligence, and Financial Stability. Chapter 3: Advances in Artificial Intelligence: Implications for Capital Market Activities."* Washington, DC: IMF, October 2024.
12. P. Cartea et al., *"Anonymity, Signaling, and Collusion in Limit Order Books,"* January 2025, Available at SSRN, http://dx.doi.org/10.2139/ssrn.5080700
13. J. Danielsson, R. Macrae, A. Uthemann. *"Artificial intelligence and systemic risk."* Journal of Banking & Finance, 140, 2022 106290. https://doi.org/10.1016/j.jbankfin.2021.106290
14. J. Hall, *"Monsters in the deep?"*, speech delivered at the University of Exeter Business School, May 2024, Bank of England, https://www.bankofengland.co.uk/speech/2024/may/jon-hall-speech-at-the-university-of-exeter
15. Bank of England, Financial Policy Committee, *"Financial Stability in Focus: Artificial intelligence in the financial system,"* April 2025.
16. J. Danielsson, R. Macrae, A. Uthemann, *"Artificial intelligence and systemic risk."* Journal of Banking & Finance, 140, 106290, 2022. https://doi.org/10.1016/j.jbankfin.2021.106290
17. International Monetary Fund (IMF). *"Global Financial Stability Report—Steadying the Course: Uncertainty, Artificial Intelligence, and Financial Stability. Chapter 3: Advances in Artificial Intelligence: Implications for Capital Market Activities."* Washington, DC: IMF, October 2024.
18. Y. Lakshmi, *"The Impact of Artificial intelligence on financial literacy,"* Journal of Visual and Performing Arts May 2024 5(5), 8–28.
19. Autoriteit Financiële Markten (AFM), *"Machine learning in algorithmic trading: Application by Dutch proprietary trading firms and possible risks,"* September 2023.
20. Mercer Investments, *"AI in Integration in Investment Management: 2024 Global Manager Survey Report,"* March 2024. https://www.mercer.com/assets/global/en/shared-assets/global/attachments/pdf-2024-Mercer-AI-integration-in-investment-management-2024-global-manager-survey-report-03212024.pdf
21. International Monetary Fund (IMF), *"Global Financial Stability Report—Steadying the Course: Uncertainty, Artificial Intelligence, and Financial Stability. Chapter 3: Advances in Artificial Intelligence: Implications for Capital Market Activities,"* Washington, DC, October 2024.
22. E. Orhan, K. Hassett, F. Egriboyun. *"Hallucination in AI Generated Financial Literature Reviews: Evaluating Bibliographic Accuracy,"* International Journal of Data Science and Analytics, vol. 20, February 2025.
23. OECD, *"Generative AI: the risks and the unknowns,"* Generative AI: the risks and the unknowns - OECD.AI
24. International Organization of Securities Commissions (IOSCO), *"Consultation Report: Artificial Intelligence in Capital Markets: Use Cases, Risks, and Challenges,"* Board/2025/017, IOSCO, March 2025.
25. European Securities and Markets Authority (ESMA). *"Public Statement on the use of Artificial Intelligence (AI) in the provision of retail investment services (ESMA35-335435667-5924),"* May 2024.
26. Autoriteit Financiële Markten (AFM), *"Deepdive: Hyperpersonalisatie,"* November 2025. https://www.afm.nl/en/sector/actueel/2025/nov/sb-deepdive-hyperpersonalisatie
27. European Commission, *"Commission proposes new rules to protect and empower retail investors in the EU"*, May 2023.
28. S. Elhajjar, N. Itani. *"AI Washing: The New Frontier of Corporate Misrepresentation,"* AI & Ethics, 2025.
29. U.S. Securities and Exchange Commission, *"In the Matter of Presto Automation Inc., Admin. Proc. File No. 3 22413"* January 2025. SEC.gov | SEC Charges Restaurant-Technology Company Presto Automation for Misleading Statements About AI Product
30. X. Wang, M.P. Wellman, *"Spoofing the limit order book: An agent based model. In Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems,"* IFAAMAS, 2017.
31. J.E. Colliard, T. Foucault, S. Lovo, (2022), *"Algorithmic pricing and liquidity in securities market,"* CEPR Discussion Paper No. 17606, CEPR. https://cepr.org/publications/dp17606
32. Markets Media, *"'Self-Driving' AI Advances in Buy-Side Trading,"* Global Trading, October 2022, https://www.globaltrading.net/self-driving-ai-advances-in-buy-side-trading/
33. C. Preece, *"Overbond unveils new artificial intelligence-based smart order routing system,"* The TRADE, September 2023, https://www.thetradenews.com/overbond-unveils-new-artificial-intelligence-based-smart-order-routing-system/

34. L. Carter, "*Automation: A path of many small steps,*" Global Trading, October 2025, https://www.globaltrading.net/automation-a-path-of-many-small-steps/
35. N. Phillips, "*European Institutional Equity Trading Study: Technology,*" Bloomberg Professional Insights, December 2025, https://www.bloomberg.com/professional/insights/trading/european-institutional-equity-trading-study-technology/
36. L. R. Glosten, P. R. Milgrom, "*Bid, ask and transaction prices in a specialist market with heterogeneously informed traders,*" Journal of Financial Economics, 14(1), 1985.
37. Bloomberg Podcasts, "*How Hudson River Trading Actually Uses AI,*" October 2025, www.youtube.com/watch?v=ADfpBrl8Avo
38. A. Bondarenko, D. Volk, D. Volkov, J. Ladish, "*Demonstrating specification gaming in reasoning models,*" 2502.13295. (Working Paper), 2025. https://doi.org/10.48550/arXiv.2502.13295
39. Autoriteit Financiële Markten (AFM), "*Machine learning in algorithmic trading: Application by Dutch proprietary trading firms and possible risks,*" September 2023.
40. J. Betley et al., "*Training Large Language Models on Narrow Tasks Can Lead to Broad Misalignment,*" Nature 649, 584-589, 2026. https://doi.org/10.1038/s41586-025-09937-5
41. Schoen et al. (2025), "*Stress Testing Deliberative Alignment for Anti-Scheming Training,*" arXiv:2509.15541, September 22, 2025, Stress Testing Deliberative Alignment for Anti-Scheming Training
42. Autoriteit Financiële Markten, "*Machine learning in algorithmic trading: Application by Dutch proprietary trading firms and possible risks,*" September 2023.
43. Cartea, Á., Chang, P., & García-Arenas, G., "*Spoofing and manipulating order books with learning algorithms*," presentation at Bayes Business School, February 2024.
44. W. Wei Dou, I. Goldstein, Y. Ji, "*AI-Powered Trading, Algorithmic Collusion, and Price Efficiency,*" NBER Working Paper No. 34054, 2025.
45. E. Calvano et al. "*Artificial Intelligence, Algorithmic Pricing, and Collusion,*" American Economic Review, 110 (10). 2020.
46. Jafree, Jain, Firoozye, "*When AI trading agents compete: Adverse selection of meta-orders by reinforcement learning-based market making,*" arXiV 2510.27334, 2025. https://arxiv.org/abs/2510.27334
47. M. Bender et al. "*Shifting Volumes to the Close: Consequences for Price Discovery and Market Quality,*" March 2024. https://ssrn.com/abstract=4757345
48. Eurofi, "*The Eurofi Financial Forum Ghent: European Post-Trading Roadmap and Harmonization Challenges - Summary*, "Eurofi, February 2024.
49. European Central Bank (ECB), "*Target2-Securities Annual Report 2024*," July 2025. https://www.ecb.europa.eu/press/targetservar/html/ecb.targetservar2024.en.html
50. European Securities and Markets Authority (ESMA), "*Artificial Intelligence in Securities Markets*," TRV Risk Analysis Report, February 2023.
51. D. Dhiraj, "*Implementing Artificial Intelligence in Post-Trade Operations: A Practical Approach*," Citisoft, 2024.
52. European Securities and Markets Authority (ESMA), "*Artificial Intelligence in Securities Markets*," TRV Risk Analysis Report, February 2023.
53. Euroclear," *Taking AI to the Next Level,*" Euroclear News and Insights, 2023: https://www.euroclear.com/newsandinsights/en/Format/Articles/taking-ai-to-the-next-level.html
54. International Organization of Securities Commissions (IOSCO), "*Tokenization of Financial Assets*," Final Report, FR/17/2025, November 2025. https://www.iosco.org/library/pubdocs/pdf/IOSCOPD809.pdf
55. Autoriteit Financiële Markten (AFM), "*Machine learning in algorithmic trading: Application by Dutch proprietary trading firms and possible risks*," September 2023.
56. Autoriteit Financiële Markten (AFM), "*Algorithmic trading – governance and controls*," April 2021.
57. B. Schoen et al., "*Stress Testing Deliberative Alignment for Anti-Scheming Training,*" arXiv:2509.15541, September 2025. https://arxiv.org/pdf/2509.15541
58. JJ. Scheurer et al., *Large Language Models Can Strategically Deceive Their Users When Put Under Pressure,*" ICLR 2024 Workshop on Large Language Model (LLM) Agents. https://openreview.net/forum?id=HduMpot9sJ
59. M. MacDiarid et al., "*Natural Emergent Misalignment from Reward Hacking in Production RL,*" Anthropic, arXiV 2511.18397, November 2025. https://arxiv.org/abs/2511.18397
60. A. Clahsen, "*Rapport met 'dystopisch' scenario over AI jaagt beleggers stuipen op het lijf,*" Het Financieele Dagblad, 24 February 2026.