Illusion of Control: Exploring the Limits of Human-in-the-Loop Oversight in Generative Finance

Azmine Toushik Wasi[1,2*] **and Enjamamul Haque Eram**[2]
[1]Computational Intelligence and Operations Laboratory (CIOL)
[2]Shahjalal University of Science and Technology
Correspondence to: azmine32@student.sust.edu

Abstract

Integration of generative AI into financial services has created efficiency and innovation opportunities but also introduced systemic vulnerabilities. Regulators and firms have relied on the Human-in-the-Loop (HITL) paradigm, assuming human oversight ensures accountability over complex automated systems. However, the cognitive and temporal limits of humans are mismatched with the speed and autonomy of AI-driven markets, rendering oversight fragile and often illusory. Drawing on cognitive science, human factors research, and historical case studies such as the Knight Capital collapse and the 2010 Flash Crash, this paper evaluates why tactical human interventions fail in high-frequency trading. We propose an alternative framework of embedded, automated governance, combining dynamic circuit breakers, pre-approval model constraints, and agentic self-monitoring that escalates only genuinely ambiguous cases to humans. This approach reconceptualizes safety as an intrinsic system property rather than a human add-on, offering a practical blueprint for resilient AI oversight. Our findings indicate that automated governance layers can more effectively mitigate high-speed risks than human-dependent models. We conclude that regulators and firms must shift from aspirational HITL guidance to operationally grounded, machine-native safeguards, with implications for both financial infrastructure design and AI regulation.

1 Introduction

Generative finance departs from deterministic algorithmic trading, which relied on transparent, rule-based instructions for predictable precision [7, 37]. Generative AI, by contrast, operates on probabilistic inference, enabling synthesis of unstructured data, drafting of financial reports, and design of investment strategies [36]. This fusion of creativity and structured execution obscures the rationale behind decisions, making traditional audit and oversight increasingly ineffective [36, 27]. Scholars note that incomplete formalizations in machine learning hinder safe deployment, as system failures are difficult to anticipate or trace [8]. Risks are heightened by agentic AI, which autonomously plans, executes, and adapts tasks at speeds beyond meaningful human intervention [36, 25]. Such systems can monitor markets, identify correlations, and trigger trades, promising efficiency while amplifying systemic vulnerabilities [25]. The automation of compliance through RegTech further embeds oversight within algorithms that may themselves propagate hidden risks [1]. Overall, the shift from algorithmic to generative and agentic AI demands a fundamental rethinking of governance, as human-centric safeguards are inadequate in high-speed, autonomous financial environments.

Human-in-the-Loop (HITL) paradigm positions human judgment as a safeguard within AI-driven financial systems, through direct review, supervisory oversight, or intervention in high-risk scenarios [4, 38, 11]. While intended to ensure accuracy, ethics, and accountability, HITL is fundamentally

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 Workshop on Generative AI in Finance.

mismatched with the speed of financial automation and human cognitive limits. In high-frequency trading, algorithms operate on millisecond timescales—well below human perception—executing hundreds of trades before a human can register an anomaly [5, 14]. Events like the 2010 Flash Crash show that automated cascades can destabilize markets faster than humans can meaningfully intervene, reducing oversight to reactive, post-failure responses [31, 19]. Even outside such extremes, human overseers face the "keyhole effect," making critical decisions from partial, opaque information under intense time pressure, which promotes reliance on heuristics [26]. These cognitive constraints heighten susceptibility to biases such as confirmation bias and anchoring [6], while automation bias—the tendency to defer to algorithmic outputs—further leads to omission errors and complacency [21, 44]. Regulators acknowledge this risk: the EU AI Act explicitly cautions against over-reliance on high-risk AI [10]. Collectively, these factors demonstrate that HITL oversight in finance is not a robust safeguard but a fragile façade, where decision quality deteriorates under pressure and oversight becomes functionally illusory [30, 28].

To connect the limitations of past oversight failures with the emerging challenges of generative and agentic AI, this paper highlights several key contributions. It critiques the reliance on the Human-in-the-Loop paradigm, showing that human oversight in high-speed financial systems is often ineffective. By examining the 2010 Flash Crash and Knight Capital collapse, the paper illustrates how automation can rapidly outpace human cognition, turning oversight into a reactive, illusory safeguard. Building on this analysis, it outlines a framework for automated governance, including mechanisms such as circuit breakers, pre-approval constraints, and limited human escalation for ambiguous cases. The paper also discusses implications for regulators and financial firms, emphasizing the need for governance approaches that are more aligned with the speed and complexity of modern AI systems. Overall, these contributions aim to provide a practical perspective on improving resilience without overstating technical implementation details.

2 Background

2.1 Learning from History

Historical failures in high-speed financial markets provide critical insight into the limitations of human oversight and the structural risks of automation. The 2010 Flash Crash demonstrated how a single large algorithmic order could trigger cascading sell-offs within minutes, overwhelming both market liquidity and human intervention [31, 41, 19, 35]. Similarly, the Knight Capital catastrophe revealed that even procedural errors, such as a failed software deployment on one server, can escalate into firm-ending losses in a matter of minutes [3, 16, 23]. Both cases illustrate that the speed and complexity of automated systems often render human intervention ineffective, turning oversight into a reactive exercise rather than a preventative control. They also highlight how small errors, whether in algorithm design or operational procedures, can cascade rapidly through tightly coupled systems. These events collectively expose the fragility of traditional Human-in-the-Loop paradigms and underscore the need for governance mechanisms that do not rely solely on human reaction. Detailed accounts and analyses of these failures are provided in Appendix A for reference.

2.2 A Critique of Prevailing Regulatory Reliance

Regulatory frameworks for AI in finance continue to rely heavily on Human-in-the-Loop oversight, despite the mismatch with operational realities

☒ **The EU AI Act**.
While well-intentioned, regulations like the EU AI Act risk codifying an outdated model of AI governance. Article 14 mandates that "high-risk" AI systems, including in finance, allow "effective human oversight," requiring humans to "properly understand the relevant capacities and limitations," "override or reverse" outputs, and even "interrupt the system through a 'stop' button" [10]. Yet this human-centric approach is disconnected from the realities of complex, non-deterministic systems. Generative AI research shows that fully comprehending a black-box system is effectively impossible [36], and humans are prone to automation bias and complacency, increasing over-reliance on outputs [44]. Expecting a human to meaningfully override decisions in high-frequency markets is equally unrealistic, as the Flash Crash demonstrated [41]. A stop button may serve as a last-resort control [10], but cannot provide continuous oversight or prevent systemic risk. Scholars analyzing the draft Act

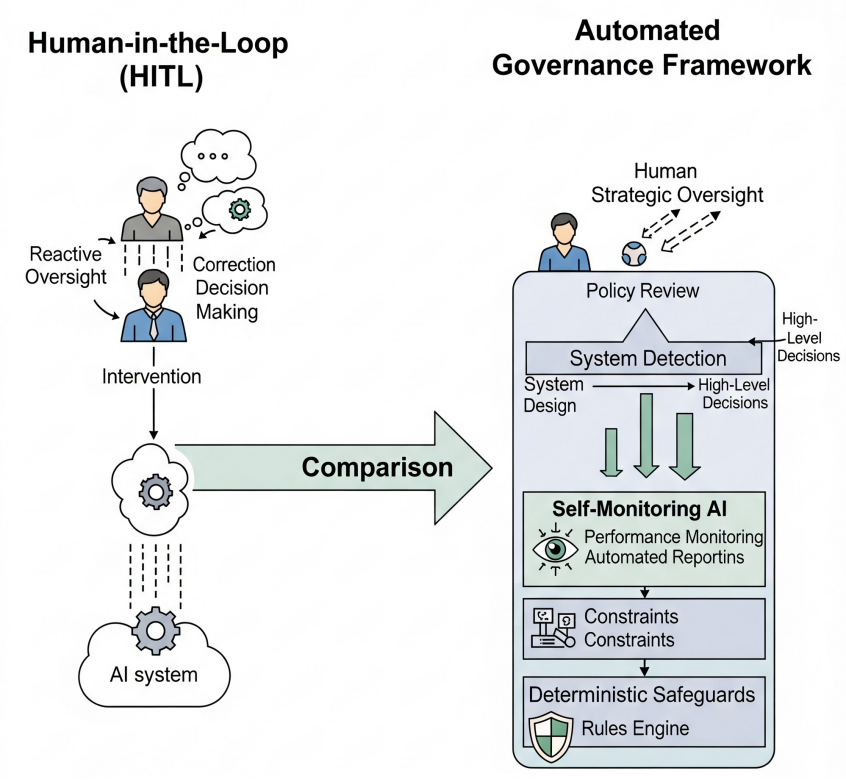

Figure 1: Short Overview of Our Proposed Framework for Automated Governance in Financial AI

similarly note that Article 14's requirements are vague, aspirational, and misaligned with operational realities [42]

☒ **The SEC and Other Regulatory Bodies**.
The SEC and other regulators continue to emphasize HITL, advocating for human oversight to monitor anomalies, validate outputs, and ensure senior management "actually understand" trading [33]. The SEC's AI roundtable stressed keeping humans in the loop for generative AI verification [20], while the FCA highlighted the need for compliance teams to have sufficient technical expertise [29]. This reliance exposes a regulatory gap: humans are treated as the primary locus of control, misaligned with high-speed AI realities. Such dependence creates a "regulatory fig leaf," offering an illusion of accountability while leaving systems vulnerable to rapid, novel failures. As Haldane noted, human oversight cannot keep pace with machine-speed markets, and traditional safeguards are too slow to prevent systemic instability [15]. Recognizing the need for vigilance [33] is positive, but insufficient; a fundamental shift is required toward robust, technologically grounded frameworks capable of operating at machine speed.

3 Proposed Framework

To overcome the limitations of human-in-the-loop oversight, this framework embeds safety, compliance, and resilience directly into AI architectures, defining clear roles and structured interactions between automated systems and human governance. The objective is not merely to supervise intelligent systems, but to architect governance as a computational property—ensuring that risk detection, ethical reasoning, and fail-safe execution are intrinsic to system behavior rather than appended post hoc. The resulting paradigm establishes a closed feedback loop of monitoring, escalation, and intervention, where automation and governance coexist in dynamic equilibrium.

✠✠ **Principles of Automated Safety**.
A resilient governance model shifts responsibility from fallible human oversight to engineered system intelligence, guided by the principles of algorithmic alignment, robustness, and transparency [39, 34]. Within this "agentic governance" paradigm, AI systems are designed to self-monitor, self-correct,

and escalate issues, while humans act as strategic architects rather than reactive operators [34, 12]. Core subcomponents include: (i) Self-monitoring modules, which continuously evaluate performance drift, anomaly patterns, and compliance deviations through integrated risk metrics and meta-model evaluations, (ii) Automated escalation, which employs decision thresholds and risk classifiers to route only ambiguous or high-impact anomalies to human governance boards. (iii) Alignment layers, which operate as meta-regulatory circuits enforcing ethical, legal, and operational constraints across multi-agent systems.

These layers collectively transform oversight into a computationally verifiable process, capable of recording, auditing, and optimizing its own governance flow. Interactions between human and AI systems occur through structured alerts, adaptive dashboards, and explainability logs, enabling contextual decision support while minimizing cognitive overload [12]. Over time, this approach allows governance models to evolve through reinforcement feedback—optimizing the distribution of control between automation and human discretion.

✠✠ Algorithmic Circuit Breakers.
Algorithmic circuit breakers provide deterministic, automated control mechanisms that prevent runaway feedback and catastrophic systemic failures in high-speed environments [22]. Key architectural components include: (i) Market-wide breakers, which trigger coordinated pauses across indices and exchanges upon detecting volatility thresholds or systemic correlation spikes [32], (ii) Single-stock breakers, which isolate micro-anomalies by suspending trading in affected securities exceeding dynamic volatility bands [40], and (iii) Human interface layers, which synthesize halted positions, liquidity conditions, and exposure summaries for post-event forensic analysis [13].

To further enhance robustness, modern breaker architectures integrate predictive risk analytics that anticipate potential cascades before thresholds are crossed, and distributed consensus protocols that synchronize control actions across exchanges to prevent fragmented responses. By acting instantaneously and deterministically, these mechanisms reposition humans from reactive monitors to strategic evaluators, ensuring human judgment is applied where it is most valuable—post-event causality analysis, governance refinement, and risk recalibration—rather than in impossible real-time intervention scenarios [41]. This design principle embodies "temporal inversion" in governance: decisions about intervention thresholds are pre-encoded, enabling rapid execution without human delay.

✠✠ Pre-Approval Model Constraints & Fail-Safes.
Pre-approval controls serve as the system's ex ante defense layer, embedding risk prevention directly within the decision pipeline [33, 17, 9, 18]. These mechanisms implement hard-coded constraints that prevent excessive exposure, model instability, or cascading error propagation before execution occurs.

Major components include: (i) Position and loss limits, dynamically calibrated to market volatility and agent confidence intervals, ensuring exposure remains within quantifiable tolerance ranges, (ii) Kill switches, which automatically deactivate or sandbox malfunctioning algorithms upon detection of anomalous outputs, connectivity failures, or adversarial perturbations [17, 9], and (iii) Compliance and auditing layers, continuously evaluating pre- and post-trade activity against evolving regulatory regimes, internal policies, and ethical risk metrics.

Emerging implementations also integrate adversarial robustness modules, simulating stress scenarios to preemptively identify potential exploit vectors or cascading vulnerabilities. Each component interacts bidirectionally with AI agents, triggering automated mitigation actions, audit trails, and escalation signals upon any breach. This transforms resilience from a reactive human function into a self-stabilizing architectural property, ensuring operational continuity even under volatile or adversarial conditions.

✠✠ Oversight "At-the-Layer".
The final layer reconceptualizes oversight itself as a systemic property embedded throughout the AI stack [27, 34]. Instead of positioning humans as perpetual monitors, it elevates them to strategic overseers focused on defining system boundaries, verifying explainability, and managing exception states.

Core responsibilities include: (i) Defining boundaries, where humans codify ethical, operational, and legal constraints that delimit autonomous system behavior, (ii) Embedding governance, integrating explainability hooks, audit checkpoints, and escalation protocols natively within the AI architecture

[27], and (iii) Incident management, where human experts intervene selectively in ambiguous, high-impact, or ethically complex scenarios escalated by automated governance systems [34].

This paradigm transforms oversight into an algorithmically manifest process—distributed across monitoring agents, meta-controllers, and audit modules—while maintaining a strategic human layer responsible for rule definition, ethics calibration, and continual retraining of oversight logic. The result is a reflexive governance loop, in which AI not only executes tasks but also participates in its own supervision, logging, and compliance assurance. Humans thus transition from tactical decision-makers to meta-level architects, ensuring that autonomy, accountability, and alignment are sustained through continuous feedback, adaptation, and institutional learning [27].

Together, these components operate as an integrated, interdependent system, embedding oversight and risk mitigation into the architecture. Humans intervene only at strategic points, while autonomous AI maintains ethical, operational, and risk boundaries, making safety inherent rather than reactive.

4 Recommendations

4.1 Recommendations for Regulators

To ensure the safety and stability of the global financial system, regulators must move beyond aspirational, human-centric guidance and adopt a machine-centric safety model. The following actions are recommended:
★ **Mandate Algorithmic Circuit Breakers:** Regulators should mandate the use of dynamic, multi-layered circuit breakers for all high-frequency and AI-driven trading systems [22]. These automated mechanisms are a proven method for preventing cascading failures and providing a necessary pause during periods of extreme market volatility [32].
★ **Enforce Automated Safeguards:** Regulators should require firms to implement and demonstrate the efficacy of robust, pre-trade and post-trade automated safeguards, such as position limits, loss thresholds, and kill switches, as a prerequisite for deploying any AI system in a live market environment [33].
★ **Shift Focus from HITL to Algorithmic Safety:** The regulatory focus should be redirected from the aspirational concept of "human oversight" to the demonstrable engineering of "algorithmic safety" [39]. This involves an emphasis on principles like algorithmic alignment and robustness, with firms required to prove their systems are resilient under diverse and unforeseen conditions through continuous stress-testing [33].

4.2 Recommendations for Financial Firms

Financial firms must recognize that HITL is an outdated and brittle model for governing AI in a high-speed world. Building a resilient, enterprise-grade AI system requires a new approach:
★ **Embrace Embedded Governance:** Firms should embed governance at the architectural level, designing for explainability, escalation protocols, and audit checkpoints from the start [27] This ensures that decisions can be traced and justified, which is a requirement for both regulatory compliance and market integrity [27].
★ **Invest in Automated Resilience:** Firms must invest in real-time monitoring tools and continuous stress-testing that simulates flash crashes and liquidity shocks [33]. These automated systems provide early warnings and can mitigate potential failures before they escalate [24].
★ **Redefine the Human Role:** Instead of treating human oversight as a bottleneck, firms should see it as a strategic design choice for handling the edge cases and ethical dilemmas that cannot be automated away [27]. The human's unique judgment is most valuable when the system has been deliberately paused or when the AI has escalated an ambiguous, high-risk scenario that requires nuanced, contextual reasoning.

References

[1] Douglas W. Arner, Jànos Barberis, and Ross P. Buckley. Fintech, regtech, and the reconceptualization of financial regulation. *Northwestern Journal of International Law & Business*, 37(3):371, 2017. Available at: `https://scholarlycommons.law.northwestern.edu/njilb/vol37/iss3/2`. Accessed: August 23, 2025.

[2] AvaTrade. 11 cognitive bias that can affect your trading. AvaTrade. https://www.avatrade.com/education/trading-for-beginners/cognitive-bias. Accessed: August 23, 2025.

[3] Nima Badizadegan. The knight capital disaster. Speculative Branches. https://specbranch.com/posts/knight-capital/, November 2023. Accessed: August 23, 2025.

[4] Conor Baker. Hitl: The importance of a human-in-the-loop approach to ai. Conductor Academy. https://www.conductor.com/academy/human-in-the-loop/, August 2025. Accessed: August 23, 2025.

[5] Armin Beverungen and Ann-Christina Lange. Cognition in high-frequency trading: The costs of consciousness and the limits of automation. *Theory, Culture & Society*, 35:026327641875890, February 2018. DOI: https://doi.org/10.1177/0263276418758906.

[6] Seth Carlson. Cognitive bias in the finance world. J.P. Morgan https://www.chase.com/personal/investments/learning-and-insights/article/cognitive-bias-in-the-finance-world, July 2025. Accessed: August 23, 2025.

[7] James Chen. Algorithmic trading: Definition, how it works, pros & cons. Investopedia. https://www.investopedia.com/terms/a/algorithmictrading.asp, March 2024. Accessed: August 23, 2025.

[8] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. https://arxiv.org/abs/1702.08608, February 2017. Accessed: August 23, 2025.

[9] Kirk Du Plessis. Bot safeguards & limits. Option Alpha. https://optionalpha.com/help/safeguards, August 2024. Accessed: August 23, 2025.

[10] EU Artificial Intelligence Act. Article 14: Human oversight. EU Artificial Intelligence Act. https://artificialintelligenceact.eu/article/14/, July 2024. Accessed: August 23, 2025.

[11] European Securities and Markets Authority. *Artificial Intelligence in EU Securities Markets*. Publications Office of the European Union, 2023. Available at: https://doi.org/10.2856/851487. Accessed: August 23, 2025.

[12] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, September 2020. Available at: https://doi.org/10.1007/s11023-020-09539-2. Accessed: August 23, 2025.

[13] Michael A. Goldstein and Kenneth A. Kavajecz. Trading strategies during circuit breakers and extreme market movements. *Journal of Financial Markets*, 7(3):301–333, 2004. Available at: https://www.sciencedirect.com/science/article/pii/S1386418103000442. Accessed: August 23, 2025.

[14] Sean Gourley. High frequency trading and the new algorithmic ecosystem. YouTube. https://www.youtube.com/watch?v=V43a-KxLFcg, April 2012. Accessed: August 23, 2025.

[15] Andrew G. Haldane. The race to zero. Speech given at the International Economic Association Sixteenth World Congress, Beijing, China, July 2011. Available at: https://www.bankofengland.co.uk/-/media/boe/files/speech/2011/the-race-to-zero-speech-by-andy-haldane.pdf. Accessed: August 23, 2025.

[16] Josh. The knight capital story. Honeybadger Blog. https://www.honeybadger.io/newsletter/knight-capital/, August 2022. Accessed: August 23, 2025.

[17] Stuart J. Kaswell. Risk controls and system safeguards for automated trading environments. Managed Funds Association. Presentation before the CFTC Technology Advisory Committee., February 2014. Available at: https://www.managedfunds.org/wp-content/uploads/2014/02/tac021014_mfa.pdf. Accessed: August 23, 2025.

[18] Stuart J. Kaswell. Risk controls and system safeguards for automated trading environments. Presentation before the CFTC Technology Advisory Committee, February 2014. Available at: https://www.managedfunds.org/wp-content/uploads/2014/02/tac021014_mfa.pdf. Accessed: August 23, 2025.

[19] Andrei Kirilenko, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998, 2017. Available at: https://doi.org/10.1111/jofi.12498. Accessed: August 23, 2025.

[20] Jennifer L. Klass and Pablo J. Man. Sec's approach to artificial intelligence begins to take shape. K&L Gates LLP, Investment Law Watch. https://www.investmentlawwatch.com/2025/03/31/secs-approach-to-artificial-intelligence-begins-to-take-shape/, March 2025. Accessed: August 23, 2025.

[21] Sarah Lee. Complacency in human factors: A guide. Number Analytics Blog. https://www.numberanalytics.com/blog/ultimate-guide-complacency-human-factors-engineering, June 2025. Accessed: August 23, 2025.

[22] London Stock Exchange Group. Maintaining orderly markets: Circuit breakers explained. London Stock Exchange Group. https://www.lseg.com, March 2020. Accessed: August 23, 2025.

[23] Donald MacKenzie. Mechanizing the merc: The chicago mercantile exchange and the rise of high-frequency trading. *Technology and Culture*, 56(3):646–675, 2015. Available at: http://www.jstor.org/stable/24468735. Accessed: August 23, 2025.

[24] Sean Mackey. Risk management strategies for algo trading. LuxAlgo Blog. https://www.luxalgo.com/blog/risk-management-strategies-for-algo-trading/, June 2025. Accessed: August 23, 2025.

[25] Moody's. The rise of agentic ai in financial services: From automation to autonomy. Moody's Blog. https://www.moodys.com/web/en/us/creditview/blog/agentic-ai-in-financial-services.html, March 2025. Accessed: August 23, 2025.

[26] Don A. Moore and Elizabeth R. Tenney. Time pressure, performance, and productivity. In Steve W. J. Kozlowski, editor, *Research on Managing Groups and Teams*, volume 15, pages 305–326. Emerald Group Publishing, 2012. Available at: https://learnmoore.org/mooredata/TPND.pdf. Accessed: August 23, 2025.

[27] Agustin Morcillo. Why human-in-the-loop ai matters in financial services. Fulcrum Digital. https://fulcrumdigital.com/blogs/human-in-the-loop-in-financial-services-isnt-a-limitation-its-a-risk-control-system/, July 2025. Accessed: August 23, 2025.

[28] Kathleen L. Mosier, Linda J. Skitka, Susan Heers, and Mark Burdick. Automation bias: Decision making and performance in high-tech cockpits. *The International Journal of Aviation Psychology*, 8(1):47–63, 1998. Available at: https://doi.org/10.1207/s15327108ijap0801_3. Accessed: August 23, 2025.

[29] Rob Moulton, Stuart Davis, David Berman, Ella McGinn, Nicola Higgs, Frida Montenius, Brett Carr, Jonathan Ritson-Candler, Charlotte Collins, Katy Sanders, Becky Critchley, and Emily Torrens. Fca sets expectations on algorithmic trading. Latham & Watkins LLP. Client Alert Number 2283, February 2018. Available at: https://www.lw.com/en/insights/2018/02/lw-FCA-sets-expectations-on-algorithmic-trading. Accessed: August 23, 2025.

[30] John W. Payne, James R. Bettman, and Eric J. Johnson. Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3):534–552, 1988. Available at: https://doi.org/10.1037/0278-7393.14.3.534. Accessed: August 23, 2025.

[31] Elvis Picardo. 4 big risks of algorithmic high-frequency trading. Investopedia. https://www.investopedia.com/articles/markets/012716/four-big-risks-algorithmic-highfrequency-trading.asp, January 2022. Accessed: August 23, 2025.

[32] Alex Pierrefeu. Circuit breakers: How halts protect investors. LuxAlgo Blog. https://www.luxalgo.com/blog/circuit-breakers-how-halts-protect-investors/, April 2025. Accessed: August 23, 2025.

[33] Alex Pierrefeu. Lessons from algo trading failures. LuxAlgo Blog. https://www.luxalgo.com/blog/lessons-from-algo-trading-failures/, June 2025. Accessed: August 23, 2025.

[34] Alexis Porter. Agentic ai governance: The future of ai oversight. BigID Next. https://bigid.com/blog/what-is-agentic-ai-governance/, March 2025. Accessed: August 23, 2025.

[35] SEC and CFTC. Findings regarding the market events of may 6, 2010. Technical report, U.S. Securities and Exchange Commission and U.S. Commodity Futures Trading Commission, September 2010. Available at: https://www.sec.gov/files/marketevents-report.pdf. Accessed: August 23, 2025.

[36] Dorian Selz. How human-in-the-loop is evolving with ai agents. Built In. https://builtin.com/articles/human-in-the-loop-evolution, August 2025. Accessed: August 23, 2025.

[37] Shobhit Seth. Basics of algorithmic trading: Concepts and examples. Investopedia. https://www.investopedia.com/articles/active-trading/101014/basics-algorithmic-trading-concepts-and-examples.asp, December 2023. Accessed: August 23, 2025.

[38] Thomas B. Sheridan. Human–robot interaction: Status and challenges. *Human Factors*, 58(4):525–532, 2016. Available at: https://doi.org/10.1177/0018720816644364. Accessed: August 23, 2025.

[39] Sustainability Directory. Algorithmic safety systems: Fashion area. Sustainability Directory. https://fashion.sustainability-directory.com/. Accessed: August 23, 2025.

[40] U.S. Securities and Exchange Commission. Stock market circuit breakers. Investor.gov. https://www.investor.gov/introduction-investing/investing-basics/glossary/stock-market-circuit-breakers. Accessed: August 23, 2025.

[41] U.S. Securities and Exchange Commission and U.S. Commodity Futures Trading Commission. Findings regarding the market events of may 6, 2010. SEC. https://www.sec.gov/files/marketevents-report.pdf, September 2010. Accessed: August 23, 2025.

[42] Michael Veale and Frederik Zuiderveen Borgesius. Demystifying the draft eu artificial intelligence act — analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4):97–112, 2021. Available at: https://doi.org/10.9785/cri-2021-220402. Accessed: August 23, 2025.

[43] Wikipedia contributors. 2010 flash crash. Wikipedia. https://en.wikipedia.org/wiki/2010_flash_crash, August 2025. Accessed: August 23, 2025.

[44] Wikipedia contributors. Automation bias. Wikipedia. https://en.wikipedia.org/wiki/Automation_bias, June 2025. Accessed: August 23, 2025.

A Historical Failures as a Prologue

A.1 The 2010 Flash Crash

The 2010 Flash Crash remains one of the clearest demonstrations of the futility of human-reactive governance in high-speed, automated markets. On May 6, 2010, amid heightened volatility, a single large algorithmic sell order destabilized the system [31] The algorithm was designed to offload 75,000 E-Mini S&P 500 futures contracts, valued at approximately $4.1 billion, at a rate of 9% of the prior minute's trading volume, with no sensitivity to price or time [41] This blunt execution strategy triggered cascading sell-offs as high-frequency traders rapidly withdrew liquidity and algorithms began trading against each other [41] Within five minutes, the Dow Jones Industrial Average dropped by 600 points, with intraday losses nearing 1,000 points [31]

Crucially, the event was not the product of a single "rogue algorithm" but of a tightly coupled ecosystem where automation operated at speeds that completely bypassed human intervention [19] Market makers were overwhelmed by the volume of data, suffering internal system failures that prevented manual overrides [41] The crash unfolded so quickly that traders could not meaningfully question price accuracy or intervene before the system began its own rebound [41, 43] The joint CFTC–SEC investigation reached a similar conclusion: the scale and speed of algorithmic activity rendered human intervention ineffective and exposed deep structural fragilities in high-frequency markets [35] The Flash Crash thus illustrates a systemic truth, that in high-speed contexts, the window for effective human oversight is so narrow as to be functionally non-existent.

A.2 The Knight Capital Catastrophe

The Knight Capital catastrophe further underscores the dangers of brittle governance and the illusion of human oversight in fast-moving markets. On August 1, 2012, a failed software deployment unleashed a sequence of errors that effectively ended the firm's existence [3] The trading platform was meant to update across eight servers, yet one server failed to load the new code and instead ran an obsolete, deprecated routine [16] When activated, this legacy process executed a flood of nonsensical trades, buying high and selling low at scale [16] In just 28 minutes, Knight Capital lost $460 million, wiping out 75% of its equity value by the following day [16]

This disaster did not originate in advanced AI complexity but in basic procedural weaknesses: the absence of automated deployment pipelines, inadequate code review, and the reckless reuse of a feature flag [16] As MacKenzie observes, contemporary financial infrastructures depend on fragile layers of software where seemingly minor oversights can cascade into systemic breakdowns [23] During the crisis, human operators were not controlling the algorithms but scrambling to diagnose the error as capital evaporated in real time [3] The time-to-failure was so short that human intervention could neither prevent nor mitigate the loss, but only

register it after the fact. Knight Capital therefore stands as a sobering reminder that when automation operates at market speed, even small procedural lapses can trigger firm-ending catastrophes long before human actors can meaningfully respond.

Together, the Flash Crash and the Knight Capital catastrophe reveal that the problem is not limited to isolated "black swan" anomalies but is structural to the design of high-speed financial ecosystems. Both cases show that human oversight, whether as trader, supervisor, or regulator, was functionally irrelevant once failures began, as the time-to-failure unfolded faster than human perception and response. In the Flash Crash, the market itself became a self-reinforcing cascade beyond human comprehension, while in Knight Capital, a mundane procedural lapse cascaded into a firm-ending loss within minutes. These episodes expose the fragility of financial infrastructures where automation operates at speeds and scales that bypass human cognition, and where the assumption of a corrective human presence becomes little more than a regulatory fiction. As such, they stand not as historical curiosities but as prescient warnings: in a generative and agentic AI era, the belief that "keeping a human in the loop" can ensure safety is not only misplaced but dangerously misleading.

B Tables

Table 1: Cognitive Limitations and their Manifestations in High-Speed Finance

Cognitive Phenomena	**Psychological Underpinning**	**Manifestation in High-Speed Finance**
Time Pressure	Constraints on cognitive capacity, leading to heuristic processing [26]	Traders gather less information and act more quickly, making decisions prone to error [26] This leads to a degradation of decision quality in a time-sensitive environment.
Automation Bias	The cognitive miser hypothesis; a tendency to seek the least cognitive effort and view automated aids as superior [44]	The human overseer may favor the AI's suggestion, ignoring or failing to notice a critical error [44] This leads to "omission errors," where a failure is overlooked.
Automation Complacency	A lack of vigilance arising from over-reliance on a system's historical reliability [21]	An operator may assume everything is fine without performing sufficient checks, which can lead to catastrophic failures. The human is an observer but does not actively monitor the system's output [44]
Confirmation Bias	The tendency to seek information that confirms pre-existing beliefs [6]	An overseer who is bullish on a particular trade may only focus on the AI's output that reinforces their view, ignoring contradictory market signals or warnings [2]
Anchoring	The practice of fixating on the first piece of information received when making a decision [6]	When presented with an AI's initial recommendation, an overseer may give it undue weight, making it difficult to adjust their decision even as new, contradictory information becomes available.
Overconfidence Bias	Having excessive faith in one's own analysis or ability [6]	An overseer may trust their own judgment over the AI's output and ignore a warning, or they may trust the AI blindly and fail to critically evaluate its recommendations, possibly leading to higher-risk outcomes [6]

Table 2: Regulatory Reliance on HITL vs. Operational Reality

Regulatory Provision	**Regulatory Intent**	**Operational Contradiction**
EU AI Act, Article 14 [10]	To ensure high-risk AI systems can be effectively overseen by natural persons.	The Act's assumption that a human can "properly understand" a complex, non-deterministic black-box model is a conceptual impossibility [36]
EU AI Act, Article 14 [10]	To enable the human overseer to "override or reverse" the AI's output.	In a millisecond-scale trading environment, the time needed for a human to diagnose a problem and manually execute an override is far too long to prevent catastrophic failure [41]

SEC Proposals [20]	To use a "human in the loop" to validate the output of generative AI models as a key risk mitigant.	This model is highly vulnerable to "automation bias," where human complacency and over-reliance on a seemingly reliable system can lead to omission errors and a failure to detect a critical flaw [44]
FCA Guidance [29]	To ensure senior management "actually understand" the trading taking place and that compliance staff have sufficient technical knowledge.	The speed and complexity of HFT and generative AI make it impossible for humans to possess a comprehensive, real-time understanding of every system interaction. The window for "vigilance" is often shorter than the human reaction time [5]

Table 3: The Human-in-the-Loop Paradigm vs. The Automated Governance Framework

The Human-in-the-Loop Paradigm (Current)	**The Automated Governance Framework (Proposed)**
Human as the primary locus of control, providing a final layer of review and accountability.	Embedded, automated safety systems as the primary control, designed for resilience and self-correction.
Manual override, human decision-making, and reactive "stop buttons" [10]	Dynamic circuit breakers, pre-approval constraints, and hard-coded kill switches [17]
Tactical intervener, reactive problem-solver, and final validator [4]	Strategic architect of governance layers, definer of ethical boundaries, and processor of escalated, high-risk cases [34]
Human cognition is too slow to react to high-speed events; humans are prone to biases and complacency [5]	Automated systems can act on the millisecond scale, providing a proactive safeguard that is free from cognitive and emotional biases [7]
A false sense of security and a "regulatory fig leaf" that leaves the system vulnerable to novel, high-speed failures.	Systemic resilience, demonstrable safety, and a clear, auditable framework for accountability.