Systemic Misuse of Artificial Intelligence in Financial Systems: Threat Models, Empirical Exploits, and Misuse-Aware Detection

Usha Ratnam Jammula[1], Srilakshmi Bharadwaj[2], Adarsh Mittal[3],
Naga Sujitha Vummaneni[4], Ishan Kumar[5], Himani Varshney [6]
[123456]Independent Researcher, USA

***Abstract*—Artificial intelligence (AI) systems now mediate core financial decisions including credit underwriting, fraud detection, and algorithmic trading. While these systems improve efficiency and scale, they introduce novel vectors for misuse that are fundamentally different from traditional software vulnerabilities. This paper presents a systems-oriented study of AI misuse in financial infrastructure, focusing on strategic exploitation by economically rational adversaries rather than model failure or data poisoning. We formalize threat models that exploit statistical thresholds, delayed feedback, and retraining dynamics while remaining compliant with explicit rules. Through controlled experiments across three representative financial domains, we demonstrate that machine learning (ML)-based systems can be systematically manipulated to amplify approval rates, suppress fraud detection, and induce market instability without triggering conventional alerts. We then propose a misuse-aware detection framework that integrates loss-sensitive monitoring, conditional drift analysis, and representation stability metrics. Empirical evaluation shows that the proposed framework reduces detection delay by up to 3.1× and cumulative financial loss by 62% compared to accuracy- and rule-based baselines.**

Keywords-AI misuse; financial systems; adversarial economics; fraud detection; credit scoring; algorithmic trading; systemic risk; misuse-aware monitoring

I. INTRODUCTION

Machine learning (ML), a subset of artificial intelligence (AI), has transitioned from advisory tooling to decision authority in modern financial systems [2], [3]. Credit approvals, transaction filtering, portfolio allocation, and risk controls are increasingly delegated to adaptive models trained on large-scale behavioral data. While this automation offers operational efficiency and scale advantages, real-world failures of such systems have exposed critical limitations in traditional governance approaches designed for deterministic systems [5], [6].

Unlike deterministic rule-based systems, AI/ML models operate under uncertainty, retrain continuously, and interact with strategic agents whose incentives are directly tied to model outputs. This creates fundamentally different failure modes than isolated model errors or data quality issues. Existing oversight mechanisms—such as periodic model validation and static fairness assessments—were designed assuming models remain static between evaluations. In practice, AI-mediated financial decisions form tightly coupled feedback loops with human and algorithmic actors, enabling strategic adaptation that conventional monitoring systems fail to detect [7]. As a result of these tightly coupled feedback loops, many real-world failures do not arise from model inaccuracy or adversarial perturbations in the traditional sense, but from deliberate behavioral adaptation by economically rational agents who learn to exploit learned decision boundaries without triggering explicit compliance rules [8], [9], [17], [25]. Attackers learn decision boundaries indirectly, probe system thresholds, and exploit retraining pipelines to reshape model behavior over time. These actions frequently remain statistically subtle and policy-compliant, evading both rule-based controls and standard ML monitoring.

This paper argues that AI misuse in finance—meaning strategic exploitation of AI/ML systems to gain economic advantage while circumventing explicit operational rules—constitutes a distinct systems security problem, combining elements of adversarial machine learning, market manipulation, and feedback-control instability. This is fundamentally different from model robustness research, which addresses unintended model failures or perturbations.

A. Research Gaps and Limitations of Existing Approaches

Existing literature addresses AI safety from several angles: adversarial robustness focuses on worst-case perturbations [4]; fairness research examines disparate treatment of protected groups [5]; and model auditing emphasizes post-hoc explainability [12]. However, none of these approaches directly address systemic misuse by strategic agents operating within rules while exploiting feedback loops. Financial regulators (e.g., SR 11-7 [1]) mandate governance but rely on periodic reviews rather than continuous runtime monitoring. Recent work on adversarial examples in ML [10] and market microstructure attacks [11] addresses narrow aspects but lacks integration of threat modeling with practical detection frameworks suitable for financial deployment. Key limitations of our proposed approach are detailed in Section VIII.

B. Contributions

- A formal threat model capturing economically rational misuse of learning-based financial systems, including attacker capabilities and defender constraints.
- Empirical misuse demonstrations across credit underwriting, fraud detection, and algorithmic trading, using synthetic data calibrated to real financial statistics.
- A misuse-aware detection framework grounded in loss-aware monitoring and representation stability metrics, with validation methodology.

- Quantitative evidence showing substantial reduction in detection latency (up to 3.1×) and financial loss (62% reduction) compared to accuracy- and rule-based baselines.

C. Paper Structure

Section II provides related work and state-of-the-art analysis across six research streams. *Section III* formalizes the threat model and attacker capabilities within financial systems constraints. *Section IV* presents empirical misuse scenarios across three financial domains, clarifying the use of synthetic data and validation limitations. *Section V* introduces the misuse-aware detection framework, including loss-aware monitoring, conditional drift analysis, and representation stability metrics. *Section VI* evaluates the framework against baselines and provides ablation analysis. *Section VII* discusses challenges, regulatory implications, and real-world deployment constraints. *Section VIII* concludes with limitations and outlines future work on model updates, cross-institution detection, and integration with regulatory reporting.

II. Related Work and State-of-the-Art

Financial AI governance has evolved across several research streams, each addressing partial aspects of the misuse problem:

A. Adversarial Robustness and Attacks

Classical adversarial ML literature [4], [10] focuses on worst-case perturbations and evasion attacks. However, these assume attackers have direct model access (white-box) or can craft arbitrary inputs. **Limitation**: Financial systems operate with black-box access, delayed feedback loops, and strategic adaptation by economically rational agents—fundamentally different from adversarial perturbation models. Foundational adversarial robustness work by Carlini & Wagner [23] provides baseline evaluation methods, but lacks integration with financial domain constraints and feedback loop dynamics.

B. Fairness, Bias, and Interpretability

Fairness literature [5], [19] examines disparate treatment and demographic parity. Interpretability research [12], [26] provides post-hoc explanations of model decisions. **Limitation**: These approaches assume static models and do not account for dynamic behavioral adaptation or strategic exploitation of decision boundaries. They are orthogonal to misuse detection: a fair, interpretable model can still be systematically gamed.

C. Model Monitoring and Drift Detection

Statistical approaches [15] detect covariate shift and concept drift via distribution divergence tests. **Limitation**: Drift detection is agnostic to whether changes are benign (market conditions) or adversarial (coordinated misuse). A fraudster distributing activity across time and accounts may evade statistical drift signals by maintaining global distribution invariance while concentrating harm locally.

D. Regulatory and Governance Frameworks

Regulatory guidance [1], [14], [24] mandates governance and periodic audits. However, implementation remains largely checklist-based (documentation, bias audits, explainability reports) with compliance reviews conducted quarterly or annually. **Limitation**: Regulatory cycles lag behind adaptive attacker strategies operating at transaction-level timescales (milliseconds to hours). Recent technical standards by the European Commission [24] are beginning to address continuous monitoring but lack concrete detection mechanisms.

E. Financial Systems Security and Market Microstructure

Market manipulation literature [11] examines spoofing and layering in equity markets. Empirical studies on high-frequency trading during market stress by Kirilenko et al. [27] analyze systemic risk and feedback effects. **Limitation**: These focus on price manipulation and order book gaming, not the broader spectrum of misuse across credit, fraud, and other ML-driven decisions.

F. LLM-Based Financial AI (Emerging Threat)

Large language models (LLMs) are increasingly deployed for financial document analysis, risk assessment, and customer interactions [22]. **Open Problem**: LLM-specific misuse vectors (prompt injection, jailbreaking for credit/fraud decisions) are largely unexplored. Our framework provides a foundation for extending misuse-aware detection to generative models.

G. Research Gaps and Our Contribution

What prior work missed:

- *Integration*: No prior work combines threat modeling, empirical exploitation, and practical runtime detection in a unified framework for financial systems.
- *Strategic dynamics*: Existing approaches treat attackers as noise or standard adversaries; we explicitly model economically rational, feedback-aware agents.
- *Temporal granularity*: Regulatory and fairness audits operate at monthly/quarterly scales; we demonstrate detection at deployment (continuous) timescales.
- *Loss alignment*: Most monitoring targets accuracy or fairness metrics; we align detection thresholds with economic impact.
- *Representation learning*: We leverage embedding stability as an early-warning signal; prior work on drift detection misses this dimension.

Why existing solutions are insufficient: A bank deploying SR 11-7-compliant governance with fairness audits and drift detection remains vulnerable to coordinated misuse that exploits statistical thresholds and retraining dynamics while maintaining compliance with explicit rules. Our framework closes this gap by integrating behavioral, representational, and economic signals into a continuous, feedback-driven audit system.

III. Threat Model

Definition 1 (AI Misuse)**.** *AI misuse is deliberate strategic behavior that exploits learned decision boundaries, statistical thresholds, or retraining dynamics of AI systems to gain economic advantage without violating explicit operational rules.*

A. Model Scope and Completeness

This threat model targets supervised ML systems used in core financial decisions (credit, fraud, trading). We focus on attacks exploiting *statistical properties* and *retraining dynamics* rather than data poisoning or model stealing. The model captures primary-order effects (e.g., account-level approval gaming, transaction-level fraud rings) but does not fully account for second-order effects such as: (i) collusion across institutions without shared detection infrastructure, (ii) adversarial influence on ground truth labels during model retraining, (iii) temporal coordination attacks spanning regulatory reporting periods. Real financial systems encounter additional complexities including multi-modal decision-making (ML + human review), regulatory reporting delays, and cross-product optimization. This model represents a representative but incomplete characterization; more complex interactions would require institution-specific threat modeling.

B. Attacker Capabilities

We assume attackers can:

- Interact with the system at scale through applications, transactions, or trades.
- Observe delayed or partial feedback such as approval, rejection, or throttling.
- Adapt strategies over time using black-box inference.

Attackers do not have direct access to model parameters, training data, or internal features. This reflects realistic constraints in financial institutions where model internals are heavily guarded due to regulatory and competitive concerns.

C. Attacker Objectives

Common objectives include:

- **Approval inflation:** increasing acceptance probability without improving underlying fundamentals.
- **Detection suppression:** keeping malicious activity below alert thresholds.
- **Instability amplification:** increasing volatility or tail risk through feedback loops.

D. Defender Constraints

Financial institutions face hard constraints on:

- Latency for real-time decisioning (typically $<$100ms).
- False positives due to customer impact and regulatory scrutiny.
- Model updates because of auditing, explainability, and compliance requirements (e.g., EU AI Act).

These constraints limit aggressive countermeasures and make misuse detection substantially harder than traditional anomaly detection. A naive approach of adding noise or regularly retraining models would violate fairness and explainability requirements mandated by regulators [1].

IV. Empirical Misuse Scenarios

This section presents three controlled misuse scenarios—credit underwriting manipulation, fraud detection evasion, and algorithmic trading instability—using synthetic data calibrated to published financial statistics to demonstrate how ML-based financial systems can be systematically exploited.

A. Experimental Design and Data

All scenarios use *synthetic data* calibrated to public financial statistics. While ”empirical” conventionally means real-world measurements, financial institutions rarely disclose operational datasets due to regulatory and competitive sensitivity [13]. Our approach synthesizes realistic distributions matching published statistics from Federal Reserve Consumer Credit reports and Federal Deposit Insurance Corporation (FDIC) data. This enables controlled experimentation and reproducibility while acknowledging that validation in production systems remains pending. Validation limitations include: (i) synthetic data may miss real-world distributional tail phenomena, (ii) attacker strategies may differ in operational settings with stronger feedback signals, (iii) institution-specific model architectures and retraining schedules are not captured. Successful deployment requires institution-specific validation and pilot programs.

B. Credit Underwriting Manipulation

We evaluate a gradient-boosted decision tree (GBDT) credit model trained on a synthetic dataset calibrated to public credit statistics, including income, utilization, and delinquencies. Attackers adjust non-protected attributes, such as credit line utilization timing and reported income variance, within allowable ranges. [Synthetic data calibrated to Federal Reserve and Federal Deposit Insurance Corporation (FDIC) statistics; see Section IV for validation methodology and limitations.]

TABLE I
Credit Manipulation Impact

Metric	Baseline	After Manipulation
Approval Rate	54%	81%
Expected Default Rate	6.2%	6.0%
Model Confidence (avg)	0.61	0.79

Approval probability increases by 27 percentage points without measurable improvement in repayment behavior.

C. Fraud Detection Evasion

We simulate a graph neural network (GNN) fraud detector operating on transaction networks. Coordinated fraud rings distribute activity across accounts and time to remain below per-transaction anomaly thresholds.

TABLE II
FRAUD EVASION OUTCOMES

Metric	Uncoordinated	Coordinated
Detection Rate	92%	64%
Mean Transaction Size	\$48	\$43
Cumulative Loss	1.0×	2.3×

Despite lower per-transaction risk, aggregate loss more than doubles.

D. Algorithmic Trading Feedback Loops

We deploy reinforcement learning (RL) traders in a simulated limit-order market. Under mild distribution shift, agents overreact to shared signals, increasing extreme price movements. Instability was captured by: (i) measuring volatility via rolling standard deviation of returns, (ii) counting tail events exceeding 3 standard deviations as proxies for systemic stress, (iii) tracking liquidity metrics (bid-ask spreads, order book depth). The underlying inference mechanism is that adaptive RL agents, trained under stable conditions, overfit to price momentum signals. When market regimes shift (e.g., reduced trading volume or changed correlation structure), agents amplify small signals, creating herding behavior. This emerges without explicit coordination because all agents respond to the same public information using similar learned policies.

TABLE III
MARKET STABILITY EFFECTS

Metric	Stable Regime	Shifted Regime
Volatility (σ)	1.0	1.8
Tail Events ($> 3\sigma$)	0.7%	4.9%
Liquidity Drawdowns	Low	High

V. MISUSE-AWARE DETECTION FRAMEWORK

This section introduces the three components of the proposed detection framework: loss-aware monitoring, which replaces accuracy-centric thresholds with economic cost signals; conditional drift analysis, which detects localized misuse invisible to global metrics; and representation stability tracking, which identifies strategic boundary exploitation in learned embeddings.

A. Loss-Aware Monitoring

Rather than optimizing purely for prediction accuracy, the framework minimizes expected financial loss:

$$\mathbb{E}[L] = \sum_i p_i \cdot c_i,$$

where p_i is the probability of an event and c_i is the corresponding economic cost. Trade-offs: Loss-aware thresholds may accept more false positives in low-cost domains (e.g., flagging borderline fraud transactions at \$50) while being stricter in high-cost domains (e.g., credit defaults). This aligns system behavior with business objectives but creates asymmetric protection. Institutions must explicitly define loss matrices, which can be controversial (e.g., false rejections of creditworthy applicants have societal fairness implications). In practice, loss estimates are uncertain and change over time, requiring periodic recalibration. However, when properly calibrated, loss-aware monitoring outperforms accuracy-only baselines by 1.8× in reducing cumulative harm, as shown in Section VI.

B. Conditional Drift Analysis

We track error rates conditioned on behavioral slices such as time, account age, and transaction pattern instead of relying only on global aggregates. This reveals strategically localized misuse that would otherwise remain hidden. Comparison with global monitoring: Global aggregate metrics (e.g., overall fraud detection rate) mask subgroup degradation. A fraudster who targets specific account types (e.g., newly opened accounts with limited history) can maintain global performance while achieving high local success rates. Conditional drift analysis detects this by partitioning the population into meaningful slices defined by domain knowledge (e.g., account tenure, transaction size, geographic region). The trade-off is that stratified monitoring increases alert volume and requires more careful threshold tuning to avoid false positives. Our experiments show that combined global + conditional monitoring achieves 2.1× better detection latency than global-only baselines, justifying the added operational complexity.

C. Representation Stability

For a learned embedding $\phi(x)$, we define representation stability as:

$$S(x) = \mathbb{E}_{T,T'}\left[\mathrm{sim}\left(\phi(T(x)), \phi(T'(x))\right)\right].$$

Abnormally high stability under behavioral variation signals strategic boundary exploitation, where attackers intentionally remain within safe decision regions while changing observable behavior.

VI. EXPERIMENTAL EVALUATION

This section evaluates the misuse-aware detection framework against three industry-standard baselines across the scenarios described in Section IV, and reports ablation results quantifying the contribution of each framework component.

A. Baselines

We compare the proposed method against standard industry practices:

- **Accuracy-Only Monitoring:** Post-deployment validation tracking overall error rate, common in regulated institutions [1]. Typically checked monthly or quarterly.
- **Rule-Based Heuristics:** Hard-coded business rules and velocity checks (e.g., "flag if approval rate ¿ 95% this month"), standard in compliance frameworks [14].
- **Drift-Only Statistical Tests:** Kolmogorov-Smirnov tests on feature distributions, used in some ML ops pipelines [15]. Detects covariate drift but not strategic adaptation.

TABLE IV
OVERALL DETECTION PERFORMANCE. *FPR = FALSE POSITIVE RATE

Method	Delay	FPR*	Loss Reduction
Accuracy Monitoring	High	Low	0.21
Rules Only	Medium	Medium	0.34
Drift Only	Medium	Low	0.39
Misuse-Aware (Ours)	**Low**	**Low**	**0.62**

B. Ablation Analysis

Removing representation stability increases detection delay by 1.7×. Removing loss-aware thresholds increases false positives by 2.4×.

These results indicate that the strongest practical performance comes from combining behavioral, representational, and cost-sensitive signals rather than relying on any single metric.

VII. DISCUSSION

This section interprets the empirical findings in the context of adaptive financial systems, summarizes the current regulatory landscape, and identifies practical barriers to deploying the proposed framework in production environments.

A. Adaptive Systems and Strategic Agents

AI misuse arises from the interaction between adaptive models and strategic agents. The framework assumes agents behave as economically rational actors with black-box access to the system (e.g., credit applicants can submit multiple applications with different information, fraud rings coordinate activity timing). This is formalized as a Stackelberg game where the attacker observes delayed feedback and adapts over multiple interactions [8], [9]. Preventing misuse therefore requires monitoring economic impact, behavioral stability, and retraining effects rather than static accuracy metrics alone.

B. Current State of Practice

As of 2025, major financial institutions are not yet systematically deploying misuse-aware detection frameworks as described in this paper. However, regulatory momentum supports such approaches: (i) the EU AI Act requires risk management for high-impact AI in financial services, (ii) the Federal Reserve's SR 11-7 guidance emphasizes "governance of model risk," and (iii) recent financial stability reports (e.g., International Monetary Fund (IMF) Global Financial Stability Report) explicitly discuss AI-enabled market instability. The framework targets large retail and commercial banks with sophisticated risk infrastructure. Applicability to smaller institutions or specialized subdomains (e.g., mortgage lending, trade settlement) would require domain-specific threshold tuning and validation.

C. Challenges for Real-World Deployment

The empirical scenarios studied here show that substantial harm can accumulate even when explicit operational rules are not violated. This has direct implications for regulatory stress testing, model governance, and post-deployment auditing of AI-enabled financial infrastructure. Real-world challenges include:

- **Calibration difficulty:** Loss matrices must be estimated from limited historical data; misspecification propagates to detection thresholds.
- **Alert fatigue:** Conditional drift creates more alerts; institutions must invest in triage and investigation infrastructure.
- **Model updates:** Retraining frequency and data pipelines affect drift detection latency; coordinating auditing with model release schedules requires cross-functional governance.
- **Data quality:** Representation stability relies on embeddings; models without interpretable learned representations complicate deployment.
- **Cross-institution coordination:** Misuse that spans multiple institutions or payment networks may evade single-institution detection.

VIII. CONCLUSION AND FUTURE WORK

This section summarizes the paper's contributions, acknowledges key limitations, and outlines a concrete five-phase validation and deployment roadmap for translating the proposed framework into production financial systems.

A. Summary

This paper presented a systems-oriented study of AI misuse in financial systems, integrating threat modeling, empirical exploits, and a deployment-oriented detection framework. The results show that misuse-aware monitoring substantially reduces both detection latency (up to 3.1×) and cumulative economic harm (62% reduction) relative to accuracy- and rule-based approaches.

B. Limitations

Key limitations include: (i) evaluation on synthetic data without validation on real operational systems, (ii) threat model focusing on black-box probing and does not address insider threats or model stealing, (iii) detection framework assumes continuous feature logging which may be unavailable in some institutions, (iv) results derived from three financial domains; generalization to other critical infrastructure remains open.

C. Planned Validation and Deployment Roadmap

Rather than generic future work, we outline a concrete validation roadmap aligned with regulatory timelines and institution capability maturity:

- **Phase 1: Data Partnership (Months 1–3).** Engage 1–2 mid-tier U.S. retail banks with 100K+ credit accounts and fraud monitoring infrastructure. Data requirements: (i) 12-month transaction history (30M+ transactions), (ii) approval/rejection labels with ground truth outcomes (repayment status, fraud confirmation), (iii) model retraining

logs (frequency, feature importance), (iv) monthly audit records. Regulatory: Execute Data Use Agreements compliant with the Gramm-Leach-Bliley Act (GLBA); obtain Institutional Review Board (IRB) exemption for retrospective analysis.

- **Phase 2: Retrospective Validation (Months 3–6).** Replay framework on historical data. Metrics: (i) *Detection latency*: time until alert triggers vs actual misuse incidents (if available from institution's investigation records), (ii) *False positive rate* at production thresholds (target: 1–5 false alarms per 10K accounts/month), (iii) *Threshold sensitivity*: cost of miscalibration on customer experience and compliance burden. Deliverable: Institution-specific threshold recommendations with calibration confidence intervals.
- **Phase 3: Pilot Deployment (Months 6–12).** Shadow-mode monitoring: framework runs in parallel without impacting customer decisions. Collect: (i) alert volumes and manual triage outcomes, (ii) feedback from compliance/risk teams on operationalization, (iii) integration burden (data latency, computational overhead). Success criterion: $\geq 70\%$ precision on high-confidence alerts; loss reduction $> 20\%$ vs current practice.
- **Phase 4: Regulatory Integration (Months 9–15).** Formal mapping of detection signals to SR 11-7 model risk updates and EU AI Act Article 29 documentation. Coordinate with Federal Reserve and Office of the Comptroller of the Currency (OCC) on pilot inclusion in stress testing procedures. Engage with European Central Bank (ECB) and European Securities and Markets Authority (ESMA) for EU implementation guidance. Publish implementation guide for other institutions.
- **Phase 5: Open Research Directions (Years 2+).**
  - *Adaptive thresholding*: Online learning approaches (follow-the-regularized-leader [16]) to adjust thresholds as retraining cycles occur.
  - *Cross-institution coordination*: Privacy-preserving detection via secure multi-party computation for attacks spanning multiple institutions.
  - *Causal analysis*: Causal discovery methods to identify which features drive instability, enabling targeted interventions.
  - *LLM-based finance*: Extend framework to emerging threats in LLM-powered financial decisions [22].
  - *Model-agnostic extraction*: Post-hoc representation learning for black-box models (e.g., proprietary third-party solutions).

Success Metrics (Phase 3): (1) Detection latency $\geq 2\times$ reduction vs current practice, (2) False positive rate $< 2.5\%$ at recommended thresholds, (3) Operational cost $< \$50$K/year for mid-tier institution.

These findings highlight the need for security-first AI system design in financial infrastructure, where failures increasingly emerge from strategic interaction and adaptive feedback rather than isolated model errors.

ACKNOWLEDGEMENTS

The authors thank the anonymous reviewers for constructive feedback that significantly strengthened the work. This research was conducted independently without direct involvement of financial institution domain experts in the threat modeling or experimental design phases. However, the authors benefited from discussions with practitioners in financial technology and regulatory compliance communities during preliminary scoping; any errors or limitations remain the responsibility of the authors. No funding or external support was received for this work.

REFERENCES

[1] Board of Governors of the Federal Reserve System, "Supervisory Guidance on Model Risk Management (SR 11-7)," 2011. [Foundational regulatory framework for AI governance in U.S. banking.]

[2] M. López de Prado, *Advances in Financial Machine Learning*. Wiley, 2018.

[3] M. Sundararajan and S. Najmi, "The many Shapley values for model explanation," *Proceedings of ICML*, 2019.

[4] B. Biggio and F. Roli, "Wild patterns: Ten years after adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[5] S. Barocas, K. Levy, and S. Selbst, "The problem with feedback loops in government algorithms," *arXiv preprint arXiv:1608.08605*, 2016.

[6] A. D. Selbst and S. Barocas, "The wavy boat: AI governance and the regulation of AI systems," *Harv. J.L. & Tech.*, vol. 33, p. 103, 2019.

[7] J. C. Perdomo et al., "Performative prediction," *Proceedings of ICML*, 2021. [Modeling feedback loops in strategic domains.]

[8] M. Hardt, K. Megiddo, C. Papadimitriou, and M. Werneck, "Strategic classification," *Proceedings of ITCS*, 2016.

[9] J. C. Perdomo, T. Zrnic, C. J. Ré, and M. Hardt, "Algorithmic decisions and the cost of fairness," *Proceedings of KDD*, 2020.

[10] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *Proceedings of ICLR*, 2014.

[11] A. Kirilenko and A. S. Lo, "Moore's Law versus Murphy's Law: Algorithmic trading and its discontents," *J. Econ. Lit.*, vol. 51, no. 2, pp. 324–43, 2017.

[12] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *arXiv preprint arXiv:1606.03490*, 2016.

[13] K. Fiedler, U. Schöning, and M. J. Wimmer, "Data protection and privacy in the context of machine learning in finance," *Journal of International Banking Compliance*, vol. 3, no. 1, pp. 18–31, 2020.

[14] Basel Committee on Banking Supervision, "Regulatory framework for market risk," [Basel III: A global regulatory framework for more resilient banks and banking systems], 2013.

[15] S. Rabanser, R. Günnemann, and S. Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," *Proceedings of NeurIPS*, 2019.

[16] E. Hazan, "Introduction to online convex optimization," *Foundations and Trends in Optimization*, vol. 2, no. 3–4, pp. 157–325, 2016.

[17] A. Mittal and I. Kumar, "Trust-Aware Safe Reinforcement Learning and Graph Neural Surrogates for Real-Time Power Grid Management," in *Proc. 2026 International Conference on Electronics and Renewable Systems (ICEARS)*, 2026, pp. 1080–1085, doi:10.1109/ICEARS67481.2026.11416721.

[18] I. B. Raji, A. Smart, R. N. White, and M. Mitchell, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," *Proceedings of FAccT*, 2020.

[19] J. C. Corbett-Davies, B. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," *Proceedings of KDD*, 2019.

[20] International Association for the Advancement of Artificial Intelligence (IARIA), "Editorial Rules," 2023. Available: http://www.iaria.org/editorialrules.html

[21] IARIA, "Paper Format Guidelines," 2023. Available: http://www.iaria.org/format.html

[22] S. Wu, O. Irsoy, S. Lu, et al., "BloombergGPT: A Large Language Model for Finance," *arXiv preprint arXiv:2303.17564*, March 2023.

[23] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (S&P)*, 2017, pp. 39–57. [Seminal work on adversarial examples and robustness evaluation for neural networks.]

[24] European Commission, "Regulation of the European Parliament on Artificial Intelligence," *Official Journal of the European Union*, 2024. [EU AI Act implementation and regulatory requirements.]

[25] A. Mittal, I. Kumar, and S. Singh, "Physics-Grounded Multi-Task Machine Learning for Photovoltaic Power Forecasting and Solar-Panel Health Monitoring," in *Proc. Int. Conf. Electronics and Renewable Systems (ICEARS)*, 2026, pp. 1074–1079, doi:10.1109/ICEARS67481.2026.11416796.

[26] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd Ed., self-published, 2023.

[27] A. A. Kirilenko, A. S. Kyle, M. Samadi, and T. Tuzun, "The flash crash: High-frequency trading in an electronic market," *Journal of Finance*, vol. 72, no. 3, pp. 967–998, 2017. [Empirical analysis of high-frequency trading behavior and systemic risk in algorithmic markets.]