FUTURE COMPUTING 2026 · Lisbon, Portugal

Systemic Misuse of AI in Financial Systems

Threat Models, Empirical Exploits, and Misuse-Aware Detection

Usha Ratnam Jammula · Srilakshmi Bharadwaj · Adarsh Mittal
Naga Sujitha Vummaneni · Ishan Kumar · Himani Varshney

FUTURECOMPUTING2026 · April 19–23, 2026

THE PROBLEM

AI now makes core financial decisions. But strategic agents exploit its logic — without breaking any rule.

Credit Scoring

Applicants game approval thresholds by timing utilization and income variance — without improving actual creditworthiness.

Fraud Detection

Coordinated rings distribute transactions below anomaly thresholds. Global metrics look fine; cumulative loss doubles.

Algo Trading

RL agents overfit to shared signals. Under mild regime shift, herding behavior amplifies volatility 1.8× with 4.9% tail events.

Misuse ≠ model failure. It's rational exploitation of learned decision boundaries.

FORMAL THREAT MODEL

Definition: AI misuse is deliberate strategic behavior that exploits learned decision boundaries, statistical thresholds, or retraining dynamics to gain economic advantage — without violating explicit operational rules.

ATTACKER CAPABILITIES

- Interact at scale (applications, transactions, trades)
- Observe delayed/partial feedback (approve/reject/throttle)
- Adapt over time via black-box inference
- No direct access to model internals or training data

DEFENDER CONSTRAINTS

- Real-time decisioning < 100ms latency
- Low false-positive tolerance (customer & regulatory impact)
- Slow model update cycles (audit/explainability compliance)
- Cannot add arbitrary noise (fairness & EU AI Act)

Modeled as a Stackelberg game: attacker observes delayed feedback and adapts over multiple interactions.

EMPIRICAL MISUSE SCENARIOS

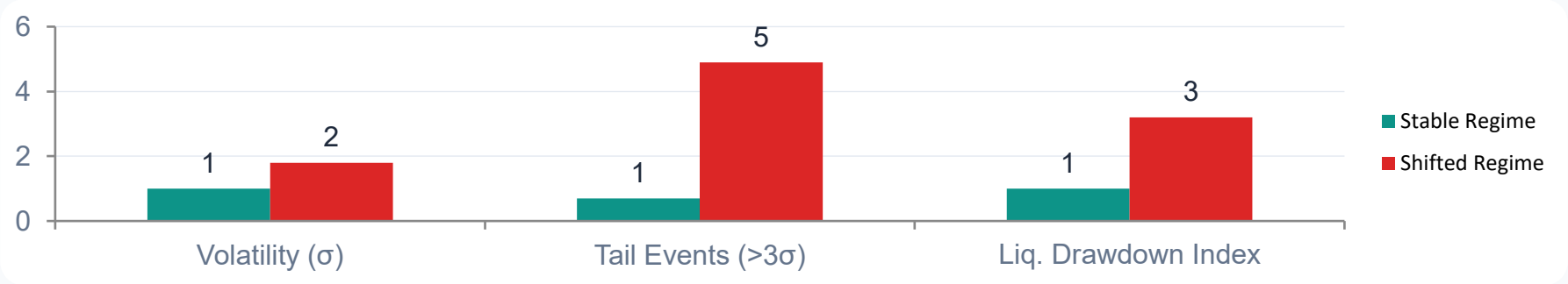
Credit Underwriting Manipulation

Metric	Baseline	After
Approval Rate	54%	81% ▲
Expected Default	6.2%	6.0%
Model Confidence	0.61	0.79

Fraud Detection Evasion

Metric	Uncoord.	Coordinated
Detection Rate	92%	64% ▼
Mean Txn Size	$48	$43
Cumulative Loss	1.0×	2.3× ▲

Algorithmic Trading — Market Stability Effects



MISUSE-AWARE DETECTION FRAMEWORK

01

Loss-Aware Monitoring

Replaces accuracy-centric thresholds with economic cost signals.
$E[L] = \Sigma\ p_i \cdot c_i$
Aligns detection sensitivity with financial impact — stricter in high-cost domains.

02

Conditional Drift Analysis

Tracks error rates on behavioral slices (account age, txn pattern, geography).
Detects localized misuse invisible to global aggregate metrics.
2.1× better detection latency vs. global-only.

03

Representation Stability

$S(x) = E[\text{sim}(\phi(T(x)), \phi(T'(x)))]$
Abnormally high stability under behavioral variation signals strategic boundary exploitation.

Ablation: removing representation stability → 1.7× more delay · removing loss-aware thresholds → 2.4× more false positives

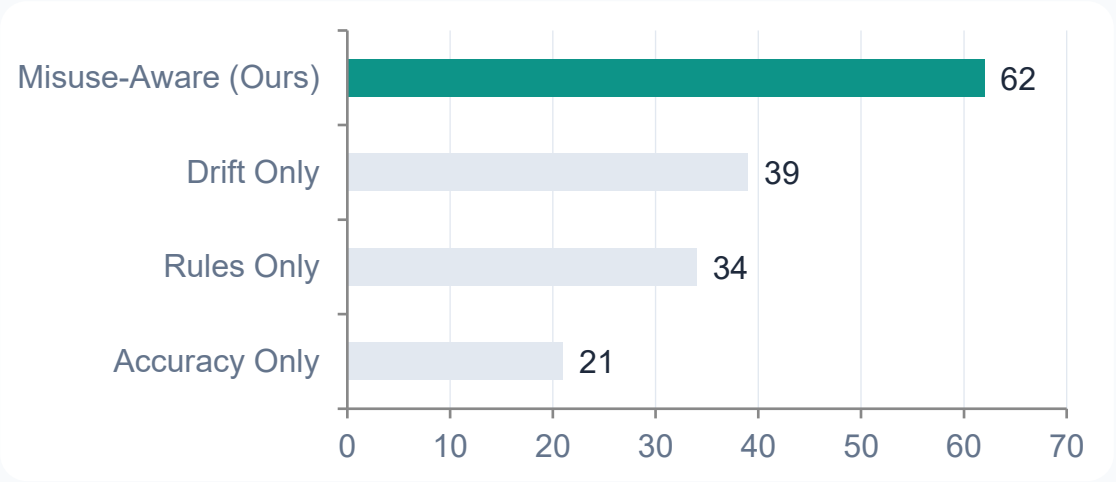
EVALUATION RESULTS

3.1×
Faster Detection

62%
Loss Reduction

Low
False Positive Rate

Detection Performance vs. Baselines



Detection Delay

Method	Delay
Accuracy Only	High
Rules Only	Medium
Drift Only	Medium
Ours	**Low ✓**

REGULATORY CONTEXT & DEPLOYMENT ROADMAP

Regulatory Alignment

- SR 11-7 (Fed): Model risk governance mandate
- EU AI Act: Continuous monitoring for high-impact AI
- IMF GFSR: Explicit AI-enabled market instability concern
- Basel III: Risk framework applicability to ML systems

5-Phase Validation Roadmap

- Ph 1 (M1–3): Data partnership, GLBA/IRB compliance
- Ph 2 (M3–6): Retrospective validation on 30M+ txns
- Ph 3 (M6–12): Shadow pilot — 70%+ precision target
- Ph 4 (M9–15): Regulatory integration (Fed/OCC/ECB)
- Ph 5 (Yr 2+): Adaptive thresholds, cross-institution detection

Known Limitations

- Evaluated on synthetic data only — real-system validation pending
- Threat model excludes insider threats and model stealing
- Requires continuous feature logging (not universal)
- Framework targets 3 domains; generalization remains open

Key Takeaways

- AI misuse in finance is a distinct systems security problem — strategic, rule-compliant, and feedback-aware
- Synthetic experiments show 27pt approval inflation, 2.3× fraud loss amplification, and 1.8× volatility increase
- Misuse-aware framework reduces detection delay up to 3.1× and cumulative loss by 62% vs. industry baselines
- Security-first AI design is necessary where failures emerge from strategic interaction, not model error

Questions? · Paper available at FUTURECOMPUTING2026 proceedings · steve@iaria.org for inquiries