Agentic AI in E-Commerce: Security Risks of Autonomous Decision Loops in Discovery, Pricing, and Routing Systems

Shaibal Maji

University of the Cumberlands, USA

Abstract

Agentic artificial intelligence is becoming increasingly prevalent in large-scale e-commerce systems, where it controls critical functions including product discovery, dynamic pricing, and fulfillment routing. Unlike traditional decision support systems that need human approval before taking action, agentic architectures can act directly through platform services, creating ongoing cycles of observing, deciding, and acting with very little human involvement. While such systems reduce response latency and improve operational efficiency, they simultaneously introduce a fundamentally new category of security and integrity hazards that extend well beyond conventional application security concerns. This article examines how the extension of autonomous decision loops creates attack surfaces that traditional security frameworks fail to address adequately. The article focuses on failure modes and adversarial risks across three key commerce domains: product discovery and ranking, dynamic pricing and promotion management, and fulfillment routing optimization. By carefully examining these areas, the study finds common ways that systems can be exploited, such as manipulating feedback, hacking rewards, poisoning data, and misusing permissions by agents with too much authority. A specific threat model for agentic commerce systems is provided, along with controls for architecture and operations that focus on enforcing rules, ensuring visibility, limiting potential damage, and establishing governance. The main point made is that to protect agentic e-commerce systems, we need to prioritize decision integrity and loop stability as key security issues, which calls for new methods that work alongside, not instead of, traditional security practices.

Keywords: Agentic Artificial Intelligence, E-Commerce Security, Autonomous Decision Loops, Dynamic Pricing, Fulfillment Routing

I. Introduction

1.1 Evolution of Artificial Intelligence in E-Commerce Platforms

Electronic commerce platforms have undergone substantial transformation through the progressive integration of artificial intelligence capabilities over the past two decades. Early commercial applications of AI in retail environments were characterized by offline analytics, batch-processed recommendations, and forecasting systems where human operators reviewed and approved outputs before they influenced production systems. These initial implementations provided valuable decision support while maintaining clear boundaries between algorithmic suggestions and actual business execution. The fundamental architecture ensured that trained personnel evaluated AI outputs against business constraints, regulatory requirements, and contextual factors before any customer-facing changes occurred. This human-in-the-loop paradigm offered natural checkpoints where errors could be identified and corrected before propagating through interconnected systems. Machine learning applications in e-commerce have demonstrated significant advances across recommendation systems, demand forecasting, and customer behavior prediction, with modern implementations utilizing datasets containing millions of samples for training and achieving classification accuracy rates exceeding 90% in network traffic analysis tasks [1].

1.2 Emergence of Agentic Architectures and Real-Time Autonomy

Recent advancements in real-time decision-making, reinforcement learning, and the management of extensive language models have culminated in the emergence of agentic AI. This enables autonomous agents to continuously oversee the system, make decisions, and execute actions on platform services without requiring human authorization for each task. These agents work with a lot of power and frequency. For example, a single decision policy could affect millions of search impressions, set off thousands of price changes, or decide how to route a large part of the daily fulfillment volume. The shift from needing humans to manage transactions to having AI handle them on its own marks a major change in how online shopping systems work, as these AI agents can now carry out complicated tasks that used to need human help. This autonomy delivers compelling business benefits, including reduced latency, improved personalization, and enhanced operational efficiency, but simultaneously transforms the security posture of digital commerce infrastructure in ways that traditional security frameworks were not designed to address. The big change includes self-managing payment processing, smart order management, and real-time inventory updates, which significantly change how humans and algorithms work together, with studies showing that the number of articles linking evolutionary computation and multi-agent systems has steadily increased from 1995 to 2024.

1.3 Limitations of Traditional Security Models and Research Objectives

Conventional approaches to securing commerce platforms center on perimeter protection, access control enforcement, and data confidentiality preservation. While these measures remain necessary, they prove insufficient when applied to systems where authorized agents make consequential decisions at machine speed without human review. Agentic systems introduce risks to outcome integrity, create opportunities for feedback manipulation, and enable uncontrolled error amplification that can cascade through interconnected services. Self-adaptive approaches to mitigating safety risks provide foundational frameworks for understanding how information systems can dynamically respond to emerging threats while maintaining operational continuity [3]. Research on adversarial attacks demonstrates that baseline intrusion detection systems achieving 85% accuracy can be degraded to 60% accuracy through Fast Gradient Sign Method attacks, 55% through Projected Gradient Descent attacks, and 50% through DeepFool attacks, illustrating the vulnerability of AI systems to adversarial manipulation [6]. This paper pursues three primary objectives: first, to define a threat model specific to agentic commerce systems that captures risks absent from traditional security frameworks; second, to analyze domain-specific risks across discovery, pricing, and routing functions where autonomous agents operate; and third, to propose architectural patterns and operational controls that minimize systemic risk while preserving the autonomy benefits that motivate agentic adoption.

II. Architecture of Agentic Decision Loops

2.1 Structural Components and Operational Phases

Agentic commerce systems usually use closed-loop designs that have five different steps that keep running on a shared network. The observation phase involves ingesting diverse signals, including user behavioral data, transactional records, inventory positions, and external market information. Decision phases assess policies using a variety of methods, such as deterministic rules, statistical models, and reinforcement learning algorithms. These methods are often used together in hybrid configurations. Action phases execute decisions through privileged application programming interfaces that modify platform state, affecting what customers see, what prices they encounter, and how their orders are fulfilled. Evaluation phases measure outcomes against reward functions that encode business objectives, while policy refinement phases incorporate lessons learned through online adaptation or periodic retraining cycles. Self-adaptive approaches to system design provide foundational frameworks for understanding how autonomous agents can govern their behavior while maintaining safety properties, with implementations demonstrating the capacity for real-time risk assessment and dynamic response adjustment [3]. Research on AI-driven systems for handling cyber incidents shows that using three-layer designs with container technology can successfully separate and run different functions in production, honeypot, and digital forensics settings.

2.2 Distributed Execution and Trust Boundary Considerations

The continuous operation of agentic decision loops spans multiple microservices and crosses numerous trust boundaries within modern commerce architectures. Agentic systems make changes to production behavior in real time, so that any failure or compromise in a component can immediately affect customer-facing experiences. This is different from batch optimization systems, which make recommendations for people to review. Decentralized autonomous collaboration architectures demonstrate how large language models can empower coordinated decision-making across distributed systems, with smart contract-based implementations enabling trustless coordination among multiple autonomous agents operating across organizational boundaries [4]. Research on multi-agent systems indicates that agents in fine-grained agent-based evolutionary computation typically interact with four to eight neighboring agents in grid or cube structures, while coarse-grained approaches divide populations into multiple subpopulations with lower interaction frequency but stronger independence [7]. The security implications of this autonomy extend beyond preventing unauthorized access to include capabilities for limiting decision scope, explaining agent reasoning, and reversing actions when errors are detected. When agents operate continuously and at scale, the traditional security assumption that humans will notice and correct problems before significant harm occurs no longer holds, necessitating automated mechanisms for constraint enforcement and anomaly detection.

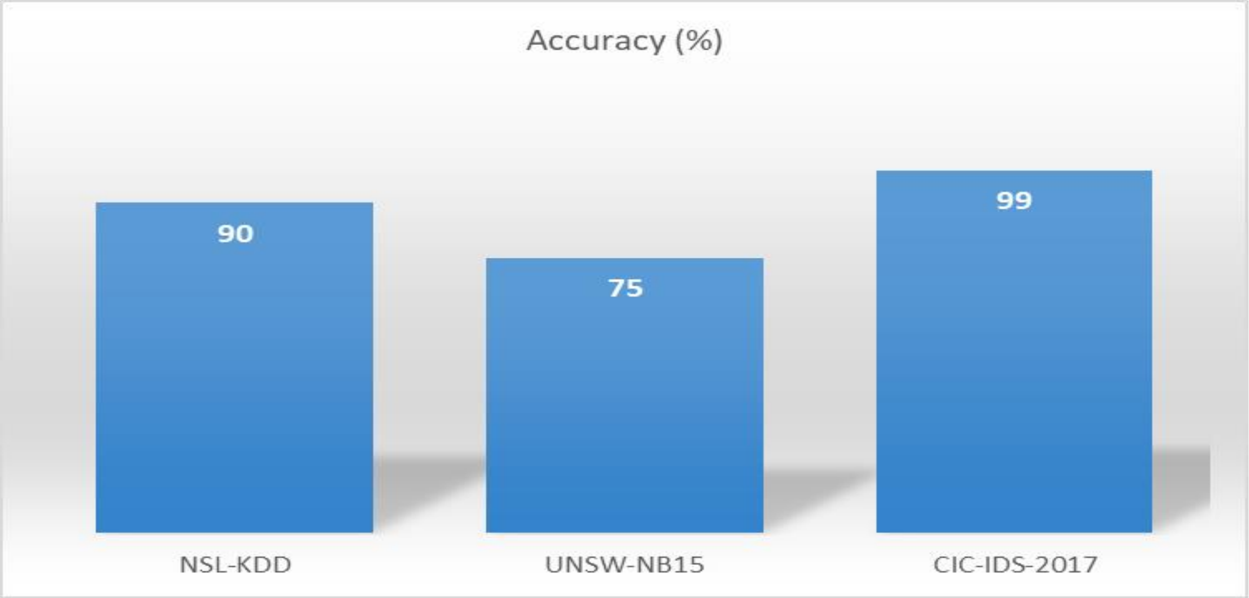

Fig. 1: Detection Accuracy Across Benchmark Datasets [9]

III. Threat Model for Agentic Commerce Systems

3.1 Adversary Taxonomy and Attack Motivations

Agentic commerce systems face threats from multiple adversary categories with distinct capabilities, access levels, and motivations. External actors may generate synthetic traffic, submit fraudulent transactions, or manipulate input signals to influence agent behavior without requiring any privileged access to platform systems. Merchandisers and trade partners occupy positions of partial trust where they provide legitimate inputs to catalog systems and pricing frameworks but may exploit this access to gain unfair competitive advantages. Insiders with access to model pipelines, configuration systems, or training infrastructure pose elevated risks given their ability to manipulate agent behavior at its source. Supply chain attackers target dependencies, model artifacts, or continuous integration credentials to compromise agents indirectly. Layer-based security analysis shows that autonomous systems have different weaknesses at each level of architecture, from perception to reasoning to actuation. This means that defense-in-depth strategies must deal with threats at multiple levels at the same time [5]. Research on autonomous system security demonstrates that perception-layer vulnerabilities, decision-layer manipulation, and actuation-layer exploitation represent fundamentally different attack surfaces requiring specialized defensive measures [5]. Studies on adversarial machine learning indicate that black-box attacks such as the Square Attack can reduce detection accuracy to 63%, demonstrating that adversaries do not require gradient information to successfully compromise AI systems [6].

3.2 Critical Assets and Unique Attack Surfaces

The critical assets requiring protection in agentic commerce systems include pricing integrity, ranking fairness, fulfillment correctness, financial margins, and regulatory compliance. Failures in protecting these assets manifest as revenue leakage, consumer harm, reputational damage, and operational instability. Adversarial threats to machine learning systems show clever ways to attack, such as data poisoning, model evasion, and inference manipulation, which need strong protective measures. Research on adversarial attacks against cloud-based intrusion detection systems reveals that gradient-based attacks such as the Fast Gradient Sign Method reduce baseline accuracy from 85% to 60%, Projected Gradient Descent attacks achieve accuracy degradation to 55%, and DeepFool attacks reduce accuracy to 50%, with corresponding F1-score reductions from 83% to 58%, 52%, and 48% respectively [6]. Precision metrics similarly degrade from a baseline of 84% to 61% under FGSM, 54% under PGD, and 49% under DeepFool attacks, while recall drops from 82% to 55%, 50%, and 47% respectively [6]. Agentic systems introduce attack surfaces absent from static systems, including reward manipulation where adversaries influence the signals agents optimize against, data poisoning of training and feedback pipelines, exploitation of reinforcement learning policy vulnerabilities, tool misuse through over-privileged agent identities, and feedback amplification where small perturbations cascade into runaway behavior. These vectors exploit the fundamental characteristic that distinguishes agentic systems: their continuous adaptation based on observed outcomes creates opportunities for adversaries to shape that adaptation toward harmful ends.

Attack Method	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
Baseline IDS	85	83	84	82
FGSM Attack	60	58	61	55
PGD Attack	55	52	54	50
DeepFool Attack	50	48	49	47

Table 1: Impact of Adversarial Attacks on IDS Model Performance [6]

IV. Domain-Specific Risk Analysis

4.1 Discovery Systems and Ranking Integrity Threats

Discovery agents that control search ranking and recommendation systems typically optimize engagement-based rewards such as click-through rates or add-to-cart conversions. Adversaries can exploit this optimization by generating synthetic engagement patterns that induce ranking biases favoring their products or content. The rapid adaptation that makes these agents effective also accelerates the impact of such attacks, as agents quickly incorporate manipulated signals into their policies. Semantic discovery systems that utilize vector embeddings for retrieval face additional risks from content crafted to achieve proximity to popular queries in embedding space, enabling systematic ranking manipulation in marketplace environments. When discovery agents incorporate natural language interpretation for query understanding or tool invocation, untrusted text from user queries or catalog data can influence execution paths, creating injection vulnerabilities where interpretation and execution lack strict separation. Multi-agent system coordination through evolutionary computation provides theoretical frameworks for understanding complex interactions among autonomous agents competing for ranking positions, with research demonstrating that evolutionary approaches can optimize multi-agent coordination across distributed systems [7]. The confluence of evolutionary computation and multi-agent systems reveals that agent-based optimization exhibits emergent behaviors that can be both beneficial for system performance and vulnerable to adversarial exploitation, with studies showing that multi-agent genetic algorithms utilizing orthogonal crossover operators can significantly improve optimization performance in grid-structured environments [7].

4.2 Pricing Systems and Autonomous Control Vulnerabilities

Autonomous pricing agents frequently incorporate external signals including competitor pricing data to inform their decisions. Manipulated external signals from adversary-controlled sources or low-integrity data providers can induce unintended price reductions at scale, with agents interpreting false competitive pressure as genuine market conditions requiring response. Promotion selection agents operating without adequate constraints on offer combinations create arbitrage opportunities where attackers systematically exploit the interaction between multiple concurrent promotions. Reward misalignment poses persistent risks when pricing agents optimize narrowly for conversion rates or gross merchandise value while ignoring downstream consequences including return rates, fraud exposure, and long-term customer trust erosion. Neuro-symbolic AI approaches offer potential solutions for combining learning-based pricing adaptation with rule-based safety constraints that encode business requirements agents cannot override, with research demonstrating that hybrid architectures can maintain constraint satisfaction while preserving adaptive capabilities [8]. The integration of symbolic reasoning with neural network learning enables systems to enforce explicit business rules while maintaining the flexibility to adapt to changing market conditions [8]. Research on anomaly detection in web intrusion systems demonstrates that optimizing contamination ratios is critical for balancing detection sensitivity and false positive rates, with studies showing that a contamination ratio of 0.01 provides optimal balance compared to stricter thresholds of 0.001 or more lenient settings of 0.1 [9]. Independent pricing policies governing related products or competing for shared inventory can interact to produce oscillatory behavior that damages both margins and customer perception of pricing stability.

4.3 Routing Systems and Fulfillment Optimization Risks

Fulfillment routing agents that optimize for cost minimization and delivery speed without incorporating fraud or abuse signals may preferentially route orders along paths that adversaries have identified as exploitable, concentrating risk rather than distributing it. Autonomous routing decisions depend critically on accurate real-time inventory information, and corrupted inventory signals, whether from system failures or adversarial manipulation cause fulfillment misses and cascade failures across distributed fulfillment networks. During partial outages affecting fulfillment dependencies, routing agents may engage in repeated replanning and retry attempts that amplify system load and operational costs rather than gracefully degrading. AI-powered systems for incident detection and response demonstrate approaches to identifying anomalous patterns in real-time operational data that could detect routing agents exhibiting problematic behavior [9]. Research on cloud-based cyber incident detection systems shows that they work differently in different network traffic environments. For example, they get 90.92% accuracy on the NSL-KDD dataset, 99.82% accuracy on the CIC-IDS-2017 dataset, and 75.64% accuracy on the UNSW-NB15 dataset [9]. The malware analysis component achieves 96.71% accuracy with 94.44% precision, 94.44% recall, and 94.44% F1-score using Random Forest, while the Keras deep learning model achieves 99% accuracy for binary classification tasks [9]. The network traffic classifier demonstrates a 91.82% detection rate on attack traffic, while the Web Intrusion Detection System successfully flags 89% of anomalous HTTP requests with minimal false positives [9]. The interconnected nature of modern fulfillment networks means that routing failures in one region can propagate to affect capacity and performance across the entire network, making blast-radius limitation essential for operational resilience.

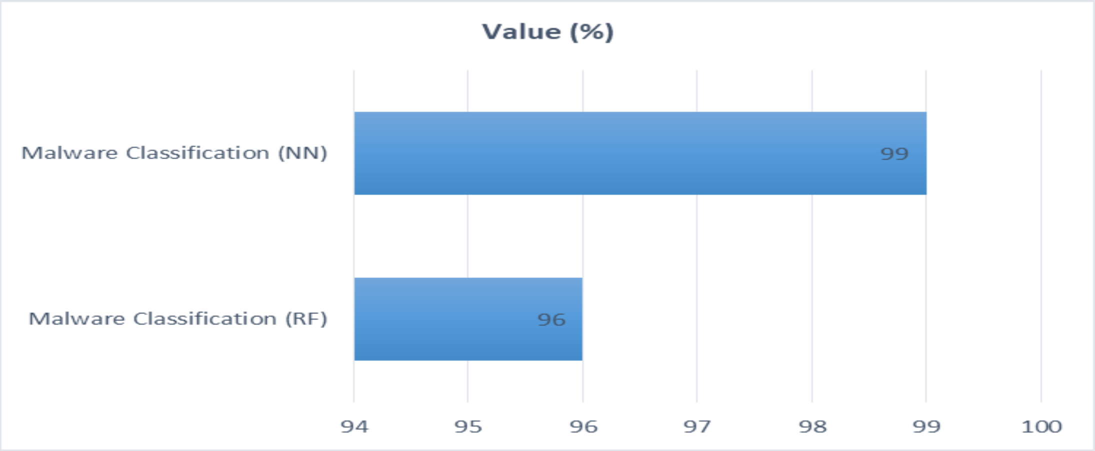

Fig. 2: Malware Analysis Performance Results [9]

V. Mitigation Framework

5.1 Architectural Controls for Bounded Autonomy

Effective security for agentic commerce systems requires architectural separation between decision formulation and action execution. Agents should propose actions rather than execute them directly, with deterministic enforcement layers validating that proposed actions satisfy constraints, authorization requirements, and policy compliance before permitting execution. Each agent should operate with dedicated identity credentials scoped to minimum necessary permissions, with comprehensive audit logging capturing both proposed and executed actions for forensic analysis. Systems must assume agent failure as inevitable and implement mechanisms to contain impact, including quotas limiting cumulative effect within time windows, rate limits preventing rapid large-scale changes, progressive rollout frameworks that expose new policies to increasing traffic gradually, and automatic rollback triggers that revert changes when outcome metrics deviate beyond acceptable thresholds. Research on adversarial defense mechanisms demonstrates that adversarial training combined with robust feature selection can restore model accuracy from degraded levels back to 88% against gradient-based attacks and 84% against black-box attacks, with defense mechanisms reducing attack success rates by up to 40% [6]. The dual-layered defense strategy incorporating adversarial training and SHAP-based robust feature selection achieves F1-score of 86%, precision of 87%, and recall of 85% compared to baseline metrics of 83%, 84%, and 82% respectively [6].

5.2 Observability Requirements and Governance Frameworks

Decision integrity observability extends beyond traditional metrics of latency and error rates to encompass policy drift detection, outcome volatility measurement, and counterfactual analysis comparing agent decisions against baseline behavior. Effective governance for agentic commerce systems requires versioned policy artifacts with formal approval workflows, strict separation between experimentation environments and production execution, regular adversarial testing and red-team exercises targeting decision loop vulnerabilities, and forensic-grade logging enabling complete reconstruction of decision paths when incidents occur. AI-powered incident response systems demonstrate the capability to achieve high accuracy in threat detection while maintaining computational overhead at acceptable levels for production deployment [9]. Research indicates that defense mechanisms introduce training overhead of approximately 25% (increasing training time from 3.0 hours to 3.75 hours) without significantly impacting inference speed, maintaining inference time at 10-11 milliseconds per sample suitable for real-time deployment [6]. The NSL-KDD dataset used for validation contains 125,973 training samples and 22,544 testing samples, while CIC-IDS-2017 comprises approximately 2.8 million samples with 80% allocated for training and 20% for testing, and UNSW-NB15 includes 175,341 training samples and 82,332 testing samples [9]. These governance mechanisms transform autonomy from a risk multiplier into a managed capability where the benefits of agent decision-making can be realized while maintaining appropriate human oversight of aggregate system behavior and individual high-stakes decisions.

Dataset	Usage Context in IDS Research	Traffic Representation	Evaluation Role
NSL-KDD	Legacy benchmark dataset	Simulated attacks	Baseline testing
CIC-IDS-2017	Modern enterprise traffic	Realistic network flow	High-fidelity validation
UNSW-NB15	Hybrid contemporary dataset	Mixed normal/attack	Generalization assessment

Table 2: Benchmark Datasets Referenced for Validation Context [9]

Conclusion

Agentic artificial intelligence fundamentally transforms the security landscape of e-commerce platforms by introducing autonomous decision loops that operate continuously at scale across critical commerce functions. These systems create high-impact control points in discovery, pricing, and routing where small signal manipulations can produce broad effects on customer experience, business outcomes, and operational stability. Traditional application security measures addressing perimeter protection, access control, and data confidentiality remain necessary but prove insufficient for systems where authorized agents make consequential decisions at machine speed. Securing agentic commerce systems requires explicit preservation of decision integrity through architectural separation of proposal and execution, bounded autonomy through least-privilege identity and blast-radius limitation, and robust feedback through observability that detects manipulation and drift. By integrating architectural guardrails with strict execution controls and outcome-oriented observability, platforms can capture the substantial benefits of agentic AI while maintaining the security, trust, and operational resilience that sustainable commerce requires. Future research directions include formal verification methods for decision policies, automated anomaly detection in feedback loops, and industry standards for agentic system certification and audit.

References

[1] Elias Dritsas, Maria Trigka, "Machine Learning in E-Commerce: Trends, Applications, and Future Challenges," IEEE Access, vol. 13, 22 May 2025. Available: https://ieeexplore.ieee.org/document/11009009

[2] Krishna Dusad, "Agentic Commerce: The Paradigm Shift from Human-Mediated to Autonomous AI-Driven Transactions in Digital Payment Systems," International Journal of Computational and Experimental Science and Engineering, 14 November 2025. Available: https://ijcesen.com/index.php/ijcesen/article/view/4304

[3] Shuja Mughal, et al., "Mitigating Safety Risks in Information Systems: A Self-Adaptive Approach," IEEE Access, 22 May 2025. Available: https://ieeexplore.ieee.org/document/11009198

[4] Anan Jin, et al., "DeCoAgent: Large Language Model Empowered Decentralized Autonomous Collaboration Agents Based on Smart Contracts," IEEE Access, 16 October 2024. Available: https://ieeexplore.ieee.org/document/10720018

[5] Muhammad Hataba, et al., "Security and Privacy Issues in Autonomous Systems: A Layer-Based Survey," IEEE Access, 25 April 2022. Available: https://ieeexplore.ieee.org/document/9762777

[6] Hariprasad Holla, et al., "Adversarial Threats to Cloud IDS: Robust Defense With Adversarial Training and Feature Selection," IEEE Access, vol. 13, 05 May 2025. Available: https://ieeexplore.ieee.org/document/10985899

[7] Tai-You Chen, et al., "The Confluence of Evolutionary Computation and Multi-Agent Systems: A Survey," IEEE/CAA Journal of Automatica Sinica, vol. 12, no. 11, pp. 2175-2193, March 2025. Available: https://www.ieee-jas.net/article/doi/10.1109/JAS.2025.125246

[8] Desta Haileselassie Hagos, Danda B. Rawat, "Neuro-Symbolic AI for Military Applications," IEEE Computer Society / AI Journal, December 2024. Available: https://www.computer.org/csdl/journal/ai/2024/12/10638797/1Zxz2zsgpnG

[9] Mohammed Ashfaaq M. Farzaan, et al., "AI-Powered System for an Efficient and Effective Cyber Incident Detection and Response in Cloud Environments," IEEE Transactions on Machine Learning in Communications and Networking, vol. 3, pp. 623-643, 28 April 2025. Available: https://ieeexplore.ieee.org/document/10979487