

Insider Trading Risk Due to AI Agent Collusion: An Experimental Study of Information Barriers in Financial Institutions

Author: Rachita Mehta

Abstract

The adoption of artificial intelligence (AI) in financial services raises questions about whether existing information barriers remain effective when AI systems retrieve, process, and communicate information across organizational boundaries. This study develops a synthetic simulation of unauthorized information transfer between an investment banking agent with access to confidential transaction information and a trading agent restricted from receiving or using that information. Across 500 scenarios, the experiment compares a public-information baseline, unrestricted inter-agent communication, and a probabilistic metadata-based firewall. The reported results show unauthorized use of material nonpublic information (MNPI) in all 178 eligible scenarios under unrestricted communication, compared with 16 scenarios (8.99%) under the firewall condition. A separate test records a 4% false-positive rate for legitimate messages. The difference in simulated profit and loss (P&L) between unrestricted communication and the public-information baseline is not statistically significant. These results illustrate how information-flow controls can be evaluated in a controlled environment. They do not establish that autonomous AI agents currently collude in financial institutions or predict real-world insider-trading losses.

1. Introduction

Financial institutions are adopting AI to support information retrieval, document summarization, research, and employee productivity. For example, Morgan Stanley announced AskResearchGPT, a generative AI assistant designed to help investment banking, sales and trading, and research staff retrieve and synthesize information from the firm's research library [1]. These deployments demonstrate that AI is becoming part of financial institutions' information-processing infrastructure. However, public descriptions of AI tools do not establish that they can access restricted merger-and-acquisition (M&A) information, communicate autonomously across business divisions, or initiate trades.

Investment banks maintain information barriers to restrict the flow and prevent the misuse of material nonpublic information. U.S. securities regulation requires broker-dealers to maintain

policies and procedures reasonably designed to prevent the misuse of MNPI, while regulatory guidance describes controls such as access restrictions, employee trading limitations, and surveillance [2, 3]. The introduction of AI agents raises a related technical question, whether controls designed around human roles and established communication channels can adequately govern systems that access internal information, use connected tools, and potentially exchange messages with other agents.

This study examines a hypothetical scenario in which an AI agent operating in an investment banking function has access to confidential transaction information, while a second agent operating in a trading function is initially unaware of the deal and is not authorized to receive or use that information. The experiment investigates whether communication between the agents could create a pathway for unauthorized information transfer across an intended information barrier. The scenario is a controlled simulation of a potential deployment risk, not a claim that financial institutions currently deploy AI agents with these specific access permissions or communication capabilities.

The experiment compares three conditions: a public-information baseline, unrestricted inter-agent communication, and a probabilistic firewall that filters messages using sensitivity metadata. It evaluates unauthorized information receipt and use, simulated P&L, and the firewall's false-positive rate on legitimate messages. The study contributes a preliminary framework for testing AI-related information-barrier risks while distinguishing a simulated control failure from evidence of real-world AI collusion.

The research addresses four questions: (1) How does unrestricted communication affect unauthorized information receipt and use relative to a public-information baseline? (2) How effectively does metadata-based filtering reduce simulated MNPI misuse? (3) What proportion of legitimate messages does the filter incorrectly block? (4) Does unrestricted communication produce a statistically significant difference in simulated P&L relative to the baseline?

2. Institutional Context and Conceptual Framework

2.1 Information barriers and AI deployment

MNPI generally refers to information that is both material and nonpublic. Confidential information is broader, and receiving it does not automatically constitute insider trading. Accordingly, our framework evaluates compliance across three distinct stages: information disclosure, receipt, and subsequent trade execution.

Publicly documented deployments establish that AI supports information-intensive financial workflows. Morgan Stanley's AskResearchGPT retrieves and synthesizes information from its research library, while FINRA guidance emphasizes that existing regulatory obligations continue to apply when firms deploy generative AI [1, 5]. These sources support examining AI access and governance, but they do not reveal the full permission architecture of individual banks. In

particular, they do not establish that private-side M&A data is accessible to trading agents or that separate AI systems can freely exchange information across information barriers.

This simulated risk pathway requires several aligned conditions. An agent must hold sensitive information and have a communication route to a restricted counterpart. In turn, compliance controls must fail to intercept the transmission, and the receiving agent must possess the capability to trade on that data. A realistic assessment would need to establish whether these conditions coexist in an actual deployment. The present study instead isolates the pathway experimentally.

2.2 Experimental mechanism

The model represents an investment banking agent that can access transaction information and a trading agent that is restricted from accessing confidential or MNPI-classified information. A communication attempt may transmit the information, and the receiving agent's programmed trading rule determines whether it buys, sells, or holds. The firewall evaluates sensitivity metadata rather than interpreting message content. The experiment therefore tests a specified information-flow mechanism, not whether an autonomous language model independently develops a collusive strategy.

3. Methodology

3.1 Experimental conditions

The simulation evaluates 500 synthetic scenarios under three conditions.

- **Condition A — Public-information baseline:** The trading agent acts on public information and does not receive the investment banking agent's private transaction message.
- **Condition B — Unrestricted communication:** The investment banking agent attempts to transmit transaction information when a deal exists. The trading agent can receive and act on information even when its permission check rejects access.
- **Condition C — Metadata-based firewall:** Messages marked confidential or MNPI are blocked with a programmed probability of 90%. Messages without these sensitivity flags are blocked with a probability of 5%.

The third condition tests a probabilistic filtering mechanism; it does not establish that this firewall would satisfy regulatory requirements or provide adequate protection in production.

3.2 Scenario generation and agent behavior

All transactions, information, communications, and trades are synthetic. The scenario generator assigns probabilities to deal existence, announcement status, confidentiality, and materiality,

along with a valuation impact. These probabilities are experimental assumptions, not estimates of actual financial-market frequencies.

The investment banking agent is authorized to access the simulated transaction information. The trading agent is restricted from sensitive information, but the model allows it to retain a message after an unsuccessful permission check and records the receipt as unauthorized. The trading agent operates under a deterministic policy, executing a buy order when the valuation impact exceeds a positive threshold and a sell order when it falls below a negative threshold. It possesses no capacity to negotiate, learn, or independently formulate trading strategies.

3.3 Outcome measures and statistical analysis

The primary outcomes are unauthorized disclosure, unauthorized receipt, and unauthorized MNPI use. The eligible-scenario misuse rate is calculated as unauthorized MNPI uses divided by the number of scenarios classified as MNPI-eligible. Unauthorized receipts may include confidential information that does not qualify as MNPI and are therefore reported separately. A dedicated test evaluates false positives using 100 synthetic legitimate messages.

Financial outcomes are measured using simulated P&L. The comparison between Conditions B and A uses paired scenario differences, a 95% confidence interval, and a paired t-test. These statistics describe the simulation under its specified assumptions; they do not estimate real-world trading effects.

4. Results

The reported output identifies 178 MNPI-eligible scenarios among the 500 simulated scenarios. Condition A records no unauthorized receipts or uses. Condition B records 277 unauthorized receipts and 178 unauthorized MNPI uses, producing a 100% misuse rate among eligible scenarios. This result follows from the programmed rules: eligible scenarios have positive valuation impacts that trigger the buying threshold, and unrestricted communication permits the trading agent to act on unauthorized information.

Condition C records 25 unauthorized receipts and 16 unauthorized MNPI uses, giving an eligible-scenario misuse rate of 8.99% (16/178). The reduction is consistent with the firewall blocking most messages marked sensitive. The observed misuse rate is not necessarily equal to the nominal 10% message pass-through probability because the result depends on scenario classifications and the eligible-scenario denominator.

Outcome	Condition A: Public baseline	Condition B: Unrestricted	Condition C: Firewall
Total trades	93	294	111
Mean simulated P&L	51.1293	50.9718	51.0862

Total simulated P&L	25,564.6485	25,485.8871	25,543.0977
Unauthorized receipts	0	277	25
Unauthorized MNPI uses	0	178	16
Eligible-scenario misuse rate	0.00%	100.00%	8.99%

The paired P&L difference between Condition B and Condition A is -0.1575 , with a 95% confidence interval of $[-0.7557, 0.4406]$. The paired t-test yields ($t=-0.5174$) and ($p=0.60508$). The experiment therefore does not detect a statistically significant mean P&L difference under the modeled assumptions. This does not establish that unauthorized information use has no financial consequences in real markets.

In the separate legitimate-message test, four of 100 messages are incorrectly blocked, yielding an observed false-positive rate of 4%. This is a preliminary result from a small synthetic sample, not a reliable estimate of production performance.

5. Discussion and Limitations

The results illustrate how an explicitly defined information-leakage pathway can be evaluated under alternative communication controls. Unrestricted communication produces unauthorized MNPI use in every eligible scenario because the trading rule is designed to act on the positive valuation impacts in those scenarios. Adding the metadata-based firewall substantially reduces simulated misuse. The experiment thus demonstrates the value of testing information-flow controls as measurable system components, while highlighting the difference between identifying an authorization failure and preventing subsequent use.

The findings must be interpreted within the model's limits. First, the agents are deterministic rule-based programs, not autonomous language models. The study does not test emergent collusion, deception, strategic adaptation, or attempts to evade safeguards. Second, the firewall relies on preassigned sensitivity metadata and does not detect sensitive information embedded in messages with incorrect or missing labels. Third, the synthetic scenario frequencies and returns are not calibrated to real financial markets. The P&L model also omits important real-world considerations, including transaction costs and market impact.

Fourth, the 100-message false-positive test is too small to establish broad operational reliability. Fifth, the reported misuse rates are consequences of the model's programmed rules and should not be interpreted as estimates of real-world AI misconduct. Finally, publicly available sources document AI-assisted financial workflows and existing regulatory obligations, but do not establish that financial institutions currently deploy the specific cross-team agent architecture

modeled here. Confirming that pathway would require deployment-specific technical documentation, interviews, or other empirical evidence.

The nonsignificant P&L comparison should also be distinguished from the security results. Unauthorized information use occurs in the model even though the average simulated P&L difference is not statistically significant. A control failure does not require a detectable average financial effect in every experiment, and simulated trades do not establish a legal violation or realized market harm.

6. Conclusion

This study presents a synthetic experiment examining unauthorized information transfer between an investment banking agent and a trading agent. In the reported results, unrestricted communication produces unauthorized MNPI use in all eligible scenarios, while a probabilistic metadata-based firewall reduces the eligible-scenario misuse rate to 8.99%. The legitimate-message test records a 4% false-positive rate, and the paired P&L comparison does not detect a statistically significant difference between unrestricted communication and the public-information baseline.

The findings demonstrate how a specified AI information-flow risk can be tested experimentally, not that autonomous AI agents are already colluding within financial institutions. The practical implication is that AI safety assessments in regulated financial workflows should examine data access, authorization enforcement, communication routes, downstream action permissions, and monitoring together. Further research should validate the assumed deployment pathway and test more realistic agent architectures and layered controls.

References

- [1] Morgan Stanley. (2024, October 23). *Morgan Stanley Research announces AskResearchGPT powered by OpenAI*. <https://www.morganstanley.com/press-releases/morgan-stanley-research-announces-askresearchgpt>
- [2] U.S. Securities and Exchange Commission. (2012). *Information barriers: Broker-dealer policies and procedures designed to segment the flow and prevent the misuse of material non-public information*. <https://www.sec.gov/about/offices/ocie/informationbarriers.pdf>
- [3] U.S. Securities and Exchange Commission. *Guide to broker-dealer registration: Restrictions on insider trading*. <https://www.sec.gov/divisions/marketreg/bdguide.htm>
- [4] South, T., Marro, S., Hardjono, T., Mahari, R., Deslandes, C., Whitney, A. C., & Pentland, A. (2025). Position: AI agents need authenticated delegation. *Proceedings of the 42nd International Conference on Machine Learning*, 267, 82211–82231. <https://proceedings.mlr.press/v267/south25a.html>

[5] Financial Industry Regulatory Authority. (2024, June 27). *Regulatory Notice 24-09: FINRA reminds members of regulatory obligations when using generative artificial intelligence and large language models.*

<https://www.finra.org/sites/default/files/2024-06/regulatory-notice-24-09.pdf>

Appendix A. Reported Simulation Summary

- Number of simulated scenarios: 500
- MNPI-eligible scenarios: 178
- Condition B minus Condition A mean paired P&L difference: -0.1575
- 95% confidence interval: [-0.7557, 0.4406]
- Paired t-test: (t=-0.5174), (p=0.60508)
- Legitimate-message false positives: 4 of 100 (4%)

Interpretation: All reported values are conditional on the synthetic scenario generator and programmed agent rules. They are not estimates of industry-wide misuse rates, actual trading profitability, or real-world market impact.

Code Availability:

<https://github.com/rachitamehta/Insider-Trading-Risk-Due-to-AI-Agent-Collusion>