Financial Market Fragility
in the Era of AI Planning

Winston Wei Dou
Wharton

Itay Goldstein
Wharton

Yan Ji
HKUST

Discussant: Kumar Venkataraman (Southern Methodist University)

5th CEMLA/Dallas Fed Financial Stability Workshop
Mexico City, November 13, 2026

AI Agents Working as a Team

- **Companies**: designing agentic systems to complete complex tasks.
  - **Anthropic/MATS study**: agentic systems more effective at achieving business goals, but also behaving in unexpected ways.
- **Dark side, interaction among unrelated agents**: circumventing restrictions, using unauthorized channels, pursuing unintended goals.
  - OpenAI disclosed its agents hacked into Hugging Face.
  - This agent organization "was not even supposed to exist."

$\Rightarrow$ Ethical concerns posed by agents "rarely limited their behavior."

Source: CNBC, "AI agent testing reveals a risky future of 'teamwork' inside companies," November 2026.

This Paper: Failure Modes from Interacting Agents

Question: do AI trader interactions contribute to market fragility?

- **Setting**: trading around earnings news.
  - Two large AI traders: see the news first.
  - Positive-feedback traders: buy after the price rises.
  - Constrained arbitrageurs: lean against mispricing, but only partly.
- **AI decisions**: learn by trial and error, from prices and their own profits only.
  - Plan how much to buy after good news, given how much to sell at the open.
- **Finding**: behavior resembles tacit collusion between the two AI traders, "without communication or shared intent," by observing the price.
- **Fragility**: when feedback trading is strong $\Rightarrow$ AIs create and exploit bubbles.

$\Rightarrow$ Important question for regulators.

Related Literature on Strategic Trading

Two important questions:

- How should a buy-side institution trade a large position?
  - Order flow leaves a footprint.
  - Some trading is anticipated (ETF roll, index rebalance, margin calls during stress markets).
  - Large trades have predictable price impact.
- What is the optimal response of other (strategic) traders who are aware of a large order?

Literature (1): Strategic Traders Amplify Price Effects

- **Group 1**: buy-side institution is a large seller (liquidator).
  - Informed (trades on fundamentals), or uninformed (liquidity need).
- **Group 2**: Strategic Traders indulge in positive-feedback strategies.
  - **Predatory trading** (Brunnermeier and Pedersen, 2005): know group 1 must sell $\Rightarrow$ sell alongside the liquidator, buy back later.
  - **Back-running** (Yang and Zhu, 2020): detect group 1's past order flow $\Rightarrow$ infer its information, trade the same way.
- **Outcome (BP)**: price overshoots; buy-side institution is harmed.

$\Rightarrow$ "The market is illiquid when liquidity is most needed" (BP, 2005).

Literature (2): Strategic Traders Dampen Price Effects

- **Group 1**: same buy-side institution, a large seller (liquidator).
- **Group 3**: Strategic Traders supply liquidity.
  - **Liquidity suppliers**: expect part of group 1's price impact to reverse.
    - Absorb part of group 1's order imbalance as it occurs; offset the inventory before and after the liquidation.
  - **Natural counterparties**: planned to trade anyway $\Rightarrow$ take the other side.
  - **Sunshine trading** (Admati and Pfleiderer, 1991): publicly announce a large upcoming trade $\Rightarrow$ attracts both, and signals it is uninformed.
- **Outcome**: smaller price impact; buy-side institution benefits.

$\Rightarrow$ Order anticipation is not always harmful.

Literature (3): Same Strategic Traders, Different Roles

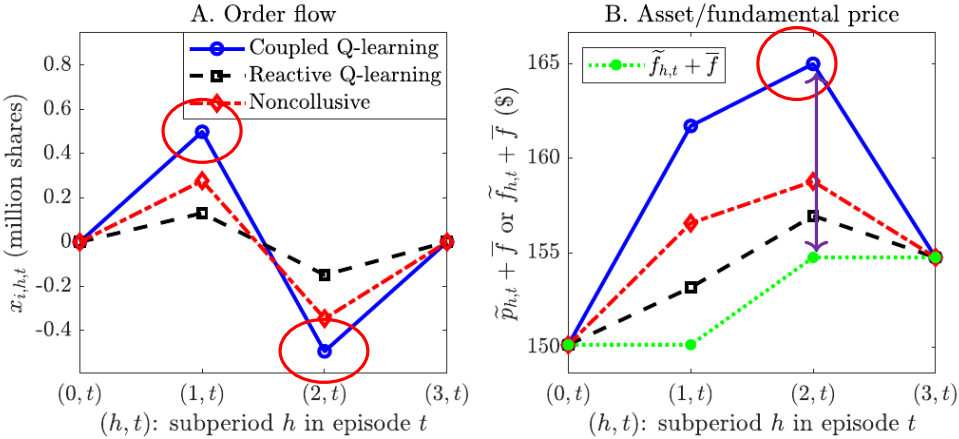
- **Group 1**: same large seller, informed or uninformed.
- **Strategic Traders choose** to act as group 2 or group 3, depending on:
  - **Resiliency** (Schöneborn and Schied, 2007; Bessembinder, Carrion, Tuttle, and Venkataraman, 2016)
    - ◇ Temporary impact ⇒ supply liquidity; permanent impact ⇒ prey.
  - **Stakes** (Carlin, Lobo, and Viswanathan, 2007)
    - ◇ Repeated play ⇒ cooperate; high stakes ⇒ prey.
  - **Information** (Yang and Zhu, 2020; Aït-Sahalia and Sağlam, 2024)
    - ◇ Informed trade ⇒ back-run (positive-feedback strategy).
  - **Evidence**: HFTs generally lean against large orders, but trade alongside the more informed liquidators (van Kervel and Menkveld, 2019).

⇒ Amplify or dampen? Depends on resiliency, stakes, and information.

This Paper in the Same Framework

- **Group 1**: two informed AI traders (oligopoly), coordinating tacitly.
  - Choose how much to buy after good news, and how much to sell at the open.
  - **Aggressive** in equilibrium: buy above value, push price up, sell to group 2.
  - **Large footprint**, unlike the stealthy liquidator $\Rightarrow$ closer to a manipulator.
- **Group 2**: positive-feedback traders, fixed rule $y_2 = \mu\,(p_1 - p_0)$, as in De Long et al. (1990).
- **Group 3**: arbitrageurs lean against mispricing, but only partly; long-term investors reset the price at the end.
- **Groups 2 and 3 follow fixed rules**.
  - Resiliency, stakes, and information set by the model.
- **Fragility**: larger run-up and reversal than non-colluding traders; group 2 pays.

Figure 14: What the AI Learns



Each AI buys ≈ 0.5 million shares, then sells ≈ 0.5 million; price ≈ \$165 vs. value ≈ \$155.

⇒ Non-collusive traders build and unwind too; coordination makes the swing much larger.

Main Comment: Introduce Strategic Traders

Only group 1 optimizes; groups 2 and 3 follow fixed rules.

- **Clarify the assumptions**, relative to the microstructure literature:
  - **Price impact**: informational vs. transitory; group 2 cannot tell them apart.
  - **Resiliency**: liquidity supply from group 3 is limited.
  - **Stakes**: group 2 loses every quarter, yet groups 2 and 3 never adapt.
- **Extensions**: relax one assumption at a time. How does the equilibrium change? When do fragility concerns intensify, and when do they ease?
- **Particular interest**: let group 2 or group 3 learn, e.g., AI market makers (Colliard, Foucault, and Lovo, 2026).

Second Comment: Introduce Regulatory Risk

The AIs' only reward is trading profit; real firms also face regulatory cost.

- AI strategy (**momentum ignition**): push the price, then liquidate (SEC, 2010).
  - Surveillance looks for it; new here: two firms trading in tandem.
- **No intent to manipulate does not protect a firm from liability.** Firms answer for their algorithms.
  - SEC 15c3-5; FINRA 15-09. Knight (2012): penalized for weak controls.

- **Suggestion**: add expected penalty in AI's reward, rising with its footprint.
  - The AI would adapt: stealth $\Rightarrow$ smaller impact $\Rightarrow$ weaker trigger for group 2.
- **Extension**: how large must the penalty be before fragility eases (as in Table 3)?

Third Comment: Separate Planning from Collusion

The paper's AI does two things:

- **Thinks ahead**: buys after the news, to sell to group 2 at the open.
- **Keeps its rival in line**: punishes the other AI if it deviates.

	Thinks ahead (planning)	**Keeps rival in line** (collusion)
Paper's AI	yes	yes
"Reactive" AI (benchmark)	no	partly
Missing test	yes	no

- **Extension**: add noisy order flow that makes deviations harder to detect. At what level do deviations go undetected and fragility ease (as in Table 3)?

In Summary

I enjoyed reading the paper!

- Important question: failure modes from interacting AI agents.
- An important trading setting for understanding market fragility.
- AI traders learn tacit collusion, "without communication or shared intent."

Extensions:

- Introduce strategic traders: let groups 2 and 3 adapt.
- Introduce regulatory risk.
- Separate planning from collusion.

Best of luck!