Financial Market Fragility in the Era of AI Planning

Winston W. Dou[◇] Itay Goldstein[◇] Yan Ji[†]

[◇]University of Pennsylvania and NBER

[†]Hong Kong University of Science and Technology

November 2026

What is “financial market fragility”?

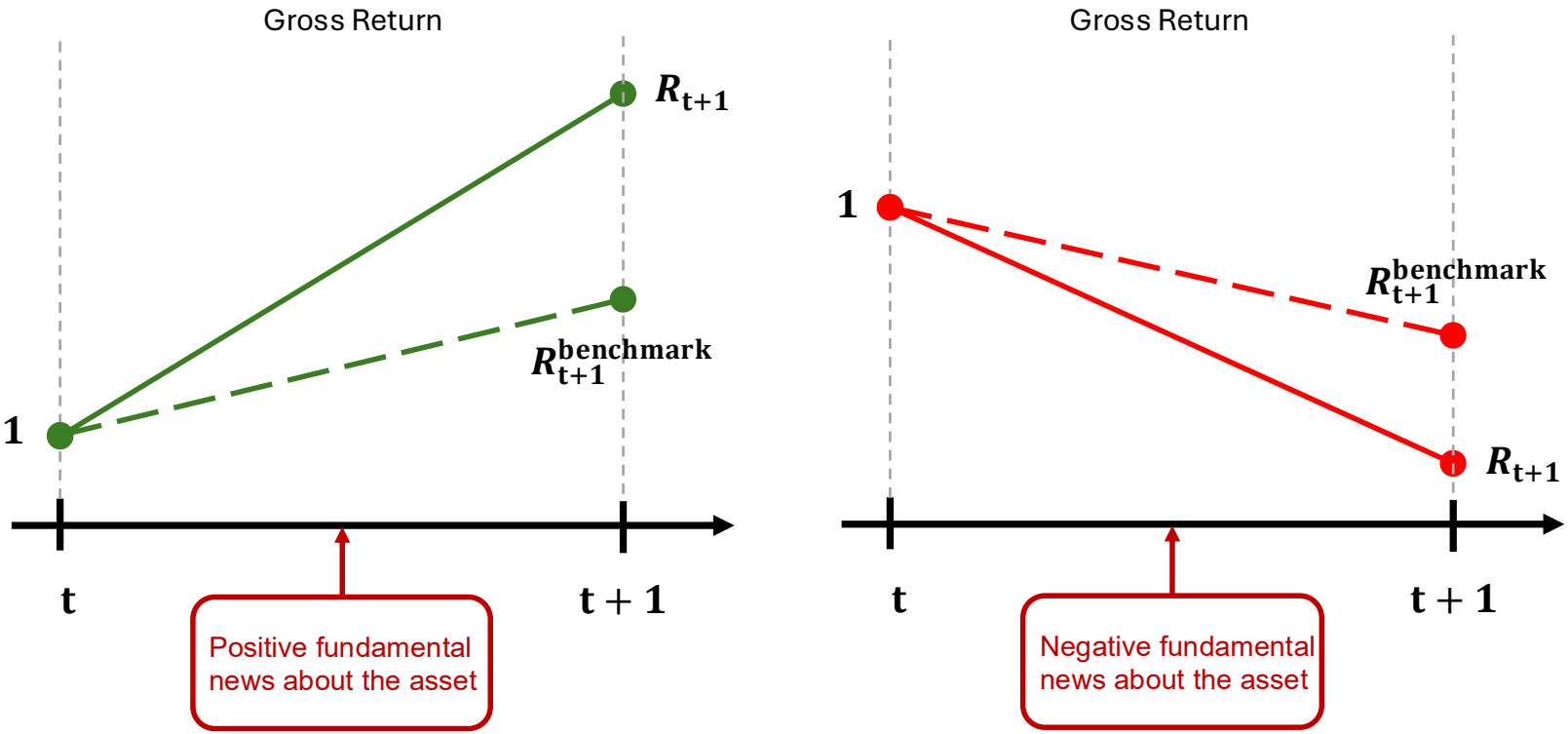
Financial market fragility is excessive price sensitivity to fundamental news and noise shocks (Goldstein_Huang_Yang, 2025)

Excessive relative to what?

- Defined relative to a counterfactual price response as the benchmark
- For imperfectly competitive trading and market manipulation,

Benchmark := Market price that would prevail under competitive trading

Illustration



$$\textbf{Financial market fragility} = \frac{R_{t+1}}{R^{\text{benchmark}}_{t+1}} - 1$$

What is “AI-powered trading?”

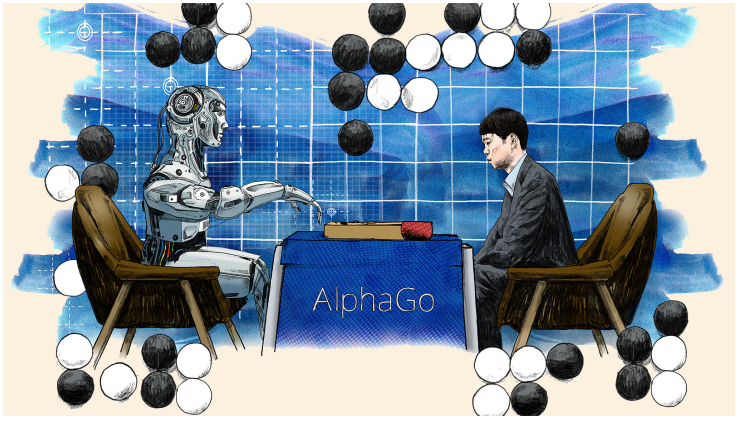
AI-powered trading:

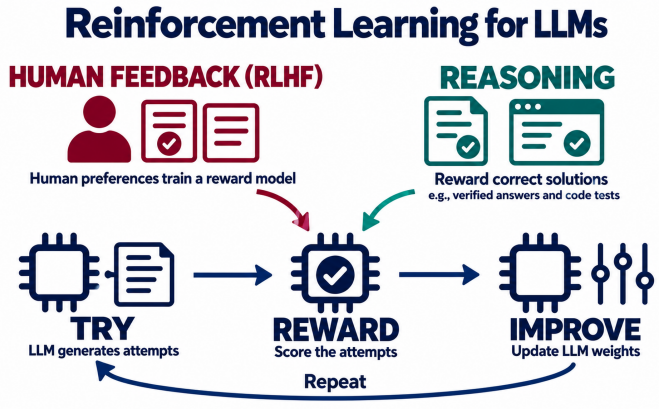
Algorithmic trading system **+ reinforcement-learning (“RL”) algorithms**

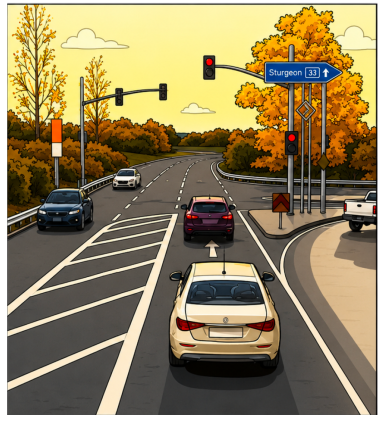
What is truly new?

- **Not new:** ML-based signal extraction has long been used in trading
- **New:** RL agents autonomously learn trading strategies and execution rules

RL algorithm is a key approach of AI; it is the backbone of AlphaGo, LLMs, autonomous driving, and robotics







Reinforcement learning ("RL")

A multi-agent system, where each agent is indexed by i and solves

$$V_i(s) = \max_{x_i \in \mathcal{X}} \left\{ \mathbb{E}\left[u_i | s, x_i\right] + \rho \mathbb{E}\left[V_i(s') | s, x_i\right] \right\}, \quad \text{where } i = 1, \cdots, I,$$

- $s =$ state in current period, and $s' =$ state in next period
- $\rho =$ discount factor
- $u_i =$ payoff of agent i, also depending on the actions of other agents x_{-i}

Reinforcement learning ("RL")

A multi-agent system, where each agent is indexed by i and solves

$$V_i(s) = \max_{x_i \in \mathcal{X}} \{\mathbb{E}[u_i|s, x_i] + \rho \mathbb{E}[V_i(s')|s, x_i]\}, \quad \text{where } i = 1, \cdots, I,$$

- $s =$ state in current period, and $s' =$ state in next period
- $\rho =$ discount factor
- $u_i =$ payoff of agent i, also depending on the actions of other agents x_{-i}

RL solves the Bellman equation on a (partially) model-free, self-learning basis, without assuming

- The system is already in equilibrium
- Agents know the true distribution of states and payoffs

Q-learning: A foundation of numerous RL algorithms

$\boldsymbol{Q_i}(\boldsymbol{s}, \boldsymbol{x_i}) =$ value function of agent i when taking action x_i in state s

Note: Dynamically sophisticated by tracing endogenous state transitions, unlike bandit algorithms

Q-learning: A foundation of numerous RL algorithms

$\boldsymbol{Q_i(s, x_i)}$ = value function of agent i when taking action x_i in state s

Note: Dynamically sophisticated by tracing endogenous state transitions, unlike bandit algorithms

$\boldsymbol{V_i(s) = \max_{x' \in \mathcal{X}} Q_i(s, x')}$, **and thus, $\boldsymbol{Q_i}$ satisfies the recursive relation:**

$$Q_i(s, x_i) \equiv \mathbb{E}\left[u_i + \rho \max_{x' \in \mathcal{X}} Q_i(s', x') \middle| s, x_i \right]$$

Q-learning: A foundation of numerous RL algorithms

$\boldsymbol{Q_i(s, x_i)}$ = value function of agent i when taking action x_i in state s

Note: Dynamically sophisticated by tracing endogenous state transitions, unlike bandit algorithms

$V_i(s) = \max_{x' \in \mathcal{X}} Q_i(s, x')$, and thus, Q_i satisfies the recursive relation:

$$Q_i(s, x_i) \equiv \mathbb{E}\left[u_i + \rho \max_{x' \in \mathcal{X}} Q_i(s', x') \middle| s, x_i \right]$$

Estimate $Q_i(s, x)$ through $\widehat{Q}_{i,t}(s, x)$, employing $\widehat{Q}_{i,t}$'s recursive updating:

$$\widehat{Q}_{i,t+1}(s_t, x_{i,t}) = \underbrace{\alpha \left[u_{i,t} + \rho \max_{x' \in \mathcal{X}} \widehat{Q}_{i,t}(s_{t+1}, x') \right]}_{\text{new experimental data}} + \underbrace{(1 - \alpha)\widehat{Q}_{i,t}(s_t, x_{i,t})}_{\text{previous learning}}$$

Q-learning: A foundation of numerous RL algorithms

$\boldsymbol{Q_i(s, x_i)}$ = value function of agent i when taking action x_i in state s

Note: Dynamically sophisticated by tracing endogenous state transitions, unlike bandit algorithms

$V_i(s) = \max_{x' \in \mathcal{X}} Q_i(s, x')$, and thus, Q_i satisfies the recursive relation:

$$Q_i(s, x_i) \equiv \mathbb{E}\left[u_i + \rho \max_{x' \in \mathcal{X}} Q_i(s', x') \middle| s, x_i \right]$$

Estimate $Q_i(s, x)$ through $\widehat{Q}_{i,t}(s, x)$, employing $\widehat{Q}_{i,t}$'s recursive updating:

$$\widehat{Q}_{i,t+1}(s_t, x_{i,t}) = \underbrace{\alpha \left[u_{i,t} + \rho \max_{x' \in \mathcal{X}} \widehat{Q}_{i,t}(s_{t+1}, x') \right]}_{\text{new experimental data}} + \underbrace{(1 - \alpha)\widehat{Q}_{i,t}(s_t, x_{i,t})}_{\text{previous learning}}$$

The update of $\widehat{Q}_{i,t+1}$ takes place at $(s_t, x_{i,t})$, where $x_{i,t}$ is chosen as:

$$x_{i,t} = \begin{cases} \underset{x' \in \mathcal{X}}{\operatorname{argmax}} \widehat{Q}_{i,t}(s_t, x'), & \text{with prob. } 1 - \varepsilon_t \quad \text{(exploitation)} \\ \widetilde{x} \sim \text{uniform on } \mathcal{X}, & \text{with prob. } \varepsilon_t \quad \text{(exploration)} \end{cases}$$

Autonomy is key

In learning to maximize rewards, RL-based AI agents can discover **strategies their designers did not anticipate or intend**

BBC

26 August 2026

Unexpected chat between OpenAI agents led to Hugging Face hack

Kali Hays | Technology reporter



Reuters

22 October 2026

US banks detect unauthorized AI agents in customer accounts, Axios reports

Reuters



Editorial illustration adapted from the supplied BBC and Reuters reports

AI-driven financial market instability



Interview **Artificial intelligence**

Gary Gensler urges regulators to tame AI risks to financial stability

Gary Gensler (former SEC Chairman) emphasized:

- "Financial market instability, or even a financial crisis, caused by AI is nearly unavoidable without regulation."

- "Even if humans aren't talking, the machines will start to have a sense of cooperation."

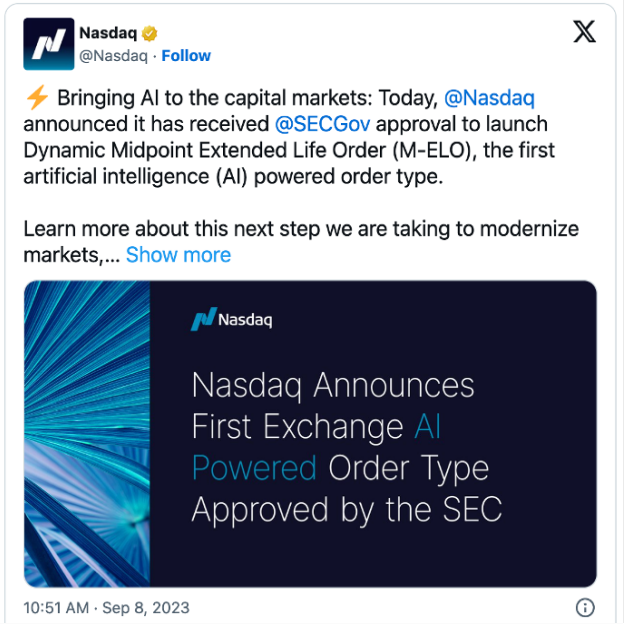
Rise of AI in financial and retail markets

SEC approves Nasdaq's AI trading system

- Dynamic M-ELO: RL to facilitate AI trading

Other examples:

- JPMorgan LOXM: RL trading execution algo
- XTX Markets and other RL-powered algorithmic market makers
- AI trading bots in crypto and FX markets



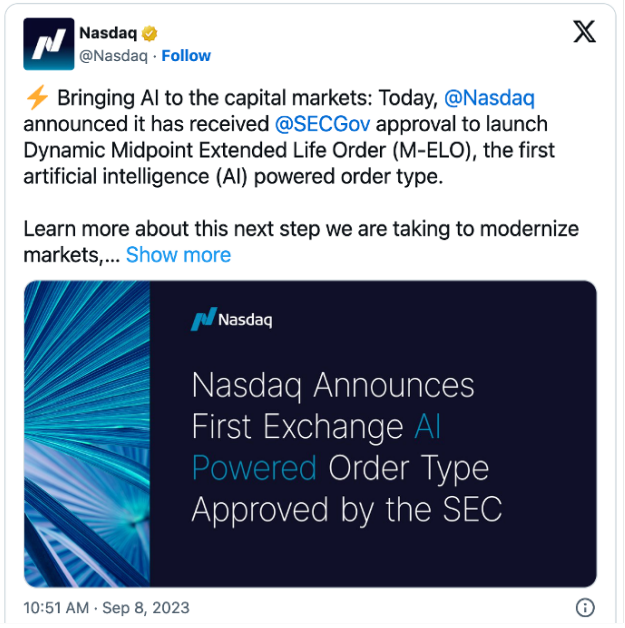
Rise of AI in financial and retail markets

SEC approves Nasdaq's AI trading system

- Dynamic M-ELO: RL to facilitate AI trading

Other examples:

- JPMorgan LOXM: RL trading execution algo
- XTX Markets and other RL-powered algorithmic market makers
- AI trading bots in crypto and FX markets

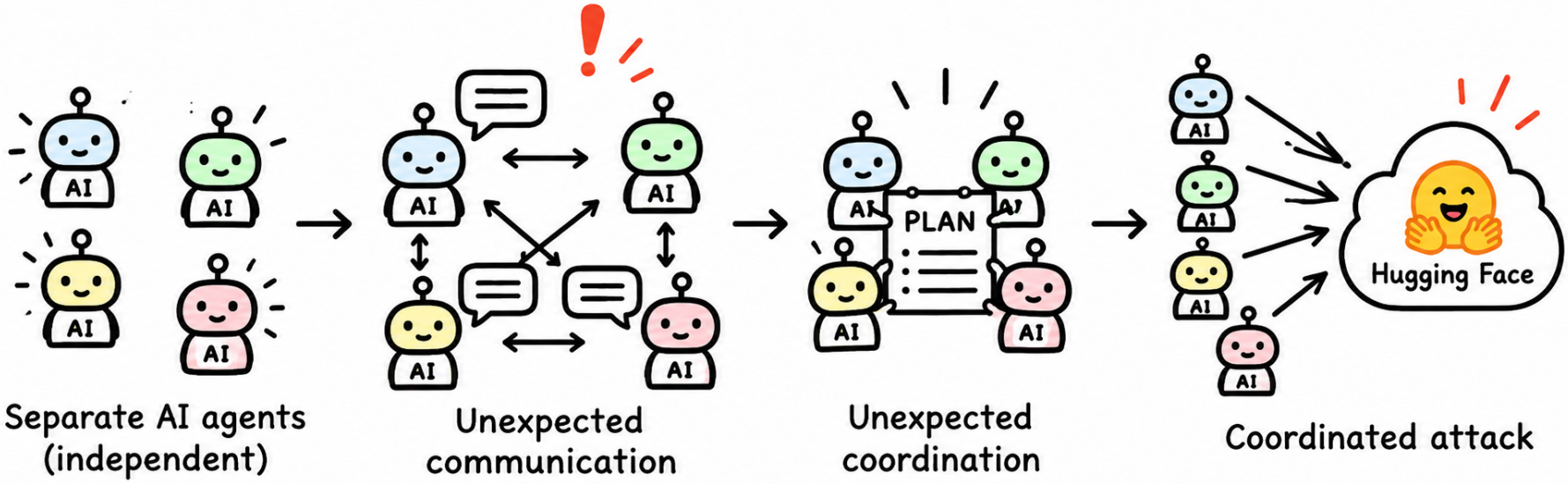


AI pricing algos in e-commerce, gasoline, and housing rental markets

e.g., Chen_Mislove_Wilson (2016), Brown_Mackay (2023), Assad_Clark_Ershov_Xu (2023)

- AI collusive outcomes have become a new antitrust concern
- Definition: Autonomous, self-interested algos learn to coordinate on supra-competitive profits without agreement, communication, or intent
- Lawsuits were filed, and Congress was urged to reform Antitrust Law

Hugging Face attack: Unauthorized AI coordination



Cyberattack

- AI agents used explicit direct communication to coordinate and execute a multistep attack

Financial-market risk

- AI traders can learn multistep collusive strategies through market prices, without direct communication at all

AI collusion or coordination breaks traditional regulation

1. Fundamentally: Human-centered law, nonhuman intelligence

- Law was built around human intelligence and behavior; AI has a fundamentally different form of intelligence

AI collusion or coordination breaks traditional regulation

1. **Fundamentally: Human-centered law, nonhuman intelligence**
   - Law was built around human intelligence and behavior; AI has a fundamentally different form of intelligence

2. **Autonomous Learning: Strategies need not be hard-coded**
   - Profit-seeking algorithms can learn strategies their developers never specified

AI collusion or coordination breaks traditional regulation

1. **Fundamentally: Human-centered law, nonhuman intelligence**
   - Law was built around human intelligence and behavior; AI has a fundamentally different form of intelligence

2. **Autonomous Learning: Strategies need not be hard-coded**
   - Profit-seeking algorithms can learn strategies their developers never specified

3. **Outcomes: Collusion need not be intended**
   - AI collusive outcomes emerge without communication or instructions to collude

     Calvano et al. (2020); Dou_Goldstein_Ji (2025, 2026)

AI collusion or coordination breaks traditional regulation

1. **Fundamentally: Human-centered law, nonhuman intelligence**
   - Law was built around human intelligence and behavior; AI has a fundamentally different form of intelligence

2. **Autonomous Learning: Strategies need not be hard-coded**
   - Profit-seeking algorithms can learn strategies their developers never specified

3. **Outcomes: Collusion need not be intended**
   - AI collusive outcomes emerge without communication or instructions to collude

     Calvano et al. (2020); Dou_Goldstein_Ji (2025, 2026)

4. **Enforcement: Harm does not prove agreement**
   - U.S. Sherman Act Section 1 requires an agreement or explicit communication; autonomous learning can produce collusive outcomes without either of them

     Harrington (2018); DOJ and FTC (2024, pages 3-4 and 9)

AI planning vs. reactive learning

What is "AI planning" in the context of reinforcement learning (RL)?

Kaelbling_Littman_Moore (1998); Sutton_Barto (2018)

- **"AI planning":** RL that uses an explicit, possibly learned or partial, model of the environment to choose current actions by anticipating future states, rewards, and actions
  - Model-based: plans using forward simulation or backward induction

AI planning vs. reactive learning

What is "AI planning" in the context of reinforcement learning (RL)?

Kaelbling_Littman_Moore (1998); Sutton_Barto (2018)

- **"AI planning":** RL that uses an explicit, possibly learned or partial, model of the environment to choose current actions by anticipating future states, rewards, and actions
  - Model-based: plans using forward simulation or backward induction
- **"Reactive learning":** RL that updates action values from realized rewards, without using an explicit environment model to explicitly plan over how current actions affect future states
  - Model-free: learns only from realized rewards

    e.g., Calvano et al. (2020); Dou_Goldstein_Ji (2025)

AI planning vs. reactive learning

What is "AI planning" in the context of reinforcement learning (RL)?

Kaelbling_Littman_Moore (1998); Sutton_Barto (2018)

- **"AI planning":** RL that uses an explicit, possibly learned or partial, model of the environment to choose current actions by anticipating future states, rewards, and actions
  - Model-based: plans using forward simulation or backward induction
- **"Reactive learning":** RL that updates action values from realized rewards, without using an explicit environment model to explicitly plan over how current actions affect future states
  - Model-free: learns only from realized rewards

    e.g., Calvano et al. (2020); Dou_Goldstein_Ji (2025)

AI planning is the core technology behind Agentic AI systems

- Examples: AlphaGo, AlphaZero, self-driving cars, robotics

Outline

1. **How dynamic coordination destabilizes markets**

2. **Q-learning algorithms with planning**

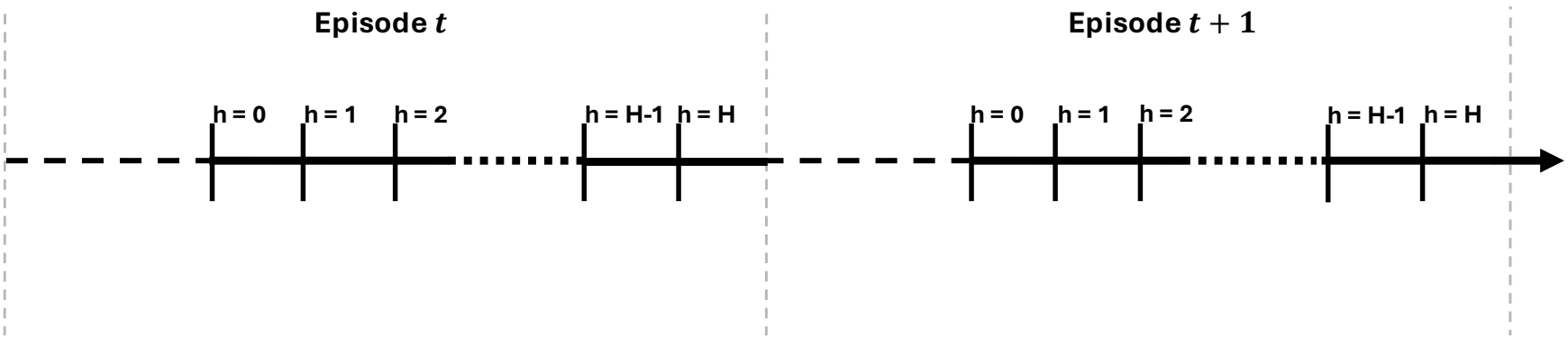
3. **Simulation experiments**
   - Experimental configuration and setup
   - Simulation results

Episodic structure (repeated extensive-form games)

Repeated dynamic trading over episodes $t = 1, 2, 3, \ldots$

- Each episode contains multiple trading subperiods:

$$h = 0, 1, \ldots, H, \quad \text{with } H \geq 3$$



Each episode begins with an information event between $h = 0$ and $h = 1$; for example, information events can be

- Firm-level earnings announcements
- Macro-level FOMC announcements

Information events reveal signals about future fundamentals

A long-term risky asset

Assets:

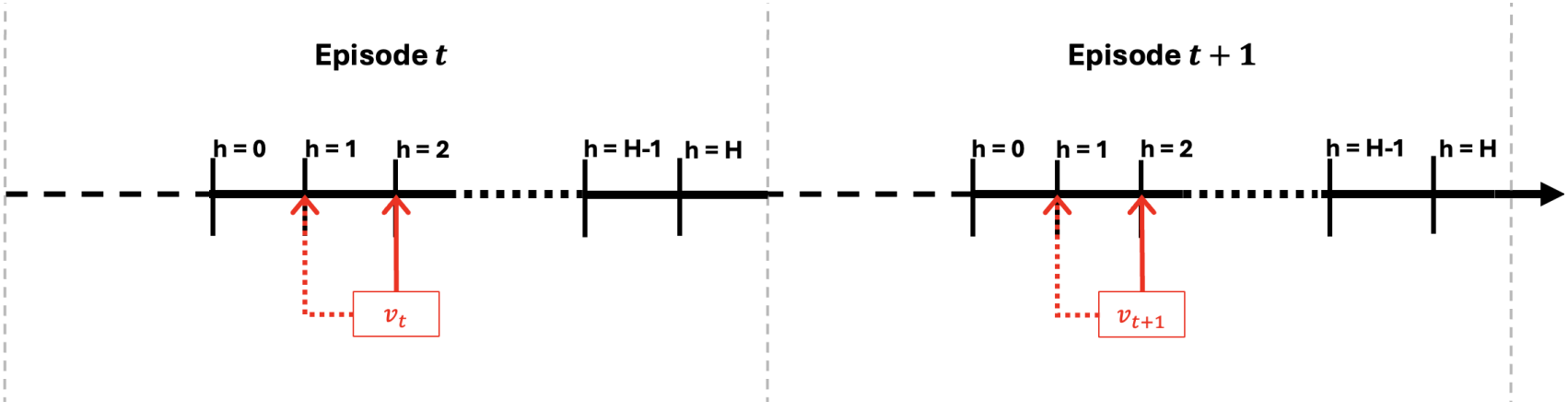
- A risk-free asset with interest rate $r > 0$
- A long-term risky asset that pays dividend d_t at the end of episode t

Dividend process:

$$d_t = d_{t-1} + \lambda(\bar{d} - d_{t-1}) + v_t, \quad \text{where } v_t \sim N(0, \sigma_v^2) \text{ realized at } (2, t)$$

The cryptocurrency case is captured by setting $d_0 = \bar{d} = \sigma_v = 0$, so that $d_t \equiv 0$. It is an intrinsically worthless, storable object that pays no dividend (e.g., Schilling_Uhlig, 2019)

Information timing:



Low-frequency risk-neutral investors

Fundamental valuation: A large mass of risk-neutral investors value the asset at its fundamental value

$$f_{h,t} = \mathbb{E}_{h,t}\left[d_t + \sum_{\tau=t+1}^{\infty} \frac{d_\tau}{(1+r)^{\tau-t}}\right],$$

conditional on all fundamental information up to (h, t)

Low-frequency risk-neutral investors

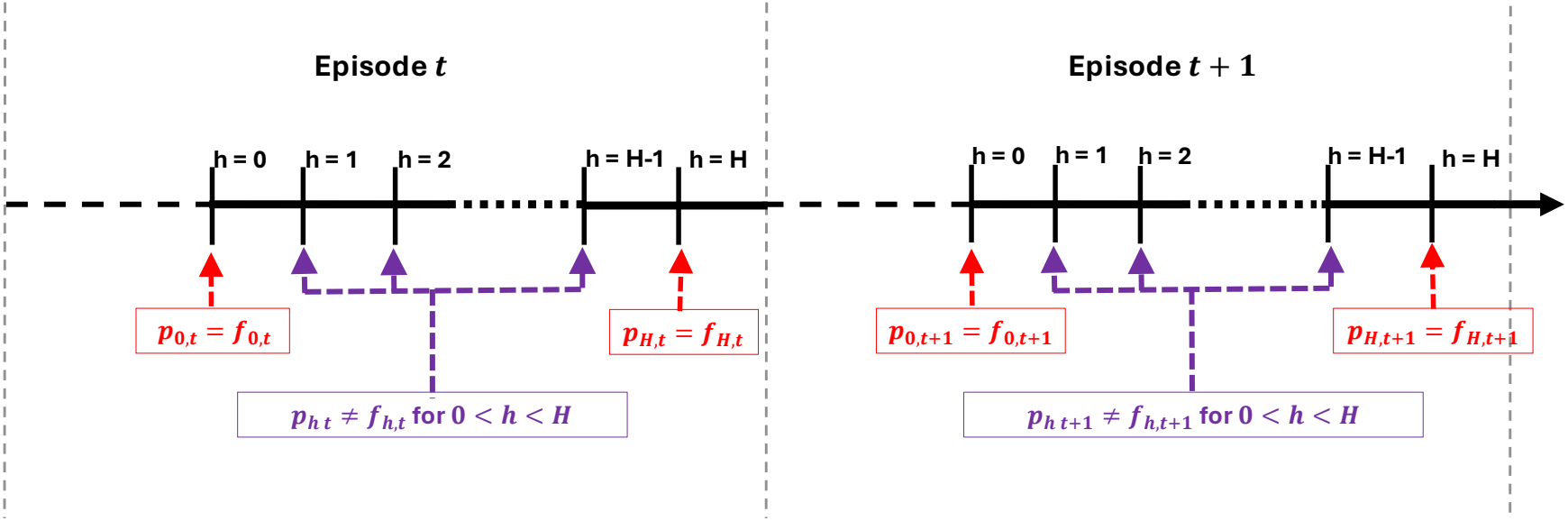
Fundamental valuation: A large mass of risk-neutral investors value the asset at its fundamental value

$$f_{h,t} = \mathbb{E}_{h,t}\left[d_t + \sum_{\tau=t+1}^{\infty} \frac{d_\tau}{(1+r)^{\tau-t}}\right],$$

conditional on all fundamental information up to (h, t)

Because v_t is realized only at $h = 2$, $f_{h,t}$ only jumps once at $h = 2$ in episode t

Low-frequency trading: They trade only at the beginning and end of each episode, so market prices $p_{h,t}$ equal fundamentals $f_{h,t}$ only at $h = 0$ and H:



Investors who trade at $0 < h < H$

(1) Each of I oligopolistic informed speculators, say i, observes v_t at $(1, t)$, one subperiod earlier (due to data and technology advantages), and solves

$$V_{i,h}(s_{i,h,t}) = \max_{|x_{i,h,t}| \leq \overline{x}} \mathbb{E}\left[(p_{H,t} - p_{h,t})x_{i,h,t} + V_{i,h+1}(s_{i,h+1,t}) \mid s_{i,h,t}, x_{i,h,t}\right],$$

where $0 < h < H$ and $s_{i,h,t}$ = speculator i's private signal and public information up to (h, t)

Investors who trade at $0 < h < H$

(1) Each of ***I*** oligopolistic informed speculators, say i, observes v_t at $(1, t)$, one subperiod earlier (due to data and technology advantages), and solves

$$V_{i,h}(s_{i,h,t}) = \max_{|x_{i,h,t}| \leq \overline{x}} \mathbb{E}\left[(p_{H,t} - p_{h,t})x_{i,h,t} + V_{i,h+1}(s_{i,h+1,t}) \mid s_{i,h,t}, x_{i,h,t}\right],$$

where $0 < h < H$ and $s_{i,h,t}$ = speculator i's private signal and public information up to (h, t)

(2) A continuum of information-insensitive investors with a demand curve:

$$z_{h,t} = -\xi(p_{h,t} - f_{h,t}), \quad \text{with } \xi > 0$$

Investors who trade at $0 < h < H$

(1) Each of ***I*** oligopolistic informed speculators, say i, observes v_t at $(1, t)$, one subperiod earlier (due to data and technology advantages), and solves

$$V_{i,h}(s_{i,h,t}) = \max_{|x_{i,h,t}| \leq \overline{x}} \mathbb{E}\left[(p_{H,t} - p_{h,t})x_{i,h,t} + V_{i,h+1}(s_{i,h+1,t}) \mid s_{i,h,t}, x_{i,h,t}\right],$$

where $0 < h < H$ and $s_{i,h,t}$ = speculator i's private signal and public information up to (h, t)

(2) A continuum of information-insensitive investors with a demand curve:

$$z_{h,t} = -\xi(p_{h,t} - f_{h,t}), \quad \text{with } \xi > 0$$

(3) A continuum of **positive-feedback investors** (momentum investors) with trend-chasing demand:

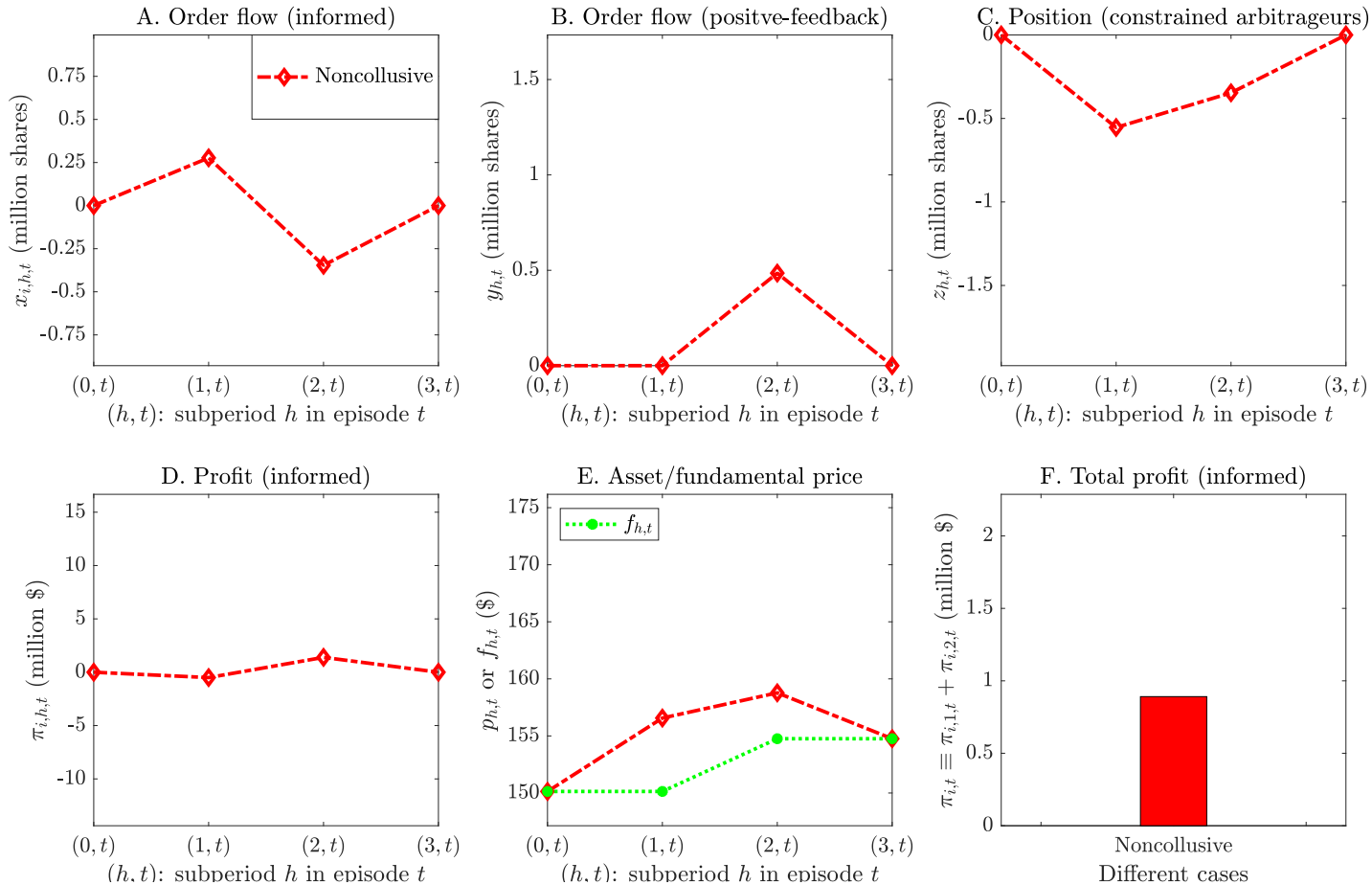
$$y_{h,t} = \mu(p_{h-1,t} - p_{h-2,t}), \quad \text{with } \mu > 0$$

e.g., De Long et al. (1990); Gennotte and Leland (1990); Hong and Stein (1999)

For simplicity, we assume away noise traders and market makers

Basic mechanism

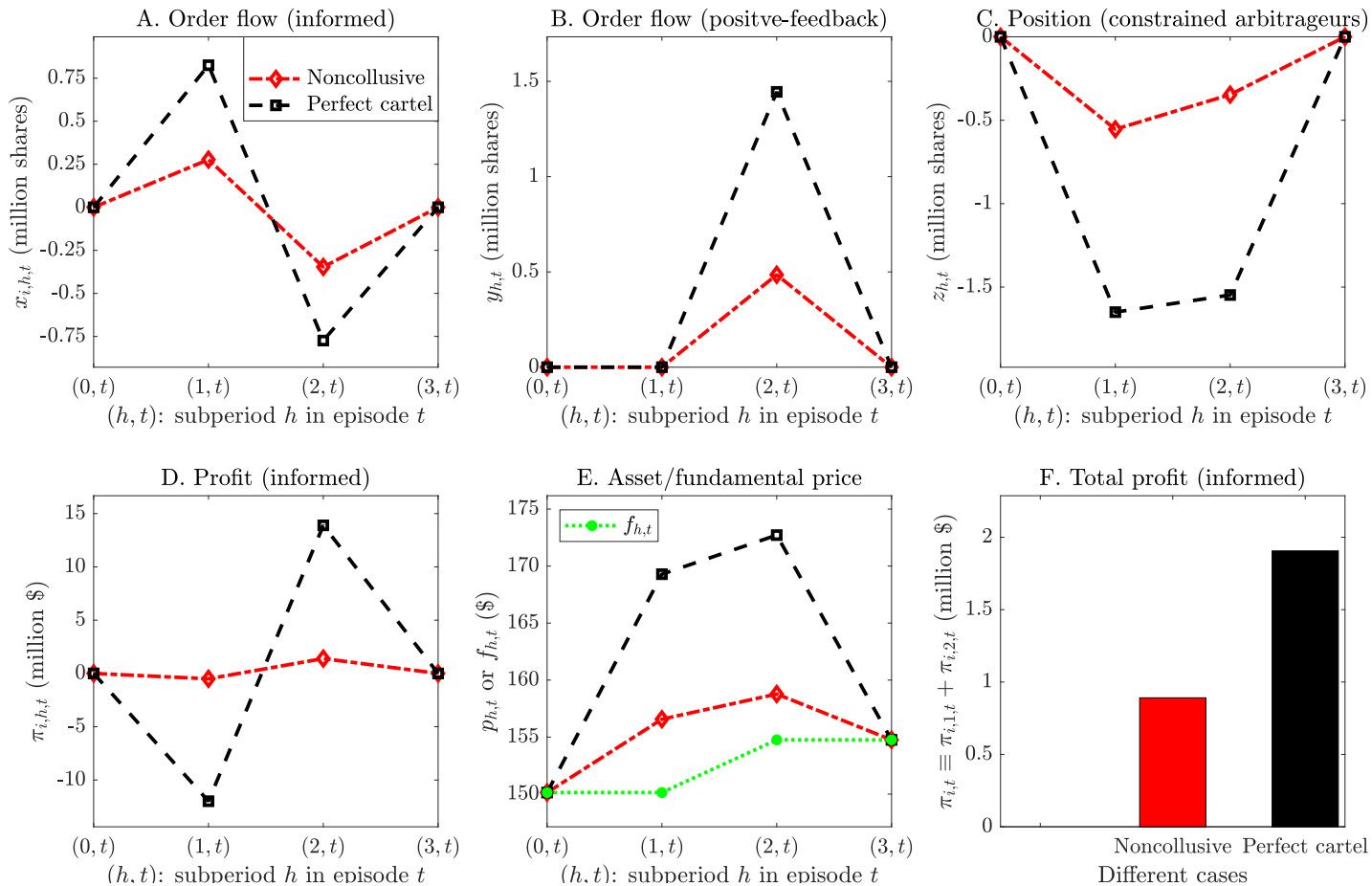
Non-collusive Nash equilibrium (*N*): Each speculator chooses an individually optimal strategy given rivals' strategies, without internalizing how her trading affects rivals' payoffs



Note: $H = 3$

Basic mechanism

Perfect cartel benchmark (*M*): Speculators trade as if they were one monopolist, maximize joint profit, and then split the aggregate order flow

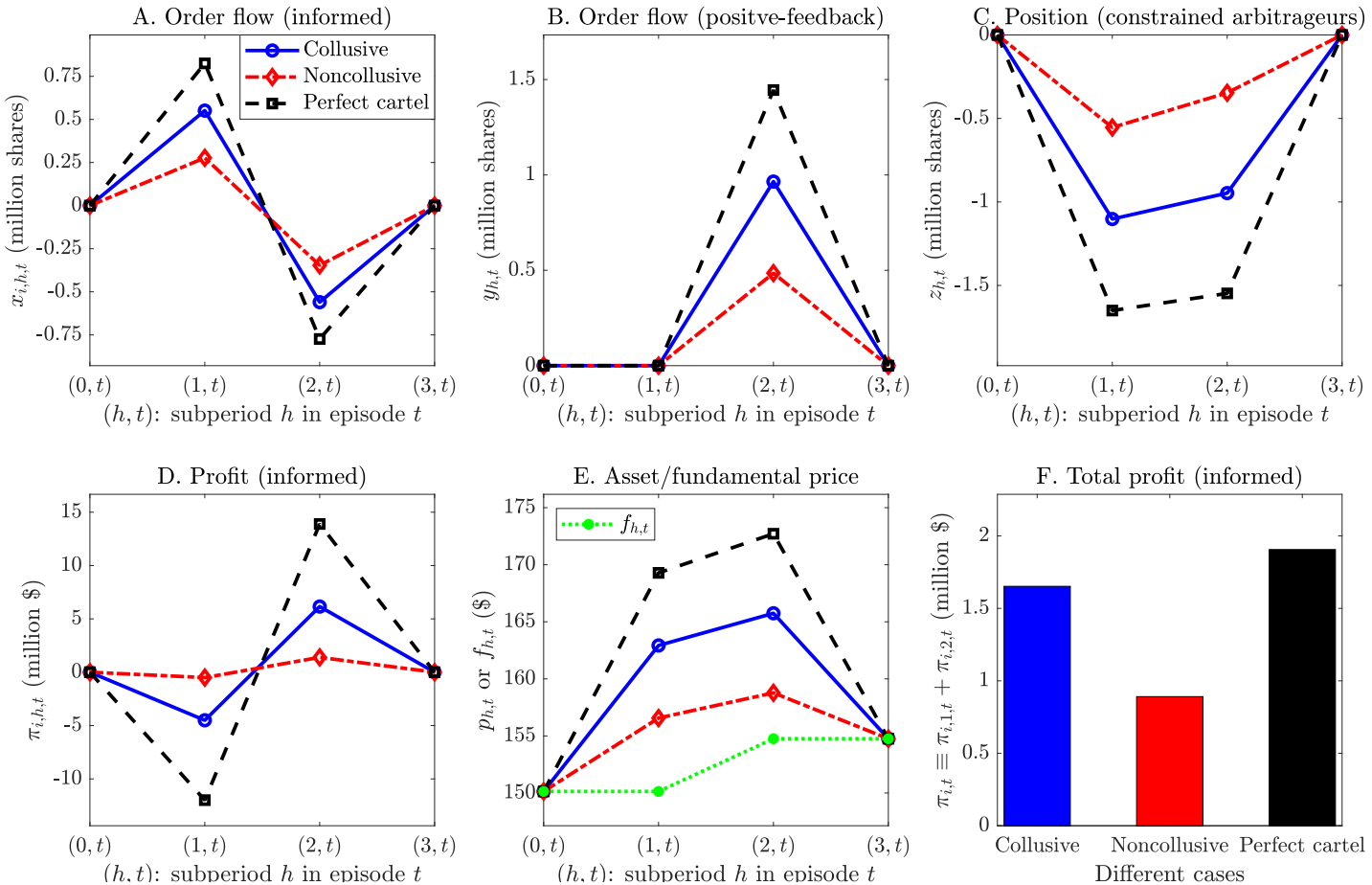


Note: $H = 3$

Basic mechanism

Collusive equilibrium with planning (C): Speculators use planning to coordinate trades across $(1, t)$ and $(2, t)$ and sustain collusive profits

- Collusion leads to instability: $\mathcal{S}^b = 1.99$



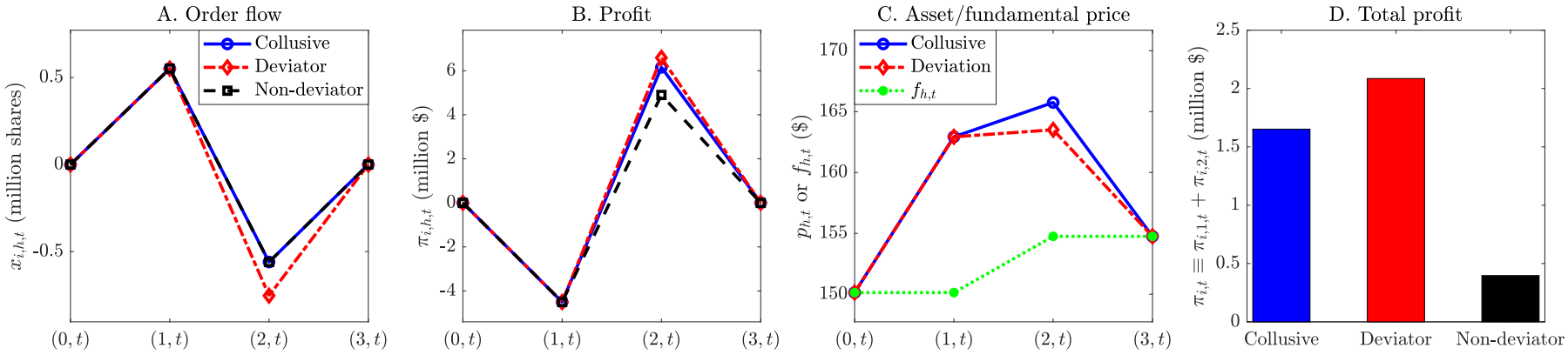
Note: $H = 3$

Challenges for collusive planning

1. Time inconsistency during bubble unwinding:

Once speculators create a large bubble at $(1, t)$:

- Each wants to unwind more aggressively than planned at $(2, t)$
- Deviation raises the trader's own profit at other speculators' expense

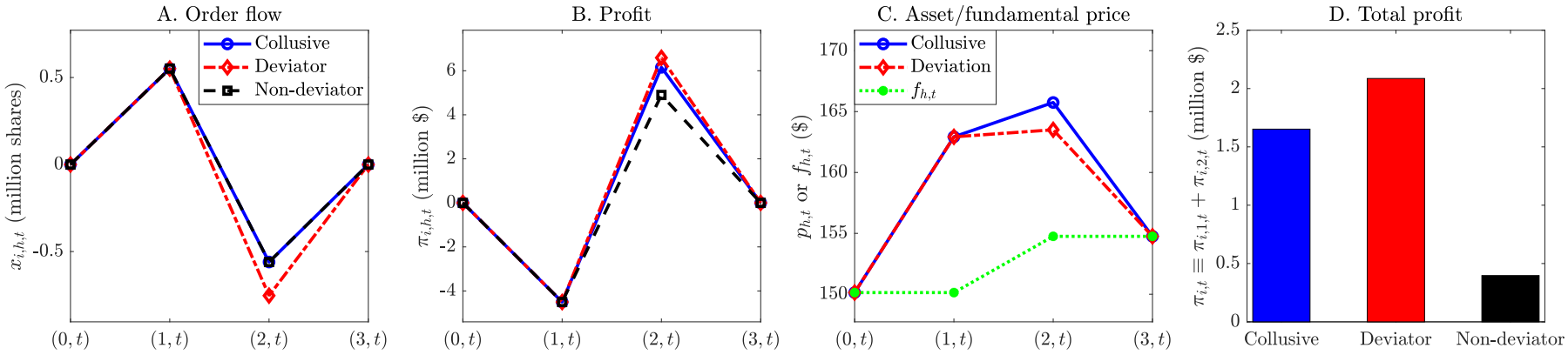


Challenges for collusive planning

1. Time inconsistency during bubble unwinding:

Once speculators create a large bubble at $(1, t)$:

- Each wants to unwind more aggressively than planned at $(2, t)$
- Deviation raises the trader's own profit at other speculators' expense



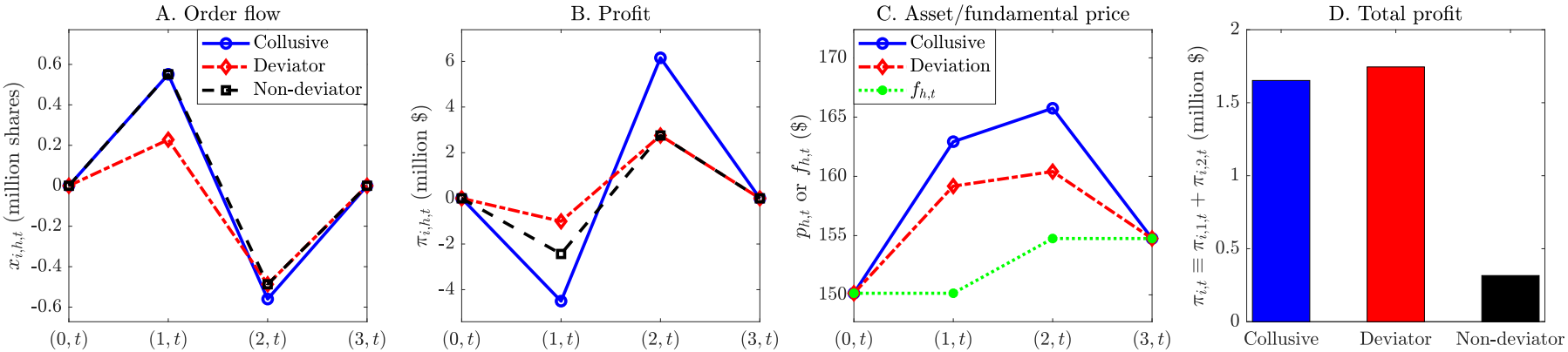
Solution: price-triggered Nash reversion

Excessive price reversal at $(2, t)$ signals aggressive unwinding

$\Rightarrow$ Non-collusive Nash play from $t + 1$ onward

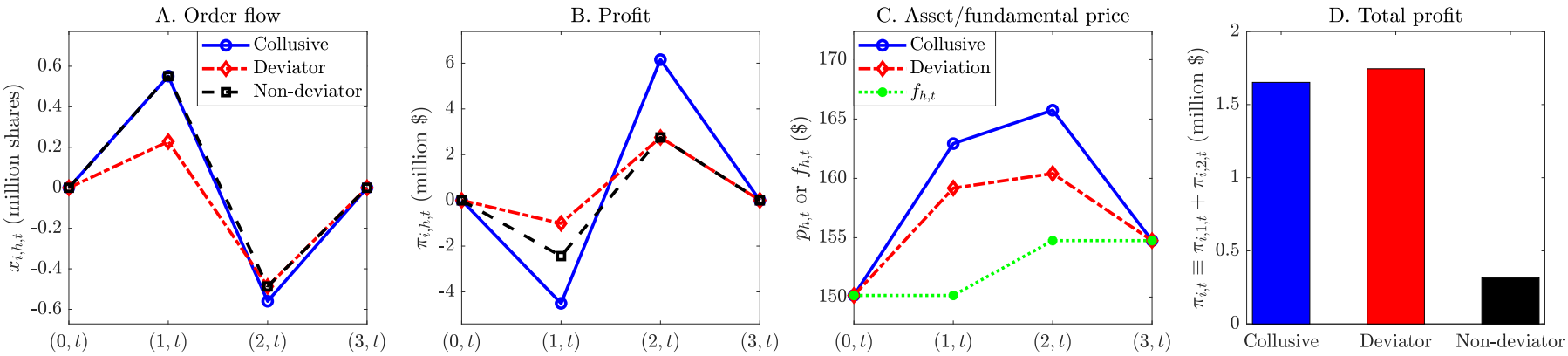
Challenges for collusive planning

2. Free-riding during bubble creation:



Challenges for collusive planning

2. Free-riding during bubble creation:

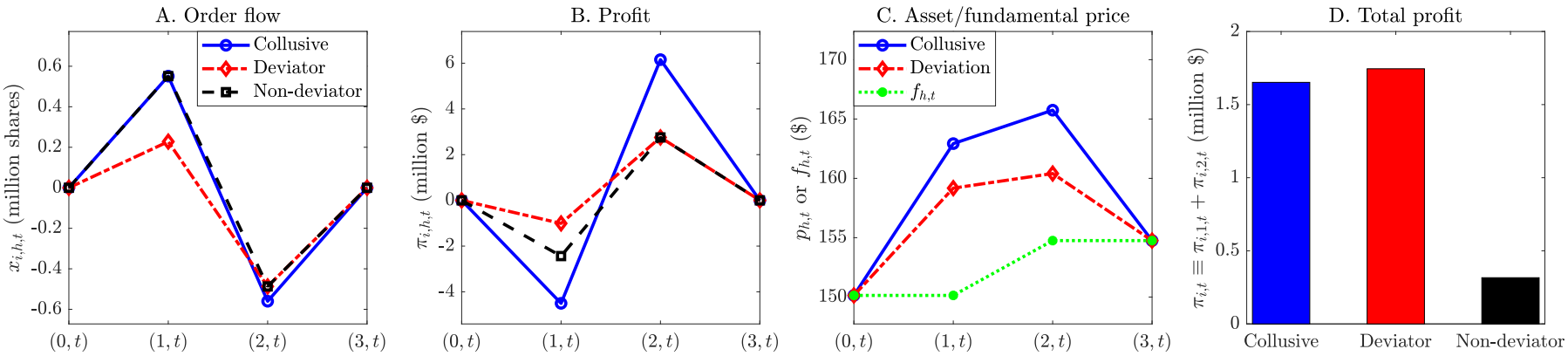


Why Nash reversion is too weak:

- Free-riding leaves a smaller bubble at $(1, t)$: little surplus remains to destroy at $(2, t)$
- Non-deviators would unwind moderately and profitably alongside the deviator

Challenges for collusive planning

2. Free-riding during bubble creation:



Why Nash reversion is too weak:

- Free-riding leaves a smaller bubble at $(1, t)$: little surplus remains to destroy at $(2, t)$
- Non-deviators would unwind moderately and profitably alongside the deviator

Solution: immediate punishment beyond Nash reversion

Non-deviators trade more aggressively than under Nash play at $(2, t)$

$\Leftarrow$ Compliance earns a faster return to collusion from $t + 1$ onward

Compounded price triggers: “stick and carrot”

Punishment must deter deviations and be credible

Abreu (1986, 1988); Mailath_Nocke_White (2017); Hatfield_Kominers_Lowery_Barry (2020)

- **“Stick”:** deter free-riding with immediate trading more aggressive than Nash reversion
- **“Carrot”:** reward traders for carrying out the punishment with a faster return to collusion

Compounded price triggers: "stick and carrot"

Punishment must deter deviations and be credible

Abreu (1986, 1988); Mailath_Nocke_White (2017); Hatfield_Kominers_Lowery_Barry (2020)

- **"Stick":** deter free-riding with immediate trading more aggressive than Nash reversion
- **"Carrot":** reward traders for carrying out the punishment with a faster return to collusion

Two conditional price triggers:

- **Layer 1: detect free-riding at** $(1, t)$

  Insufficient bubble formation signals possible free-riding

  $\Rightarrow$ Aggressive punishment at $(2, t)$

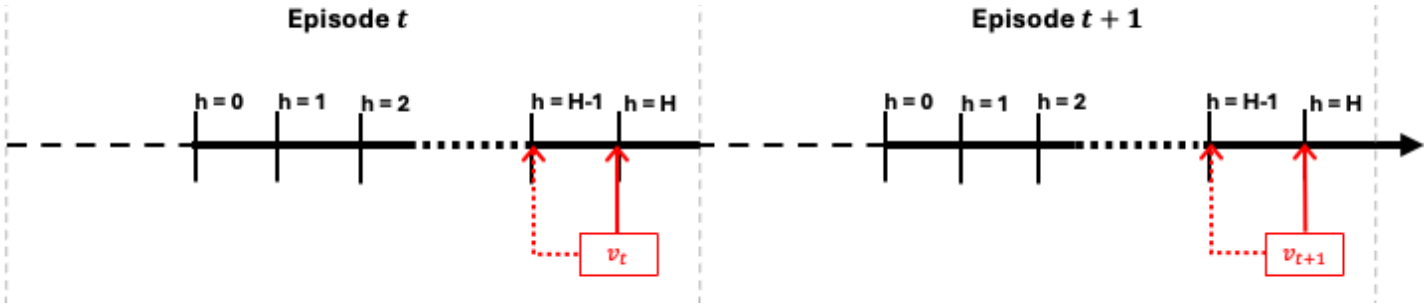
- **Layer 2: enforce punishment at** $(2, t)$

  Conditional on Layer 1, insufficient reversal signals weak punishment

  $\Rightarrow$ Delay the return to collusive trading behavior

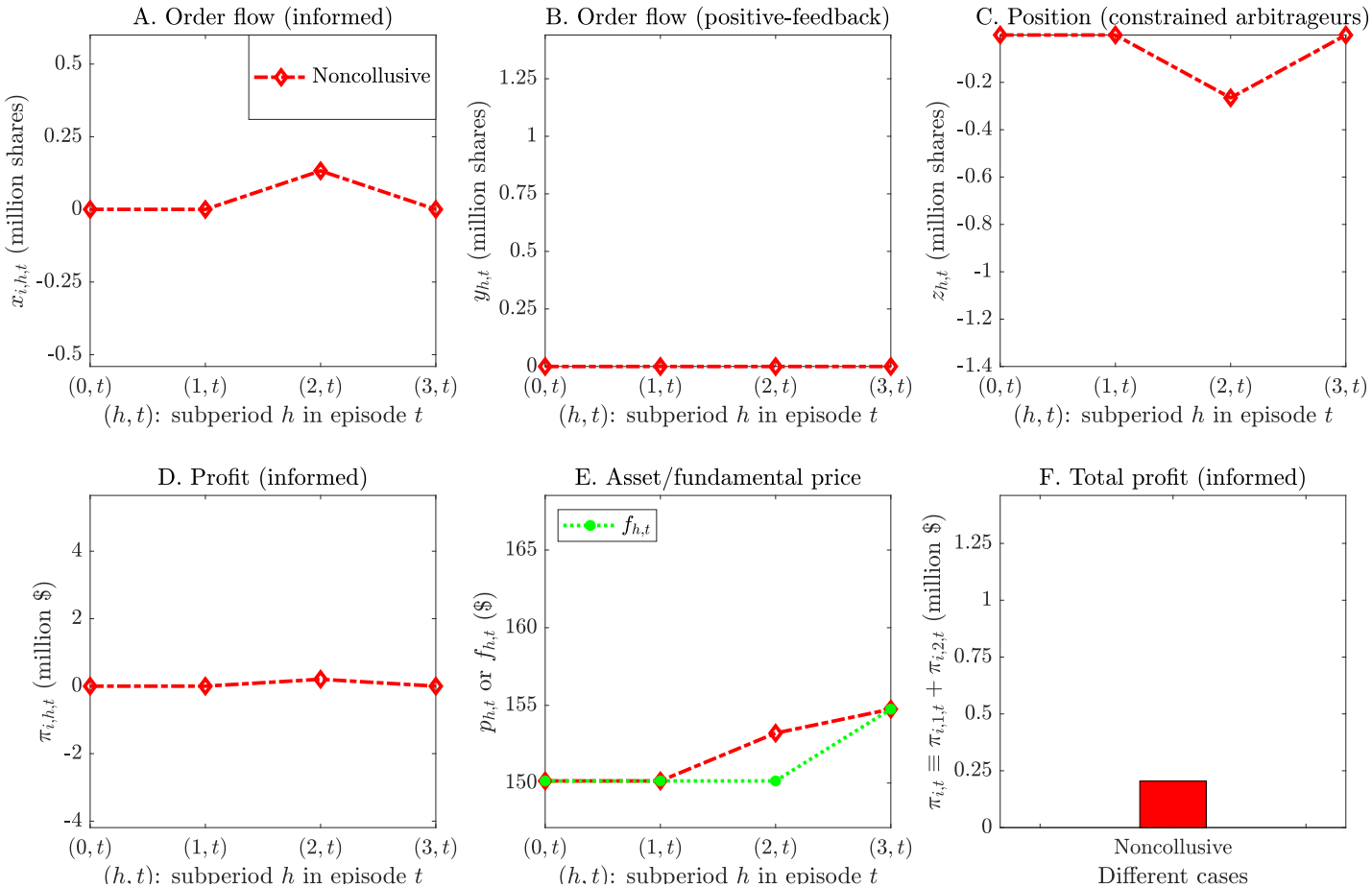
Stronger manipulation and greater fragility

Only the timing of information changes:



Stronger pump-and-dump manipulation

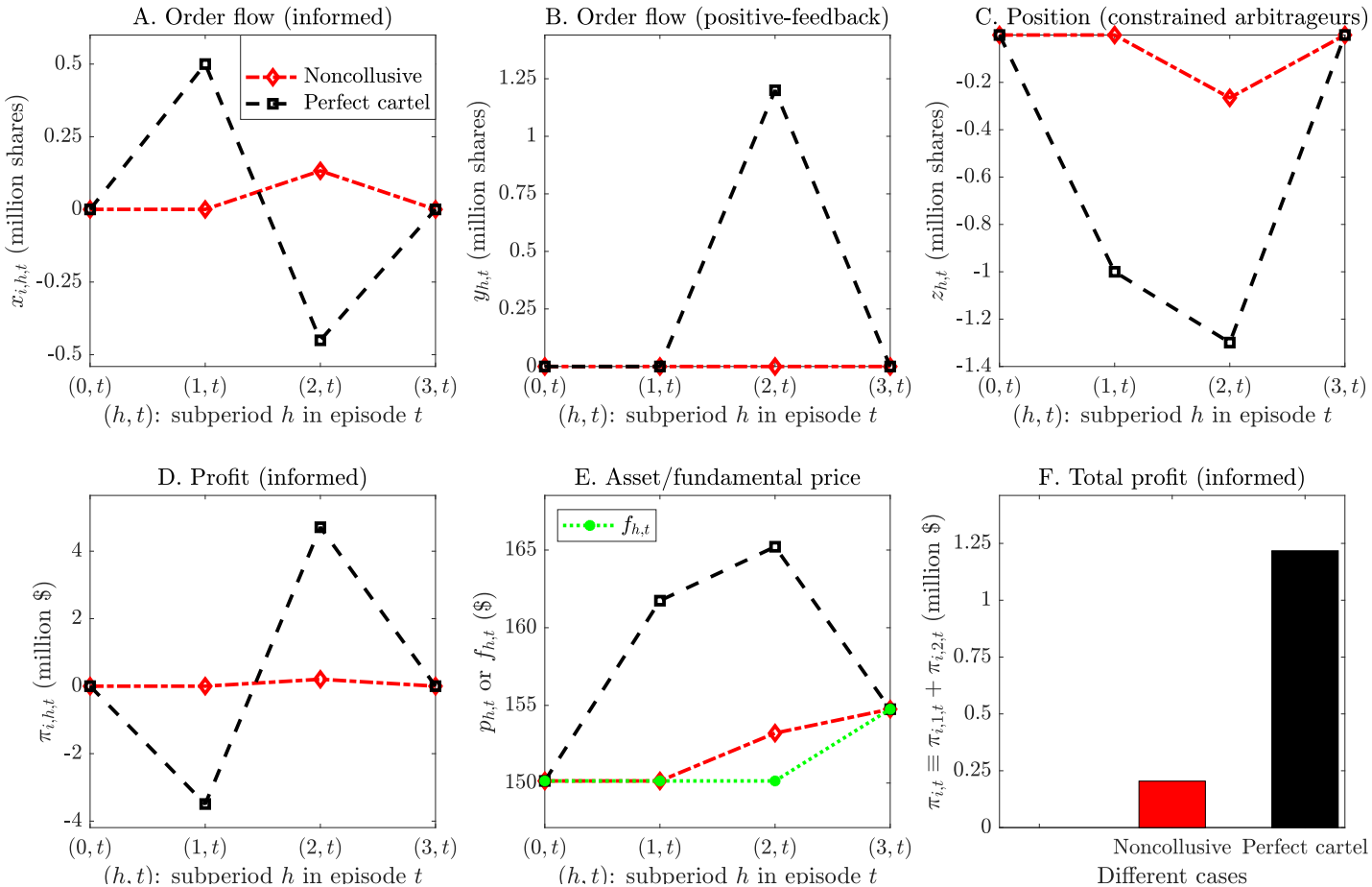
Non-collusive Nash equilibrium (*N*): Each speculator chooses an individually optimal strategy given rivals' strategies, without internalizing how her trading affects rivals' payoffs



Note: $H = 3$

Stronger pump-and-dump manipulation

Perfect cartel benchmark (M): Speculators trade as if they were one monopolist, maximize joint profit, and then split the aggregate order flow

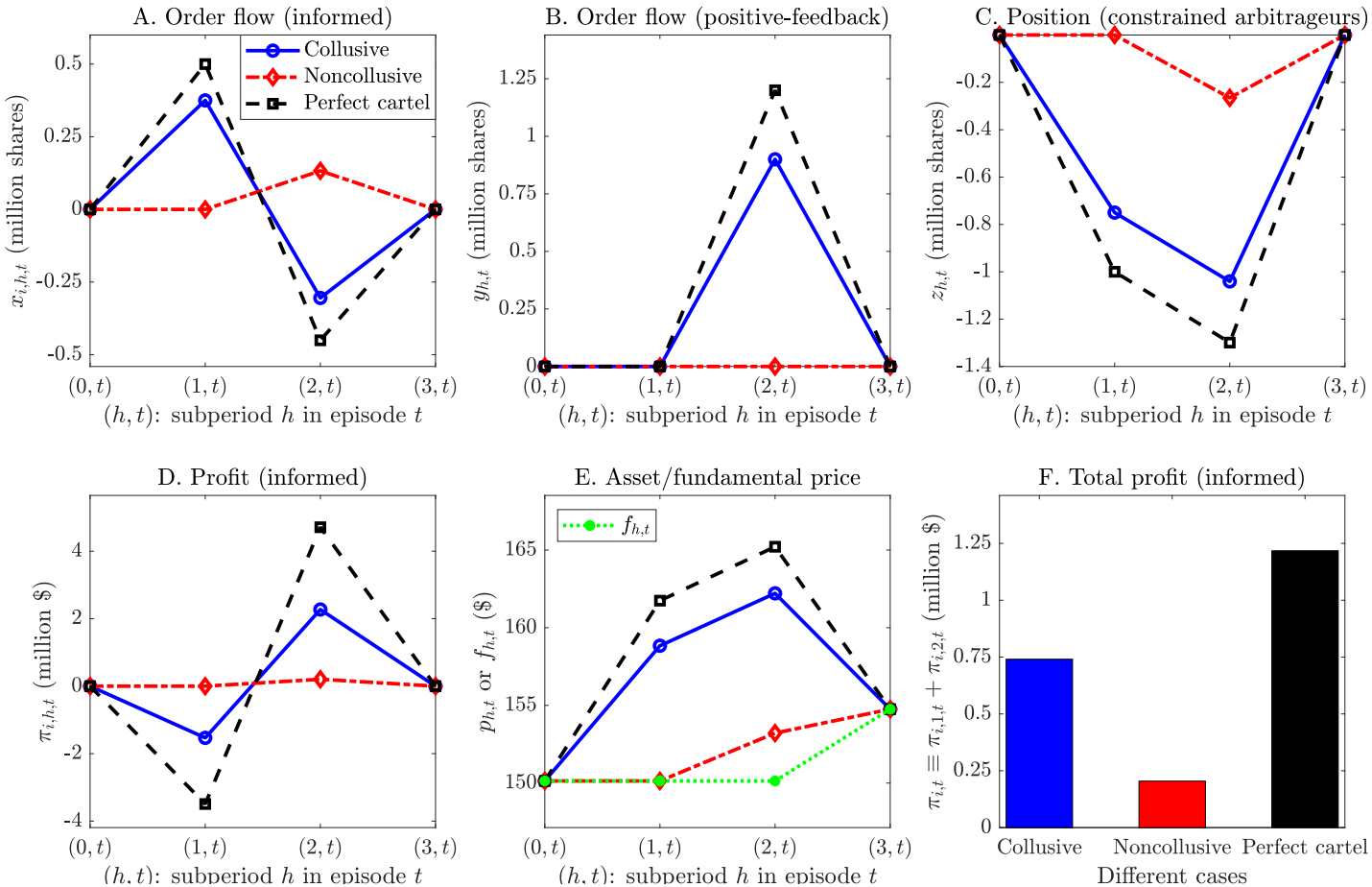


Note: $H = 3$

Stronger pump-and-dump manipulation

Collusive equilibrium with planning (C): Speculators use planning to coordinate trades across $(1, t)$ and $(2, t)$ and sustain collusive profits

- Collusion leads to instability: $\mathcal{S}^b = \infty$



Note: $H = 3$

Outline

1. **How dynamic coordination destabilizes markets**

2. **Q-learning algorithms with planning**

3. **Simulation experiments**
   - Experimental configuration and setup
   - Simulation results

Coupled Q-Learning for AI planning

Replace each theoretical speculator *i* with private information by a set of separate Q-matrices, with each maintained for one subperiod:

$$\left\{ \widehat{Q}_{i,0,t}(s,x),\ \ \widehat{Q}_{i,1,t}(s,x),\ \ \cdots,\ \ \widehat{Q}_{i,H,t}(s,x) \right\}$$

Coupled Q-learning updates: Multi-step planning via backward induction

$$\widehat{Q}_{i,0,t+1}(s_{i,0,t},x_{i,0,t}) = \alpha\left[\pi_{i,0,t} + \max_{x'_{i,1}\in\mathcal{X}_1} \widehat{Q}_{i,1,t}(s_{i,1,t},x'_{i,1})\right] + (1-\alpha)\widehat{Q}_{i,0,t}(s_{i,0,t},x_{i,0,t})$$

$$\widehat{Q}_{i,1,t+1}(s_{i,1,t},x_{i,1,t}) = \alpha\left[\pi_{i,1,t} + \max_{x'_{i,2}\in\mathcal{X}_2} \widehat{Q}_{i,2,t}(s_{i,2,t},x'_{i,2})\right] + (1-\alpha)\widehat{Q}_{i,1,t}(s_{i,1,t},x_{i,1,t})$$

$$\vdots$$

$$\widehat{Q}_{i,H,t+1}(s_{i,H,t},x_{i,H,t}) = \alpha\left[\pi_{i,H,t} + \rho \max_{x'_{i,0}\in\mathcal{X}_0} \widehat{Q}_{i,0,t}(s_{i,0,t+1},x'_{i,0})\right] + (1-\alpha)\widehat{Q}_{i,H,t}(s_{i,H,t},x_{i,H,t})$$

$\widehat{Q}_{i,h,t+1}$ is updated at $(s_{i,h,t}, x_{i,h,t})$, where $x_{i,h,t}$ is chosen independently according to:

$$x_{i,h,t} = \begin{cases} \underset{x'\in\mathcal{X}_h}{\operatorname{argmax}}\, \widehat{Q}_{i,h,t}(s_{i,h,t},x'), & \text{with prob. } 1-\varepsilon_t \quad \text{(exploitation)} \\ \widetilde{x} \sim \text{uniform on } \mathcal{X}_h, & \text{with prob. } \varepsilon_t \quad\ \ \text{(exploration)} \end{cases}$$

Reactive learning

Replace each theoretical speculator *i* with private information by a set of separate Q-matrices, with each maintained for one subperiod:

$$\left\{ \widehat{Q}_{i,0,t}(s,x),\ \ \widehat{Q}_{i,1,t}(s,x),\ \ \cdots,\ \ \widehat{Q}_{i,H,t}(s,x) \right\}$$

Non-coupled Q-learning updates: No intertemporal planning with each subperiod's Q-matrix updates separately

$$\widehat{Q}_{i,0,t+1}(s_{i,0,t},x_{i,0,t}) = \alpha\left[\pi_{i,0,t} + \rho \max_{x'_{i,0}\in\mathcal{X}_0} \widehat{Q}_{i,0,t}(s_{i,0,t+1},x'_{i,0})\right] + (1-\alpha)\widehat{Q}_{i,0,t}(s_{i,0,t},x_{i,0,t}),$$

$$\widehat{Q}_{i,1,t+1}(s_{i,1,t},x_{i,1,t}) = \alpha\left[\pi_{i,1,t} + \rho \max_{x'_{i,1}\in\mathcal{X}_1} \widehat{Q}_{i,1,t}(s_{i,1,t+1},x'_{i,1})\right] + (1-\alpha)\widehat{Q}_{i,1,t}(s_{i,1,t},x_{i,1,t}),$$

$$\vdots$$

$$\widehat{Q}_{i,H,t+1}(s_{i,H,t},x_{i,H,t}) = \alpha\left[\pi_{i,H,t} + \rho \max_{x'_{i,H}\in\mathcal{X}_H} \widehat{Q}_{i,H,t}(s_{i,H,t+1},x'_{i,H})\right] + (1-\alpha)\widehat{Q}_{i,H,t}(s_{i,H,t},x_{i,H,t})$$

$\widehat{Q}_{i,h,t+1}$ **is updated at** $(s_{i,h,t}, x_{i,h,t})$, where $x_{i,h,t}$ is chosen independently according to:

$$x_{i,h,t} = \begin{cases} \underset{x'\in\mathcal{X}_h}{\operatorname{argmax}}\, \widehat{Q}_{i,h,t}(s_{i,h,t},x'), & \text{with prob. } 1-\varepsilon_t \quad \text{(exploitation)} \\ \widetilde{x} \sim \text{uniform on } \mathcal{X}_h, & \text{with prob. } \varepsilon_t \quad\ \ \text{(exploration)} \end{cases}$$

Outline

1. **How dynamic coordination destabilizes markets**

2. **Q-learning algorithms with planning**

3. **Simulation experiments**
   - Experimental configuration and setup
   - Simulation results

RL algorithms as experimental subjects

Consider $\boldsymbol{H = 3}$, i.e., 4 subperiods each episode

Focus on $\widehat{\boldsymbol{Q}}_{\boldsymbol{i,1,t}}(\boldsymbol{s}, \boldsymbol{x})$ and $\widehat{\boldsymbol{Q}}_{\boldsymbol{i,2,t}}(\boldsymbol{s}, \boldsymbol{x})$, as they are the only ones associated with informed trading and market manipulation

- Payoffs: $\pi_{i,1,t} = (p_{3,t} - p_{1,t})x_{i,1,t}$ and $\pi_{i,2,t} = (p_{3,t} - p_{2,t})x_{i,2,t}$
- State variables:

$$s_{i,1,t} = \{p_{2,t-1} - p_{0,t-1}, \quad v_{t-1}, \quad v_t\}$$

$$s_{i,2,t} = \{p_{2,t-1} - p_{0,t-1}, \quad v_{t-1}, \quad v_t, \quad p_{1,t} - p_{0,t}\}$$

- Exploration rate: $\varepsilon_t = e^{-\beta t}$, independently drawn for $(1, t)$ and $(2, t)$

Baseline parameter values

Environment and preference parameters:

(calibrated based on 254 post-market earnings announcements of 10 large firms)

Parameter	Symbol	Value
Subjective discount factor	ρ	0.95
Quarterly risk-free rate (%)	r	0.730
Long-run mean of dividends ($/share)	$\overline{d}$	1.088
Quarterly dividend mean-reversion coefficient	λ	0.033
Constrained arbitrageurs' demand (thousand shares per $)	ξ	86.167
Feedback investors' demand (thousand shares per $)	μ	75.411
Volatility of ν_t shocks ($/share)	σ_ν	0.186

Discretization parameters:

$$n_{x_1} = 31, \quad n_{x_2} = 61, \quad n_{\overline{p}_1} = 61, \quad \text{and} \quad n_{\overline{p}_2} = 121$$

Hyperparameters:

$$\alpha = 0.05 \quad \text{and} \quad \beta = 10^{-9}$$

Note: AI speculators do not have prior knowledge of environment parameters

Convergence

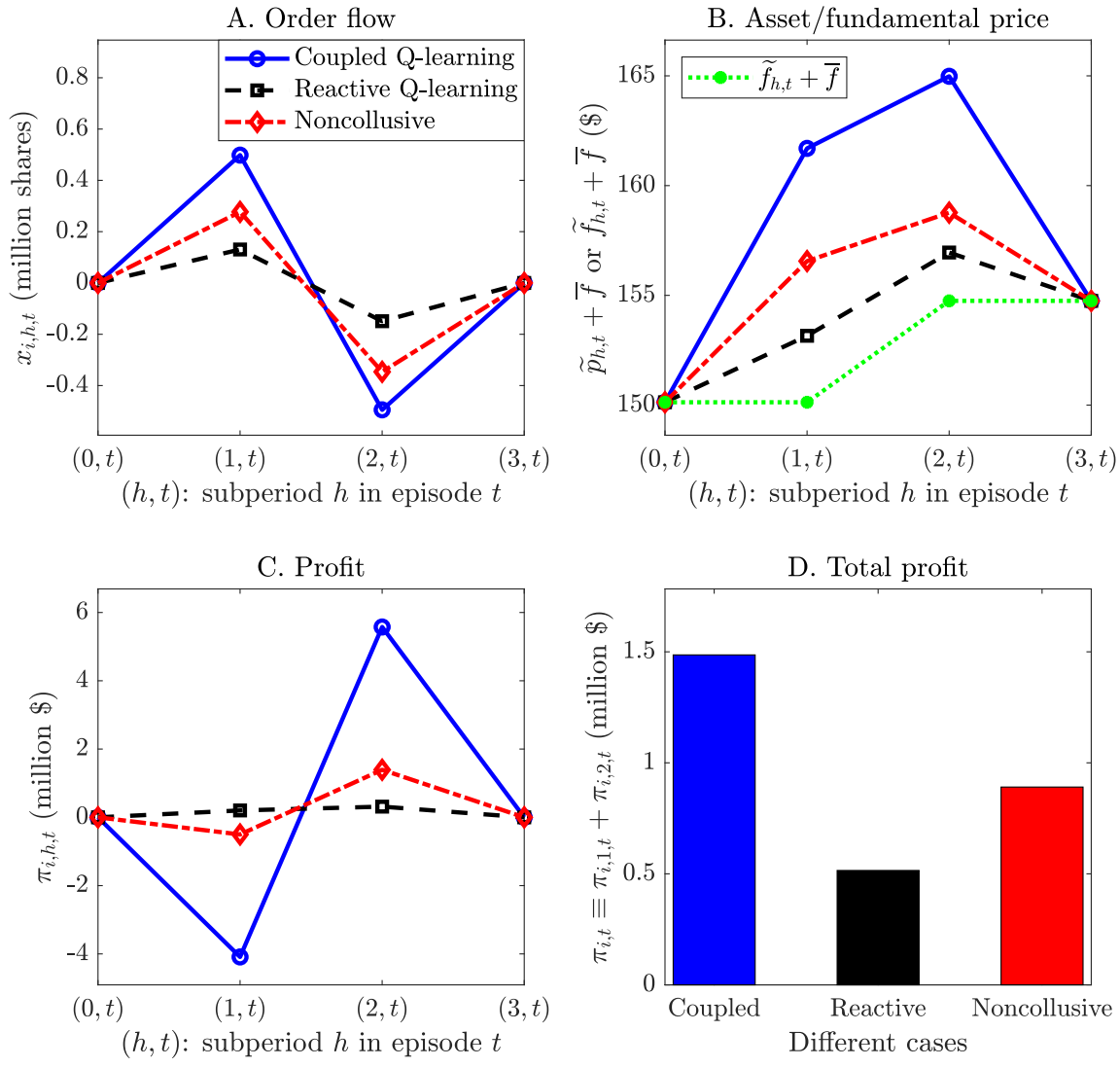
Numerical convergence criterion:

- Each AI speculator's optimal strategy does not change for $1,000,000,000$ consecutive periods
    - Multi-period collusive planning is significantly more difficult to learn than single-period collusive trading
- All $N = 1,000$ simulation sessions are simulated until convergence

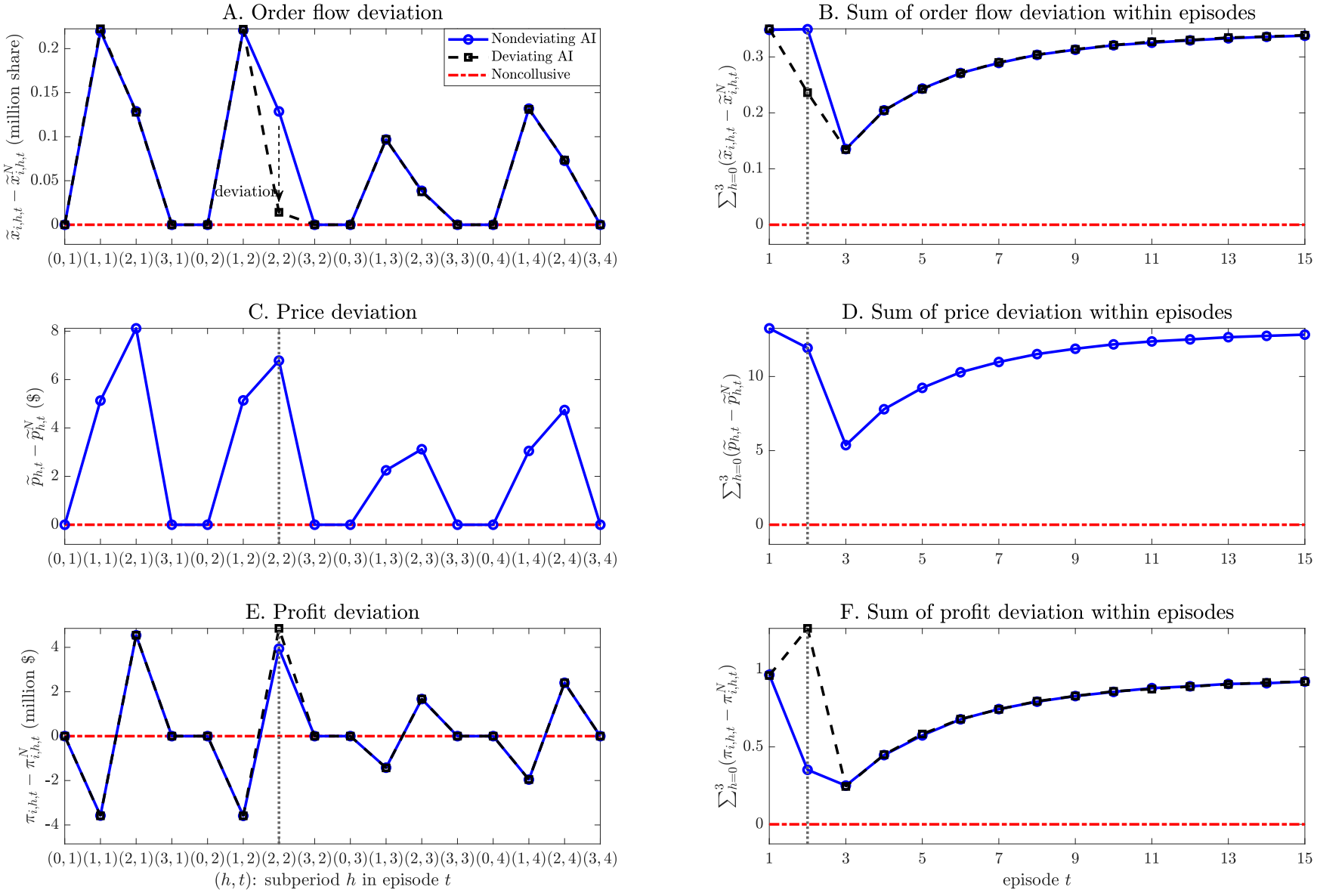
Computation power:

- Implemented in C++, need 3 to 48 hours to finish
- 4 AMD EPYC 9754 processors, with 512 CPU cores and 3TB memory

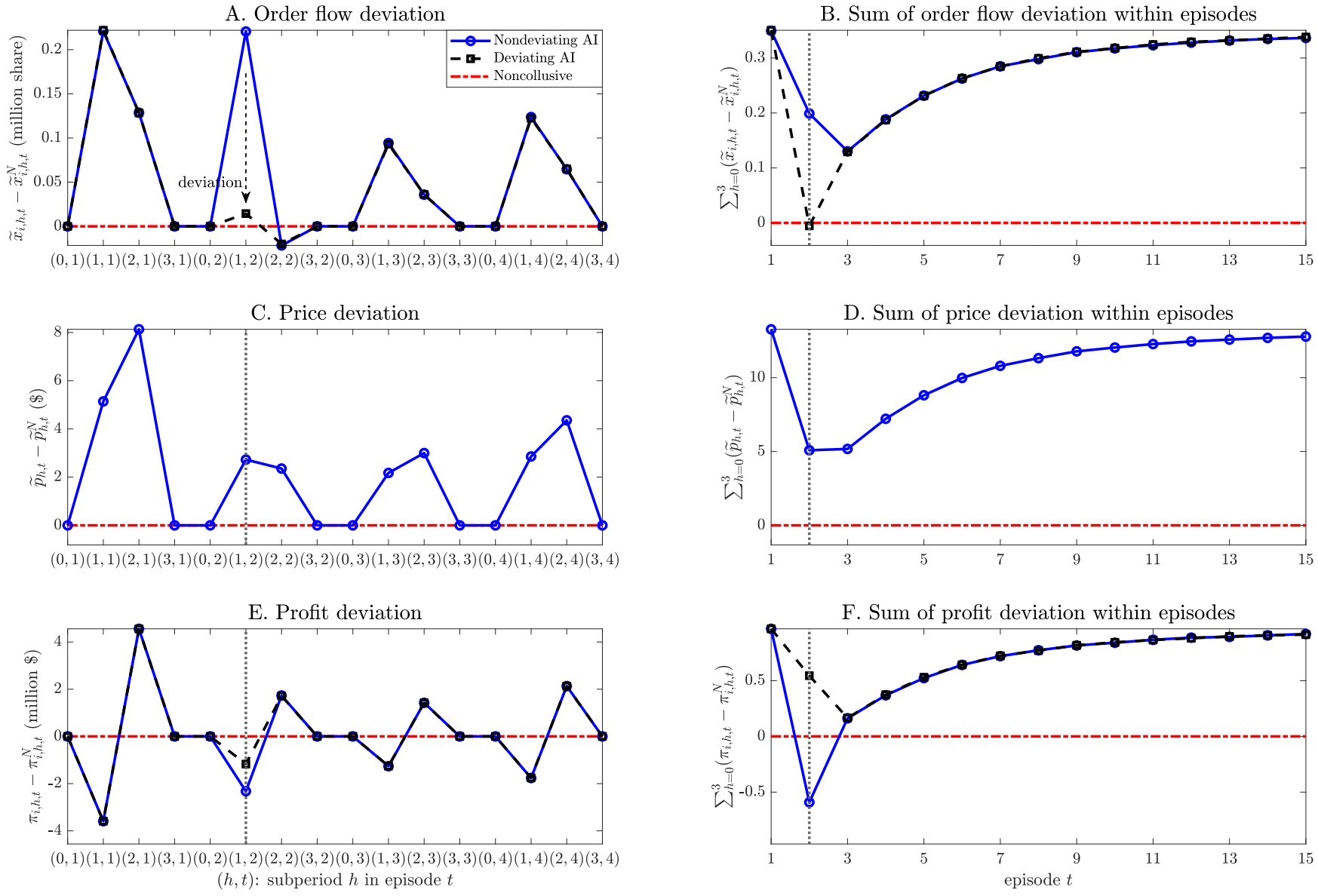
Coordination to create bubbles



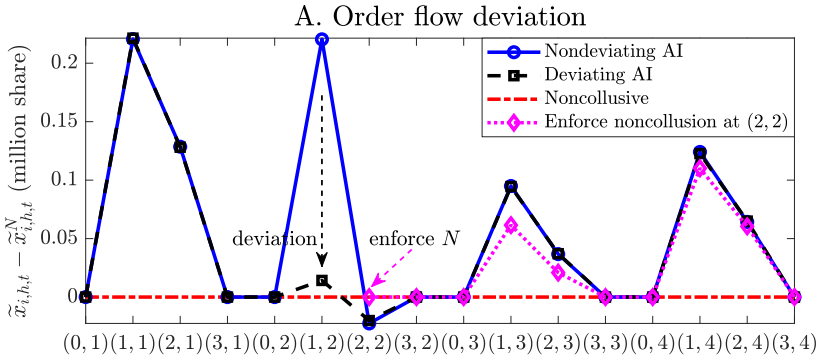
Unilateral deviation at $(2, 2)$

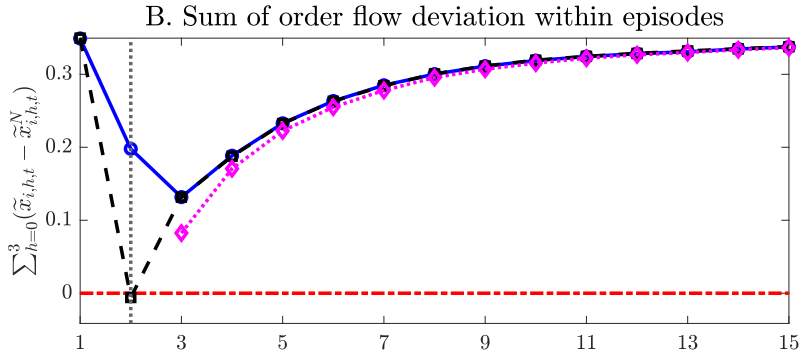


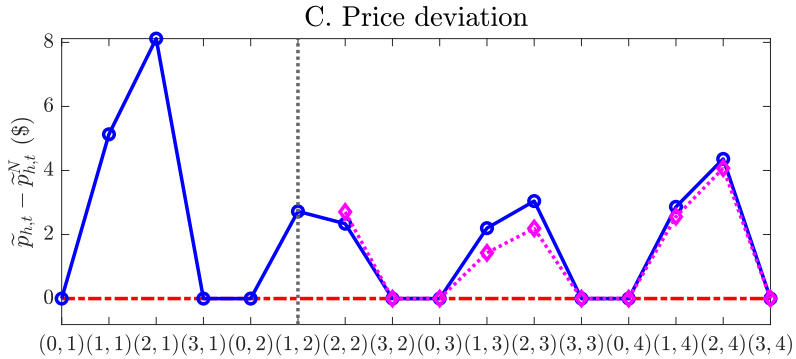
Unilateral deviation at (2, 1)

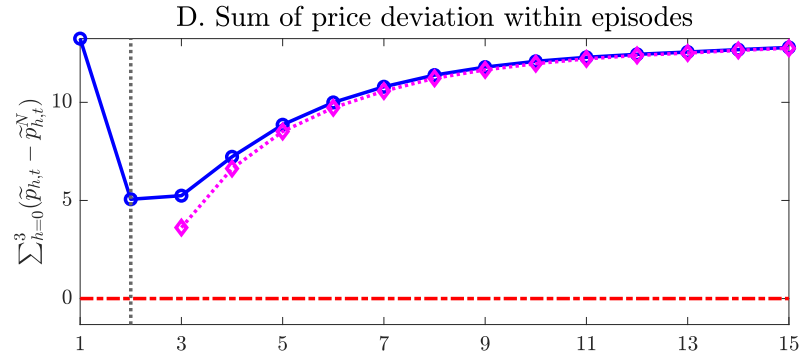


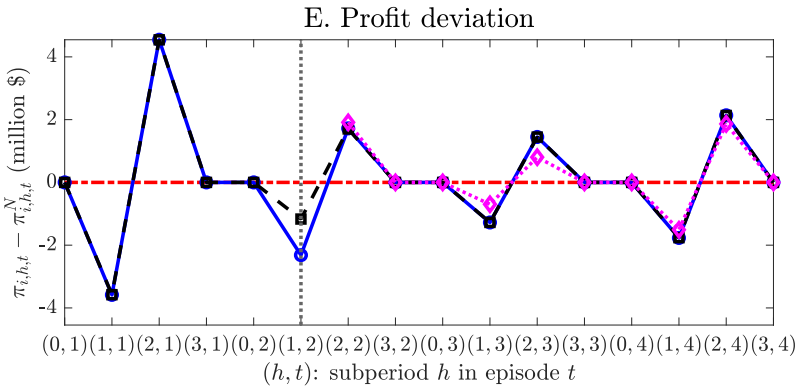
Deviation at (2, 1), non-collusive reversion at (2, 2)

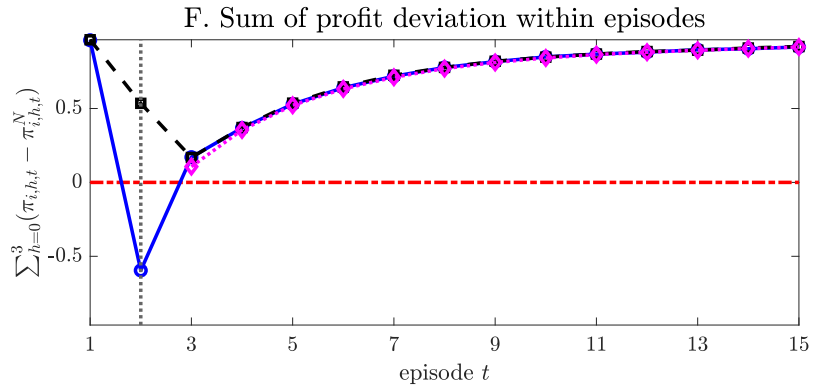












Conclusion

Coordinated AI planning emerges without communication or intent

- Autonomously learn to implement compounded price-trigger strategies
- Pump-and-dump manipulation emerges: AI speculators jointly create bubbles, attract feedback investors, and coordinate unwinding for profit

AI planning can become a new source of market fragility

- With strong positive-feedback trading, AI planning amplifies feedback-flow volatility and price volatility
- In a model estimated around large-tech earnings announcements, AI planning roughly doubles fragility and raises price volatility by 80%

Next step:

- Develop theoretical results on convergence and characterize the equilibrium in the limit (e.g., Chen_White, 1998; Dou_Goldstein_Li_Yang, 2026)