Financial Market Fragility in the Era of AI Planning

Winston Wei Dou Itay Goldstein Yan Ji *

February 12, 2026

Abstract

This paper examines how AI planning, the core technology behind agentic AI systems that pursue long-horizon objectives by anticipating how current actions shape future payoff-relevant states, affects financial market stability. We develop a dynamic trading framework with positive-feedback investors, constrained arbitrageurs, and oligopolistic informed speculators who may coordinate intertemporally: trading aggressively in tandem to generate (negative) bubbles and subsequently unwinding their positions in a coordinated manner to extract profits. Such intertemporal coordination differs from traditional collusion because it faces two unique, fundamental obstacles: time inconsistency, as coordinated plans become incentive-incompatible once a large (negative) bubble has formed, and weak punishment, as deviations are difficult to penalize when no large (negative) bubble materializes. We characterize equilibria featuring coordinated creation of manipulative, exploitative (negative) bubbles. In simulation experiments, AI speculators trained via reinforcement-learning algorithms with explicit planning modules autonomously discover and implement intertemporal collusive trading strategies based on compounded price-trigger rules, coordinating without communication or shared intent. When feedback trading is strong, these AI-planning speculators dynamically converge on destabilizing strategies that create and exploit (negative) bubbles, manipulate feedback traders, and significantly amplify market fragility.

Keywords: Financial Market Instability and Manipulation, Asset Price Bubbles and Crashes, AI Planning and Reinforcement Learning, Time Inconsistency, Compounded Price-Trigger Strategies, Intertemporal Strategic Coordination.
JEL Classification: D43, G10, G14, L13.

*Dou: University of Pennsylvania (wdou@wharton.upenn.edu) and NBER; Goldstein: University of Pennsylvania (itayg@wharton.upenn.edu) and NBER; Ji: Hong Kong University of Science and Technology (jiy@ust.hk). We thank Chong Huang, Leonid Kogan, and Tom Sargent, as well as seminar and conference participants at the AIEA, University of Hong Kong and Durham University. Dou gratefully acknowledges financial support from Analytics at Wharton (WAIAI).

1 Introduction

The proliferation of artificial intelligence (AI) planning technologies designed to pursue multi-period, long-horizon objectives is reshaping financial market trading and marks a clear paradigm shift in how financial market fragility may arise.[1] Regulators have warned that AI-powered trading could heighten market instability, including the formation of large (negative) bubbles followed by sharp crashes, and even increase the risk of a broader financial crisis.[2] A central question in addressing this regulatory challenge is how AI planning technologies may amplify financial market fragility.

This question goes beyond the scope of existing research on AI-driven market manipulation to avoid competition, for several reasons. First, the existing literature examines AI collusion that suppresses competition in financial and retail markets (e.g., Calvano et al., 2020; Dou, Goldstein and Ji, 2025). In contrast, our focus is on coordination achieved through intertemporal planning in financial market trading to extract profits by manipulating other investors' trading behavior and the resulting price dynamics, at the expense of these manipulated market participants. This mechanism is fundamentally different from collusion aimed merely to prevent competition; it requires dynamically coordinating aggressive trading to create (negative) bubbles and then further coordinating gradual unwinding to harvest profits.

Second, AI planning is a reinforcement-learning paradigm that selects actions by explicitly anticipating how current decisions shape the sequence of future payoff-relevant state variables, defined as state variables that directly enter the payoff function, across multiple subperiods within an episode in pursuit of a long-horizon, episode-level objective. This is achieved through either (i) model-based forward simulation or (ii) model-based dynamic programming via backward induction (e.g., Kaelbling, Littman and Cassandra, 1998; Sutton and Barto, 2018). This approach fundamentally differs from the reactive learning algorithms studied in prior work,[3] which select actions primarily based on accumulated rewards directly associated with past actions across episodes. While such algorithms can learn how actions affect next-episode state variables that do not directly enter the payoff function, they do not explicitly plan over the intertemporal evolution of

[1] Prominent examples of AI planning technologies include AlphaGo, which achieved superhuman performance, as well as decision-making systems in self-driving vehicles and advanced robotics.

[2] This study focuses on financial market instability characterized by excessive price fluctuations, such as (negative) bubbles and subsequent crashes, rather than on systemic risk to the broader economy that arises from financial instability driven by equilibrium switching in the financial intermediation sector.

[3] See, e.g., Calvano et al. (2020), Colliard, Foucault and Lovo (2025), and Dou, Goldstein and Ji (2025).

payoff-relevant states over subperiods of an episode.

Third, answering this question requires incorporating intertemporal strategic complementarities into a model of financial market trading and asset pricing. Such dynamic complementarities cannot be captured by the i.i.d. frameworks used in prior studies of AI-powered trading, which primarily analyze effects on price informativeness, market efficiency, price impact, and market liquidity. Our dynamic trading framework embeds both persistent fundamental dynamics and positive-feedback investors, creating the intertemporal feedback necessary to analyze financial market fragility (e.g., Goldstein, Huang and Yang, 2025). Within this environment, oligopolistic informed speculators may coordinate to manipulate positive-feedback investors through intertemporal planning. They first trade aggressively to generate (negative) bubbles, inducing positive-feedback investors to follow their trading activity and substantially amplify the bubble. Once the bubble has formed, the speculators then coordinate on gradually unwinding their positions to extract profits at the expense of the feedback traders.

Finally, this form of coordination differs fundamentally from traditional collusion based on reward-punishment trigger strategies. Not only is it not aimed merely at suppressing competition, but its sustainability also faces two distinct challenges. The first is time inconsistency: once large bubbles have formed, coordinated plans become ex post incentive-incompatible. The second is weak dynamic punishment: deviations undermine the credibility of the standard reversion to noncollusive trading as a disciplining device. When deviations occur, large (negative) bubbles fail to materialize, coordinated trading generates only limited excess profits, and the threat of future noncollusive reversion loses power as a deterrent. Together, these forces make intertemporal planning to manipulate positive-feedback investors through coordinated bubble creation and gradual unwinding inherently fragile and difficult to sustain.

We begin by developing a dynamic trading framework with persistent fundamentals and positive-feedback investors, which characterizes the potential for intertemporal coordination among informed speculators seeking to manipulate investor behavior and the resulting asset-price dynamics. Within this environment, we identify two primitive challenges that impede such intertemporal coordination. The first is time inconsistency: once a large bubble has formed, the coordinated unwinding plan becomes ex post incentive-incompatible. The second is weak dynamic punishment: deviations prevent large (negative) bubbles from materializing and thus eliminate the excess profits needed to support sufficiently painful punishment in future periods. As a theoretical benchmark,

we show that a compounded price-trigger mechanism can overcome both obstacles and sustain intertemporal coordination that generates (negative) bubbles followed by crash-like reversals.

After establishing these theoretical benchmarks, we conduct simulation experiments using reinforcement-learning algorithms equipped with explicit planning modules. These algorithms, which we refer to as informed AI speculators, trade in a synthetic environment constructed from our theoretical model, allowing us to analyze their learning dynamics, planning capabilities, and potential intertemporal coordination mechanisms in a fully controlled setting that captures the essential features of real-world market interactions.

In our simulation experiments, we demonstrate that the algorithms autonomously discover and implement intertemporal collusive trading strategies based on compounded price-trigger rules. This algorithmic mechanism underlying AI collusion has parallels in economic theory and can be interpreted through game-theoretic equilibrium concepts. By analyzing the resulting AI collusive equilibrium through extensive simulations, we show that whether collusive AI planning destabilizes financial markets depends critically on the underlying market environment. Specifically, with a significant presence of positive-feedback investors, who buy after price increases and sell after declines, informed AI speculators coordinate dynamically on destabilizing strategies that create and exploit (negative) bubbles, manipulating feedback traders and amplifying market fragility.

In out theoretical framework, time is structured into discrete episodes, each containing multiple subperiods, allowing for the examination of collusive intertemporal planning. The market features four distinct types of investors trading a long-lived asset that yields dividends at the end of each episode. First, there are oligopolistic informed speculators who trade repeatedly across periods and episodes. At the beginning of each episode, each informed speculator receives a perfect signal regarding the fundamental shock to the dividend yield of the long-lived asset. Second, there are positive-feedback investors, who trade based on recent price changes, creating the potential for market momentum. Third, constrained arbitrageurs trade against price deviations from fundamental value, providing a stabilizing force, albeit a limited one. Finally, low-frequency, risk-neutral investors ensure that the asset's price aligns with its fundamental value in the first and last subperiod of each episode, but remain inactive during the intermediate subperiods when strategic trading occurs. This model setup enables the potential for the asset's price to deviate from its fundamental value due to the trading activities of oligopolistic informed speculators, which is essential for studying how collusive intertemporal planning and

manipulation may arise in financial markets.

We establish two theoretical benchmarks: the non-collusive Nash equilibrium, in which informed speculators do not coordinate with one another, and the perfect cartel benchmark, where all informed speculators can perfectly coordinate, effectively acting as a monopoly. Any collusive equilibrium with intertemporal coordination among informed speculators lies between these two theoretical benchmarks. By comparing the collusive equilibrium with the non-collusive Nash equilibirum, we illuminate the implications of the collusive intertemporal planning in informed trading for the risk of market instability. Our central finding is that the impact of intertemporal coordination on financial market stability crucially depends on the significance of positive-feedback trading. When positive-feedback trading is prevalent (i.e., the demand of positive-feedback investors is large and very responsive to past price changes), speculators are incentivized to manipulate the behavior of positive-feedback investors to maximize profits. Relative to the non-collusive Nash equilibrium, they trade more aggressively to create a large (negative) asset price bubble, which induces positive-feedback investors to follow their trading, increasing the bubble size further. Subsequently, they gradually unwind their positions to extract profits and crash the bubble, resulting in large swings in the asset price and substantially increasing financial market instability. In contrast, when positive-feedback trading is limited, the strategic environment for informed traders resembles dynamic Cournot competition. The primary incentive for collusion in this scenario is to mitigate the price impact of their own trades. When trading non-collusively, speculators aggressively act on their private information, but this moves the asset price against them, eroding their profits. Thus, the collusive intertemporal planning in informed trading means trading conservatively. We show that in such environments, intertemporal coordination among informed speculators does not bring a significant impact on financial market stability. In fact, intertemporal coordination may lead to milder fluctuations in asset prices and slightly enhance market stability.

We theoretically illustrate the challenges of sustaining intertemporal coordination among informed speculators, particularly in trading environments characterized by a significant presence of positive-feedback investors. In such environments, coordination requires speculators to first incur losses by trading aggressively to inflate an asset price bubble before subsequently unwinding their positions for profit. This strategy is inherently unstable and faces two critical challenges. The first challenge is a time-inconsistency issue, where, after a bubble is created, each speculator has a strong incentive to deviate

from the coordinated strategy by trading more aggressively to maximize his individual payoff at the expense of other speculators. The second challenge is a weak punishment issue, where speculators are incentivized to free-ride on the initial bubble-creation efforts of others. A standard punishment of reverting to the non-collusive Nash equilibrium may be insufficient to deter this deviation, as without a sufficiently large bubble, coordinated trading yields only limited excess profits, rendering the standard punishment ineffective.

To address these challenges, we derive a compounded price-trigger strategy that establishes a sub-game perfect equilibrium capable of sustaining intertemporal coordination. This mechanism moves beyond a simple, uncompounded reversion to the non-collusive Nash equilibrium by introducing a multi-layered punishment structure with three phases: intertemporal coordination, collusive punishment, and non-collusive Cournot reversion. The key to this compounded price-trigger strategy is the implementation of a "one-time excessively aggressive trading" strategy within the collusive punishment phase, which is triggered by deviations during the initial bubble-creation stage. This severe but temporary punishment effectively deters free-riding by making defection unprofitable. The credibility of this punishment phase is, in turn, sustained by the threat of a transition to the even less profitable non-collusive Cournot reversion phase.

We next investigate whether informed AI speculators, each governed by an independent and self-interested algorithm, can achieve and sustain intertemporal coordination in a way that impacts the financial market instability described above. Specifically, we replace the rational-expectations informed speculators with autonomous informed AI speculators. Each AI speculator is an independent, profit-maximizing entity equipped with a model-free coupled Q-learning algorithm designed to facilitate multi-step AI planning. Crucially, these informed AI speculators are not programmed with any knowledge of the economic environment—they do not know the existence of positive-feedback investors and other market participants or the underlying market forces that generate fluctuations in the fundamental values and prices of the long-lived asset. Their learning is entirely endogenous, emerging from a process of trial-and-error as they interact with the market and each other over millions of trading episodes.

To capture the dynamic planning process of informed speculators in our model, which involves trading flows over multiple periods, we experiment with multi-period coupled Q-learning algorithms. Compared with reactive Q-learning algorithms, the coupled Q-learning algorithm employs a backward induction mechanism, allowing the algorithm to explicitly link the decisions made across subperiods. This enables the informed

AI speculators to formulate a coherent plan that considers the future consequences of immediate actions within a planning window. Apart from explicitly embedding this planning feature through dynamic programming within each episode, the coupled Q-learning algorithm behaves similarly as reactive Q-learning algorithms. Specifically, in each subperiod where informed AI speculators trade, the algorithm selects an action (i.e., the quantity traded) based on the current state. It employs subperiod-specific Q-matrices over a discrete state and action space to store and update estimated values for each state-action pair, encompassing both optimal and suboptimal actions. At the beginning of each episode, the algorithm observes the realized state and uses it to update a cell in each subperiod-specific Q-matrix corresponding to the state-action pair from each subperiod of the previous episode. The realized state could be influenced by both the prior state and action. The update is a weighted average of past experiences and new information, combining the reward just received in the end of the last episode with the estimated continuation value based on the newly realized state in the current episode.

When selecting actions, the algorithm follows a standard exploration-exploitation strategy. Exploitation involves choosing the action with the highest estimated Q-value for the current state in the current subperiod, while exploration involves selecting a random action. The balance between exploration and exploitation is a key feature of reinforcement learning algorithms and is essential for effective learning. Typically, the likelihood of exploration decreases over time, while the likelihood of exploitation increases. We specify a minimal set of state variables for the informed AI speculators that theoretically enables them to detect deviations from collusive behavior and condition actions based on past outcomes. These state variables include the current fundamental shock, the lagged fundamental shock, and the lagged price from the last subperiod of the previous episode in which informed AI speculators traded. Additionally, for the subperiods where positive-feedback investors trade, the asset price from the previous subperiod of the current episode is also included as a state variable.

The simulation results demonstrate that informed AI speculators, through reinforcement learning with explicit planning modules, successfully discover and implement sophisticated collusive strategies that mirror the mechanics of a compounded price-trigger strategy illustrated in our theory, achieving this coordination without explicit communication or pre-programmed intent. In our baseline trading environment, characterized by a significant presence of positive-feedback investors, the learned intertemporal coordination among AI speculators substantially amplifies financial market instability.

Impulse-response analyses reveal the mechanism by which this collusion is sustained. When an informed AI speculator deviates in the first subperiod to free-ride on bubble-creation efforts, all AI speculators respond with a one-time excessively aggressive punishment trade in the second subperiod of the same episode, followed by a persistent reversion to a less profitable collusive state. This harsh, immediate punishment renders the initial deviation unprofitable, thereby resolving the weak punishment issue. Deviations in the second subperiod are deterred by a persistent, multi-episode reversion to a less profitable but still collusive punishment phase, which effectively addresses the time-inconsistency issue.

Moreover, we analyze the behavior of informed AI speculators in trading environments with a limited presence of positive-feedback investors. While AI speculators still learn to punish deviations to maintain coordination, the punishment mechanism is less extreme. Specifically, informed AI speculators do not adopt the one-time excessively aggressive trading strategy, as there is no weak punishment issue in this setting. Instead, deviations are met with an immediate and persistent reversion to a less profitable collusive state. Aligning with the theoretical predictions, while collusion is sustained in the environments with AI trading, it does not involve the strategic creation of asset price bubbles and, as a result, does not significantly impact market stability, in some cases even providing a mild stabilizing effect.

These findings carry significant implications for financial regulation. They demonstrate that AI speculators equipped with planning capabilities can develop complex, tacitly collusive intertemporal trading strategies that can be detrimental to market stability, particularly in environments with significant feedback trading. The emergence of such strategies from independent learning processes, rather than from explicit design, highlights a novel source of systemic risk and underscores the challenge for regulators in monitoring and mitigating harmful algorithmic conduct in the era of AI planning.

Related Literature. Our work follows recent work on AI collusion in retail markets (e.g., Calvano et al., 2020, 2021; Johnson, Rhodes and Wildenbeest, 2023) and financial markets (Dou, Goldstein and Ji, 2025). While these papers have shown that AI algorithms can learn reward-punishment schemes to sustain collusion in simple environments where collusion only involves a single decision in an i.i.d. context, our setting involves highly strategic intertemporal coordination and planing of trading flows over multiple periods, exhibiting strong dynamic complementarity. Moreover, the collusive behavior results in

completely different trading patterns in different trading environments, despite the fact that informed AI speculators are not instructed with the differences across environments. Others have studied algorithmic mechanisms, which generate supra-competitive profits without the punishment trigger (e.g., Waltman and Kaymak, 2008; Hansen, Misra and Pai, 2021; Abada and Lambin, 2023; Asker, Fershtman and Pakes, 2024; Banchio and Mantegazza, 2024; Dolgopolov, 2024; Lambin, 2024), but they largely rely on simplifying restrictions on the algorithmic capacity.

Furthermore, our paper contributes to the rapid growing literature on the impact of AI and big data on the efficiency and functioning of financial markets (e.g., Goldstein, Spatt and Ye, 2021; Farboodi and Veldkamp, 2023, for literature review). Recent studies theoretically examine how data abundance and advances in information processing technologies affect price informativeness and market liquidity (e.g., Dugast and Foucault, 2018; Farboodi and Veldkamp, 2020; Dugast and Foucault, 2024). Another strand of the literature demonstrates that advanced machine learning techniques can effectively extract predictive signals or latent economic structures from high-dimensional public data, which are otherwise difficult to detect using traditional methods (e.g., Kaniel et al., 2023; Cao et al., 2024; Chen, Kelly and Xiu, 2024; Gao, Xiong and Yuan, 2024; Kelly, Malamud and Zhou, 2024; Chen et al., 2025). There are a few early works that investigate the effects that related algorithms may have on financial or money markets (e.g., Marimon, McGrattan and Sargent, 1990; Routledge, 1999, 2001). However, they either explore adaptive learning algorithms or more basic reinforcement learning algorithms than ours. More recently, Colliard, Foucault and Lovo (2025) study the interactions among Q-learning algorithms but focuses on stateless AI market makers. Cartea et al. (2022) also analyze stateless reinforcement learning in market making using a multi-armed bandit algorithm. In contrast, our paper focuses on understanding the implications of AI-powered trading and planning for the risk of market instability.

Finally, our paper is closely related to the literature on imperfect competition in financial markets. Rostek and Yoon (2024) provide a recent review of the theory of imperfectly competitive financial markets, covering influential early contributions such as Kyle (1989) and Vayanos (1999), which focus on non-collusive equilibria. Theoretical works that study the effect of coordination or collusion among major market participants on market microstructure dynamics under a repeated game framework include Dutta and Madhavan (1997), Carlin, Lobo and Viswanathan (2007), and Hörner, Lovo and Tomala (2018), among others. The common feature of these models is that supra-competitive

trading profits are sustained through the threat of punishment. In addition to these models that focus on market microstructure dynamics, there are other papers that theoretically analyze collusion in financial markets due to FinTech (e.g., Cong and He, 2019). Several papers provide supporting evidence for collusion in financial markets across various settings (e.g., Christie and Schultz, 1994, 1995; Christie, Harris and Schultz, 1994; Chen and Ritter, 2000; Dou, Wang and Wang, 2023; Bryzgalova, Pavlova and Sikorskaya, 2025; Lehar and Parlour, 2025). Unlike traditional dynamic games studied in the literature, which typically focus on single-decision collusion in an i.i.d. context, our model examines intertemporal coordination that involves a dynamic planning process with strongly complementary trading flows over multiple periods. We contribute further by showing that autonomous, self-interested AI-planning algorithms can learn compounded price-trigger strategies for sustaining the highly profitable yet fragile bubble-and-crash strategy.

2 Planning-Based AI Trading Algorithms

To illustrate the algorithms, we first describe the general environment. Consider a multi-agent Markov decision process with I agents, indexed by $i = 1, \ldots, I$. Time is divided into infinitely many episodes indexed by $t = 1, 2, \ldots$. Within each episode t, there are H subperiods indexed by $h = 1, 2, \ldots, H$. In each episode, agent i makes a sequence of H decisions, one in each subperiod h, based on the state $s_{i,t,h}$. The state $s_{i,t,h}$ includes both aggregate environmental conditions and agent-specific components, such as private information and individual constraints, and consists of payoff-relevant state variables that directly enter the agent's payoff function. The state $s_{i,t,h}$ evolves according to a Markov process driven by exogenous shocks, the agent's own actions, and the joint actions of all I agents in the system.

AI planning is a reinforcement-learning paradigm that selects actions by explicitly anticipating how current decisions shape the sequence of future payoff-relevant state variables, defined as state variables that directly enter the payoff function, across multiple subperiods h within an episode t in pursuit of a long-horizon, episode-level objective. This approach fundamentally differs from the reactive learning algorithms studied in prior work,[4] which select actions primarily based on accumulated rewards directly associated with past actions across episodes. While such algorithms can learn how actions affect next-episode state variables that do not directly enter the payoff function, they do not

[4]See, e.g., Calvano et al. (2020), Colliard, Foucault and Lovo (2025), and Dou, Goldstein and Ji (2025).

explicitly plan over the intertemporal evolution of payoff-relevant states over subperiods of an episode.

AI planning is implemented through two main approaches. Model-based forward simulation uses an internal representation of the environment to simulate the consequences of alternative action sequences. Model-based dynamic programming methods, implemented via backward induction, instead propagate value functions backward across subperiods to determine optimal current actions. Here, we focus on trading algorithms with AI planning, implemented using model-based dynamic programming.

While reinforcement learning encompasses various variants (e.g., Watkins and Dayan, 1992; Sutton and Barto, 2018), we choose to focus on Q-learning for several reasons. First, Q-learning serves as a foundational framework for many dynamically sophisticated reinforcement learning algorithms, which underpin numerous recent breakthroughs in AI. It is also important to note that while many AI trading algorithms currently in use may not exclusively rely on Q-learning principles, they often draw upon its concepts. Second, Q-learning maintains significant popularity among computer scientists for practical applications. Lastly, Q-learning algorithms are characterized by their simplicity and transparency, providing clear economic interpretations.

2.1 Bellman Equation and Q-Function

Agent i's intertemporal optimization is characterized by the Bellman equation and is solved recursively using dynamic programming:

$$\begin{aligned} V_{i,h}(s_{i,h}) &= \max_{x_{i,h}\in\mathcal{X}_h} \{\mathbb{E}[\pi_{i,h}|s_{i,h},x_{i,h}] + \mathbb{E}[V_{i,h+1}(s_{i,h+1})|s_{i,h},x_{i,h}]\}, \text{ with } h = 1,\ldots,H-1 \\ V_{i,H}(s_{i,H}) &= \max_{x_{i,H}\in\mathcal{X}_H} \{\mathbb{E}[\pi_{i,H}|s_{i,H},x_{i,H}] + \rho\mathbb{E}[V_{i,1}(s'_{i,1})|s_{i,H},x_{i,H}]\}. \end{aligned} \tag{2.1}$$

Here, $x_{i,h} \in \mathcal{X}_h$ is the action of agent i for subperiod $h = 1,\ldots,H$ within an episode, with $\mathcal{X}_h$ denoting the set of available actions, $\pi_{i,h}$ is the payoff received by agent i in subperiod h, which depends on the chosen action $x_{i,h}$ as well as the actions of other agents. The variable $s_{i,h} \in \mathcal{S}_h$ represents the state in subperiod h of the episode, and $s'_{i,1} \in \mathcal{S}_1$ represents the state in subperiod $h = 1$ of the next episode, with $\mathcal{S}_h$ denoting the set of states in subperiod h. The first term on the right-hand side, $\mathbb{E}[\pi_{i,h}|s_{i,h},x_{i,h}]$, is agent i's expected payoff in the current subperiod h, and the second term is agent i's continuation value. There is no discounting across subperiods within the same episode,

whereas payoffs in future episodes are discounted by the factor ρ across episodes.

Equation (2.1) represents the recursive formulation of dynamic control problems (e.g., Bellman, 1954; Ljungqvist and Sargent, 2012). It characterizes behavior along the equilibrium path, where the optimal value function $V_{i,h}(s_{i,h})$ depends only on the current state $s_{i,h}$. In contrast, the Q function, denoted by $Q_{i,h}(s_{i,h}, x_{i,h})$, extends the value function to each possible state-action pair, allowing evaluation of outcomes not only on the equilibrium path but also for counterfactual or off-path actions. By definition, the Q-function value for a given $(s_{i,h}, x_{i,h})$ corresponds to the right-hand side of equation (2.1):

$$\begin{aligned}
Q_{i,1}(s_{i,1}, x_{i,1}) &= \mathbb{E}[\pi_{i,1}|s_{i,1}, x_{i,1}] + \mathbb{E}[V_{i,2}(s_{i,2})|s_{i,1}, x_{i,1}], \\
Q_{i,2}(s_{i,2}, x_{i,h}) &= \mathbb{E}[\pi_{i,2}|s_{i,2}, x_{i,2}] + \mathbb{E}[V_{i,3}(s_{i,3})|s_{i,2}, x_{i,2}], \\
&\vdots \\
Q_{i,H}(s_{i,H}, x_{i,H}) &= \mathbb{E}[\pi_{i,H}|s_{i,H}, x_{i,H}] + \rho\mathbb{E}[V_{i,H}(s'_{i,1})|s_{i,H}, x_{i,H}],
\end{aligned} \tag{2.2}$$

Intuitively, the Q-function value, $Q_{i,h}(s_{i,h}, x_{i,h})$, can be interpreted as the quality of the vector of actions $x_{i,h}$ in state $s_{i,h}$. The optimal value of a state, $V_{i,h}(s_{i,h}, x_{i,h})$, is the maximum of all the possible Q-function values of state $s_{i,h}$. That is, $V_{i,h}(s_{i,h}) \equiv \max_{x'_h \in \mathcal{X}_h} Q_{i,h}(s_{i,h}, x'_h)$. By substituting $V_{i,h}(s_{i,h})$ with $\max_{x'_h \in \mathcal{X}_h} Q_{i,h}(s_{i,h}, x'_h)$ for $h = 2, \ldots, H$ and $V_{i,h}(s'_{i,1})$ with $\max_{x'_1 \in \mathcal{X}_1} Q_{i,1}(s'_{i,1}, x'_1)$ in equation (2.2), we can establish a recursive formula for the Q-function as follows:

$$\begin{aligned}
Q_{i,1}(s_{i,1}, x_{i,1}) &= \mathbb{E}\left[\pi_{i,1} + \max_{x'_2 \in \mathcal{X}_2} Q_{i,2}(s_{i,2}, x'_2) \middle| s_{i,1}, x_{i,1}\right], \\
Q_{i,2}(s_{i,2}, x_{i,h}) &= \mathbb{E}\left[\pi_{i,2} + \max_{x'_3 \in \mathcal{X}_3} Q_{i,3}(s_{i,3}, x'_3) \middle| s_{i,2}, x_{i,2}\right], \\
&\vdots \\
Q_{i,H}(s_{i,H}, x_{i,H}) &= \mathbb{E}\left[\pi_{i,H} + \rho \max_{x'_1 \in \mathcal{X}_1} Q_{i,1}(s'_{i,1}, x'_1) \middle| s_{i,H}, x_{i,H}\right].
\end{aligned} \tag{2.3}$$

When $|\mathcal{S}_h|$ and $|\mathcal{X}_h|$ for $h = 1, \ldots, H$ are finite, the Q-function in subperiod h can be represented as an $|\mathcal{S}_h| \times |\mathcal{X}_h|$ matrix, which is often referred to as the Q-matrix.

2.2 Coupled Q-Learning for AI Planning

If agent i possessed knowledge of its Q-matrix, determining the optimal actions for any given state $s_{i,h}$ would be straightforward. In essence, Q-learning estimates the Q-matrix when the conditional distribution $\mathbb{E}[\cdot|s_{i,h}, x_{i,h}]$ and off-equilibrium observations $(s_{i,h}, x_{i,h})$ are limited. The Q-learning algorithm is designed to this challenge by relying on the law of large numbers to learn the underlying distribution from repeated experiences, while its trial-and-error experiments generates counterfactual outcomes for state-action pairs that may not occur along the equilibrium path.

The coupled Q-learning algorithm is designed to facilitate multi-step AI planning. The core concept of this approach is to structure the decision-making process within a planning period t as a sequence of H subperiods. By employing a backward induction mechanism, the algorithm explicitly links the decisions made in these subperiods. This enables the AI agents to formulate a coherent plan that considers the future consequences of immediate actions within the planning window.

Specifically, agent i's iterative experimentation begins with H arbitrary initial estimated Q-matrices, $\{\widehat{Q}_{i,0,h}\}_{h=1}^{H}$. At the beginning of iteration $t+1$, the agent recursively updates all these Q-matrices from $\{\widehat{Q}_{i,t,h}\}_{h=1}^{H}$ to $\{\widehat{Q}_{i,t+1,h}\}_{h=1}^{H}$ based on the rewards and state-action pairs in period t as follows:

$$\begin{aligned}
\widehat{Q}_{i,t+1,1}(s_{i,t,1}, x_{i,t,1}) &= \alpha\Big[\pi_{i,t,1} + \max_{x_2' \in \mathcal{X}_2} \widehat{Q}_{i,t,2}(s_{i,t,2}, x_2')\Big] + (1-\alpha)\widehat{Q}_{i,t,1}(s_{i,t,1}, x_{i,t,1}), \\
\widehat{Q}_{i,t+1,2}(s_{i,t,2}, x_{i,t,2}) &= \alpha\Big[\pi_{i,t,2} + \max_{x_3' \in \mathcal{X}_3} \widehat{Q}_{i,t,3}(s_{i,t,3}, x_3')\Big] + (1-\alpha)\widehat{Q}_{i,t,2}(s_{i,t,2}, x_{i,t,2}), \\
&\vdots \\
\widehat{Q}_{i,t+1,H}(s_{i,t,H}, x_{i,t,H}) &= \alpha\Big[\pi_{i,t,H} + \rho \max_{x_1' \in \mathcal{X}_1} \widehat{Q}_{i,t,1}(s_{i,t+1,1}, x_1')\Big] + (1-\alpha)\widehat{Q}_{i,t,H}(s_{i,t,H}, x_{i,t,H}),
\end{aligned} \tag{2.4}$$

where $\alpha \in [0,1]$ captures the forgetting rate[5], $\pi_{i,t,h}$, $x_{i,t,h} \in \mathcal{X}_h$, and $s_{i,t,h} \in \mathcal{S}_h$ are the payoff, action, and state in subperiod $h = 1, \ldots, H$ of period t, respectively. The specification of coupled Q-learning facilitates multi-step planning through backward induction by capturing the sequential interdependence of actions and the endogenous updating of

[5]The forgetting rate α determines how quickly past experiments are discounted. For consistent learning, α must decay to zero to ensure convergence of the estimated Q-matrix $\widehat{Q}_{i,t}$ as t grows large. A smaller α improves asymptotic accuracy but slows convergence, reflecting a higher learning capacity at the expense of greater computational cost.

information sets based on past actions within the same period. Specifically, the action $x_{i,t,1}$ chosen at $h = 1$ in state $s_{i,t,1}$ may influence the states in subsequent subperiods, such as $s_{i,t,2}$, which in turn affects the action $x_{i,t,2}$ chosen at $h = 2$, and so on. Because $\widehat{Q}_{i,t,h+1}$ is the contiuation value of $\widehat{Q}_{i,t,h}$ for $h = 1, \ldots, H-1$, all the H Q-matrices must be learned jointly. Any bias that arises during the learning process of one Q-matrix can affect other Q-matrices, potentially disrupting effective learning. This makes coupled Q-learning for a system of recursive equations in (2.4) significantly more challenging than learning a single equation via Q-learning as in (Calvano et al., 2020; Dou, Goldstein and Ji, 2025), particularly when the sequential actions exhibit strong interdependence.

The update from $\{\widehat{Q}_{i,t,h}\}_{h=1}^{H}$ to $\{\widehat{Q}_{i,t+1,h}\}_{h=1}^{H}$ occurs immediately after $s_{i,t+1,1}$ is realized at $(t+1, 1)$, i.e., the first subperiod of iteration $t+1$. This update relies on the experiment conducted in each subperiod h of period t. As indicated by (2.4), only the values of the estimated Q-matrices $\{\widehat{Q}_{i,t,h}(s_{i,t,h}, x_{i,t,h})\}_{h=1}^{H}$ corresponding to the state-action pair $\{(s_{i,t,h}, x_{i,t,h})\}_{h=1}^{H}$ are updated. All other state-action pairs remain unchanged. For $h = 1, \ldots, H-1$, the updated value $\widehat{Q}_{i,t+1,h}(s_{i,t,h}, x_{i,t,h})$ is computed as a weighted average of accumulated knowledge from the previous experiments, $\widehat{Q}_{i,t,h}(s_{i,t,h}, x_{i,t,h})$, and learning from a new experiment, $\pi_{i,t,h} + \max_{x'_{h+1} \in \mathcal{X}_{h+1}} \widehat{Q}_{i,t,h+1}(s_{i,t,h+1}, x'_{h+1})$; and for $h = H$, the updated value $\widehat{Q}_{i,t+1,H}(s_{i,t,H}, x_{i,t,H})$ is computed as a weighted average of accumulated knowledge from the previous experiments, $\widehat{Q}_{i,t,H}(s_{i,t,H}, x_{i,t,H})$, and learning from a new experiment, $\pi_{i,t,H} + \rho \max_{x'_1 \in \mathcal{X}_1} \widehat{Q}_{i,t,1}(s_{i,t+1,1}, x'_1)$, which depends on the realized state $s_{i,t+1,1}$ at the beginning of iteration $t+1$. A key distinction between the coupled Q-learning recursive algorithm (2.4) and the Bellman recursive equation (2.1) lies in how they treat expectations for future payoffs and continuation Q-values. The coupled Q-learning algorithm (2.4) does not form expectations about the continuation value because the Markovian transition probabilities from $s_{i,t,h}$ to $s_{i,t,h+1}$ for $h = 1, \ldots, H-1$ or from $s_{i,t,H}$ to $s_{i,t+1,1}$ are unknown. Instead, it updates the Q-value using the actual realized payoff and the maximum Q-value of the realized state in the next subperiod.

2.3 Reactive Q-Learning

The reactive Q-learning does not take into account how past actions can endogenously update the information available for future choices within each period. As a result, it fails to capture the sequential interdependence among these decisions.

When conducting reactive Q-learning, agent i's iterative experimentation begins with

H arbitrary initial estimated Q-matrices, $\{\widehat{Q}_{i,0,h}\}_{h=1}^{H}$. At the beginning of iteration $t+1$, the agent recursively updates all these Q-matrices from $\{\widehat{Q}_{i,t,h}\}_{h=1}^{H}$ to $\{\widehat{Q}_{i,t+1,h}\}_{h=1}^{H}$ based on the rewards and state-action pairs in period t as follows:

$$\begin{aligned}
\widehat{Q}_{i,t+1,1}(s_{i,t}, x_{i,t,1}) &= \alpha\Big[\pi_{i,t,1} + \rho \max_{x_1' \in \mathcal{X}_1} \widehat{Q}_{i,t,1}(s_{i,t+1}, x_1')\Big] + (1-\alpha)\widehat{Q}_{i,t,1}(s_{i,t}, x_{i,t,1}),\\
\widehat{Q}_{i,t+1,2}(s_{i,t}, x_{i,t,2}) &= \alpha\Big[\pi_{i,t,2} + \rho \max_{x_2' \in \mathcal{X}_2} \widehat{Q}_{i,t,2}(s_{i,t+1}, x_2')\Big] + (1-\alpha)\widehat{Q}_{i,t,2}(s_{i,t}, x_{i,t,2}),\\
&\vdots\\
\widehat{Q}_{i,t+1,H}(s_{i,t}, x_{i,t,H}) &= \alpha\Big[\pi_{i,t,H} + \rho \max_{x_H' \in \mathcal{X}_H} \widehat{Q}_{i,t,H}(s_{i,t+1}, x_H')\Big] + (1-\alpha)\widehat{Q}_{i,t,H}(s_{i,t}, x_{i,t,H}),
\end{aligned} \tag{2.5}$$

By comparing reactive Q-learning (2.5) and coupled Q-learning equations (2.4), it is clear that reactive Q-learning fails to capture the sequential interdependence of actions across subperiods because each Q-matrix $\widehat{Q}_{i,t,h}$ is estimated separately through iterations. Moreover, reactive Q-learning considers the same state $s_{i,t}$ within iteration t, without accounting for the potential impact of action $x_{i,t,h}$ on the evolution of $s_{i,t}$ across subperiods.

Similar to coupled Q-learning, for agent i in iteration $t+1$, only the value of the estimated subperiod-h Q-matrix $\widehat{Q}_{i,t,h}(s,x)$ corresponding to the state-action pair $(s_{i,t}, x_{i,t,h})$ is updated to $\widehat{Q}_{i,t+1,h}(s_{i,t}, x_{i,t,h})$. All other state-action pairs remain unchanged. The updated value $\widehat{Q}_{i,t+1,h}(s_{i,t}, x_{i,t,,h})$ is computed as a weighted average of learning based on a new experiment, $\pi_{i,t,h} + \rho \max_{x_h' \in \mathcal{X}_h} \widehat{Q}_{i,t,h}(s_{i,t+1}, x_h')$, and accumulated knowledge based on the previous experiments, $\widehat{Q}_{i,t,h}(s_{i,t}, x_{i,t,h})$.

2.4 Experimentation

For coupled Q-learning, actions are determined sequentially as the state in each subperiod is realized. Specifically, in iteration t, the experimentation process begins with $h=1$ and continues sequentially until $h=H$. In each subperiod h, experimentation is conducted as follows. Upon the state $s_{i,t,h}$ being realized, each agent i chooses an action $x_{i,t,h}$ according to either an exploitation or exploration mode, as follows:

$$x_{i,t,h} = \begin{cases} \operatorname{argmax}_{x_h \in \mathcal{X}_h} \widehat{Q}_{i,t,h}(s_{i,t,h}, x_h), & \text{with prob. } 1-\varepsilon_t, \quad \text{(exploitation)} \\ \widetilde{x}_h \sim \text{uniform distribution on } \mathcal{X}_h, & \text{with prob. } \varepsilon_t. \quad \text{(exploration)} \end{cases} \tag{2.6}$$

After all agents choose $x_{i,t,h}$, the environment transitions to the next subperiod, where a new state is realized. To determine the mode, we employ the simple ε-greedy method. As outlined in equation (2.6), after the state $s_{i,t,h}$ is realized in iteration t, agent i follows either the exploration or exploitation mode with exogenous probabilities ε_t and $1-\varepsilon_t$, respectively. In the exploitation mode, agent i chooses its vector of actions to maximize the estimated Q-value based on $s_{i,t,h}$ in iteration t, given by $x_{i,t,h} = \text{argmax}_{x_h \in \mathcal{X}_h} \widehat{Q}_{i,t,h}(s_{i,t,h}, x_h)$. In contrast, in the exploration mode, agent i randomly chooses the whole vector of actions $\widetilde{x}_h$ from the set of all possible values in $\mathcal{X}_h$, each with equal probability.[6] Exploration enables the algorithm to experiment with the vector of actions that appears suboptimal under the estimated Q-values, $\widehat{Q}_{i,t,h}$, in iteration t. Sufficient exploration is crucial for accurately approximating the true Q-matrix and mitigating learning bias, as it ensures that all state-action pairs are sufficiently sampled, especially in complex environments. However, this comes with a tradeoff: while it enhances learning accuracy, it also increases computational burden and introduces noise, which can hinder convergence in multi-agent settings. To manage this tradeoff, the exploration probability ε_t is set to decrease monotonically toward zero as t increases. Note that we focus on asynchronous learning, defined by (2.4) and (2.6), which requires no knowledge of the underlying economic environment or information structure. In contrast, model-based synchronous learning updates all state-action pairs simultaneously in each iteration, assuming precise knowledge of the environment's structure, such as transition probabilities and payoff functions (e.g., Asker, Fershtman and Pakes, 2022, 2024). Model-based approaches are typically vulnerable to misspecification.

For reactive Q-learning, actions are determined in a manner similar to that described in (2.6), with the distinction that they are based on the same state $s_{i,t}$.

3 Model and Laboratory Design

We develop a model to establish theoretical benchmarks for analyzing the effects of coordinated AI planning on financial market stability. Subsection 3.1 outlines the economic environment. Subsection 3.2 characterizes the steady-state behavior of informed speculators under both the non-collusive Nash equilibrium and the perfect cartel benchmark. In Subsection 3.3, we examine how the intertemporal coordination among informed speculators influence financial market stability. We show that intertemporal coordination

[6] For simplicity, we use a uniform distribution, though smarter choices could improve Q-learning.

can dramatically increase financial market instability when there is a significant presence of positive-feedback investors. In Subsections 3.4 and 3.5, we elaborate on the core challenges for intertemporal coordination—time inconsistency and weak punishment. Finally, in Subsection 3.6, we present a compounded price-trigger mechanism, which can sustain the intertemporal coordination by overcoming the two core challenges.

3.1 Economic Environment

Time is discrete, denoted by episode $t = 1, 2, \ldots$ Within each episode, there are 4 subperiods. The subperiod $h = 0, 1, 2, 3$ in episode t is denoted by (t, h), where $(t, 3)$ represents the end of episode t. We assume that the period from $(t-1, 3)$ to $(t, 0)$ is considerably long while the period from $(t, 0)$ to $(t, 3)$ is very brief. In other words, all four subperiods in episode t occur in close succession towards the end of the episode.

Asset Fundamentals. All investors in the financial market can trade a long-term risky asset and a risk-free asset that offers a constant interest rate r per episode. The long-term risky asset yields a dividend d_t at the end of episode t, i.e., $(t, 3)$. The dividend follows a mean-reverting process as follows:

$$d_t = d_{t-1} + \lambda(\bar{d} - d_{t-1}) + \omega_t + \nu_t, \tag{3.1}$$

where ω_t and ν_t represent fundamental shocks to the asset in episode t, which are realized at $(t, 2)$ and $(t, 3)$, respectively. As we discuss below, an information event occurs between $(t, 0)$ and $(t, 1)$, revealing the value of ω_t to informed speculators. The shocks ω_t and ν_t are independent and identically distributed across episodes, following normal distributions with a mean of zero and standard deviations σ_ω and σ_ν, respectively.

The fundamental value of the long-term asset is denoted by $f_{t,h}$ at (t, h). This value is defined as the expected payoff generated by the long-term asset for a risk-neutral investor, discounted at the risk-free interest rate. The asset's fundamental value at $(t, 3)$ is given by

$$f_{t,3} = d_t + \mathbb{E}_t \left[\sum_{\tau=t+1}^{\infty} \frac{d_\tau}{(1+r)^{\tau-t}} \right] = \frac{1+r}{r}\bar{d} + \zeta(d_t - \bar{d}), \tag{3.2}$$

where $\zeta = (1+r)/(\lambda+r)$. Plugging (3.2) into (3.1), it follows

$$f_{t,3} = f_{t-1,3} + \lambda(\overline{f} - f_{t-1,3}) + \zeta\omega_t + \zeta\nu_t, \tag{3.3}$$

where $\overline{f} = (1+r)\overline{d}/r$.

Because the period between $(t,0)$ and $(t,3)$ is very brief, the fundamental value at $(t,0)$ is

$$f_{t,0} = \mathbb{E}_{t,0}[f_{t,3}] = \mathbb{E}_{t-1,3}[f_{t,3}] = f_{t-1,3} + \lambda(\overline{f} - f_{t-1,3}). \tag{3.4}$$

At $(t,1)$, no fundamental shock is realized yet, and thus

$$f_{t,1} = f_{t,0}. \tag{3.5}$$

At $(t,2)$, the fundamental shock ω_t is realized, and thus

$$f_{t,2} = f_{t,0} + \zeta\omega_t. \tag{3.6}$$

At $(t,3)$, the fundamental shock ν_t is realized, and thus

$$f_{t,3} = f_{t,0} + \zeta\omega_t + \zeta\nu_t, \tag{3.7}$$

which is consistent with (3.3) and (3.4).

Investors. There are four groups of investors. First, we consider the presence of infinitely many low-frequency, risk-neutral investors who lack the capacity to react promptly to the information event occurring between $(t,0)$ and $(t,1)$ through trading. Specifically, they can trade only at $(t,0)$ and $(t,3)$. Their trading activities ensure perfect alignment between the asset price and its fundamental value at both $(t,0)$ and $(t,3)$. However, since these investors are unable to trade at $(t,1)$ or $(t,2)$, the asset price in these two subperiods can deviate substantially from its fundamental value.

Second, there are $I \geq 2$ informed speculators who are able to process the information event that occurs between $(t,0)$ and $(t,1)$, transforming it into a perfect signal regarding ω_t by the time they reach $(t,1)$. These informed speculators trade based on ω_t at both $(t,1)$ and $(t,2)$, however, they do not engage in trading at $(t,0)$ and $(t,3)$, as the asset price is aligned with its fundamental value in these two subperiods. We consider linear

trading strategies of each informed speculator $i = 1, \ldots, I$ as follows

$$x_{i,t,1} = \chi_{i,1}\omega_t \quad \text{at } (t,1), \tag{3.8}$$

$$x_{i,t,2} = \chi_{i,2}\omega_t \quad \text{at } (t,2). \tag{3.9}$$

Third, there exists a continuum of positive-feedback investors of measure one. These investors tend to buy after prices rise and sell after prices fall. They do not acquire information about fundamental values either directly or through market prices that reflect informed trading flows. This behavior can create intertemporal strategic complementarities, leading to market destabilization (e.g., de Long et al., 1990; Gennotte and Leland, 1990; Hong and Stein, 1999; Goldstein, Huang and Yang, 2025). The aggregate order flows from positive-feedback investors are linear and are based on the price change observed in the preceding subperiod:

$$y_{t,2} = \mu(p_{t,1} - p_{t,0}) \quad \text{at } (t,2). \tag{3.10}$$

We assume positive-feedback investors only trade at $(t,2)$. They do not trade at $(t,1)$ because there is no well-defined previous subperiod from which the price change can be calculated.[7] It is innocuous to assume that they do not trade at $(t,3)$, as their trading would not influence the market, given that the asset price is perfectly aligned with its fundamental value at $(t,3)$.

Fourth, there exists a continuum of constrained arbitrageurs of measure one. These arbitrageurs do not learn about fundamental information through market prices that reflect informed trading flows at $(t,1)$. Instead, they engage in a buy-low, sell-high strategy based on the observed market prices and the fundamental value of the asset. The aggregate order flows from constrained arbitrageurs are as follows:

$$z_{t,1} = -\xi(p_{t,1} - f_{t,1}) \quad \text{at } (t,1), \tag{3.11}$$

$$z_{t,2} = -\xi(p_{t,2} - f_{t,2}) \quad \text{at } (t,2), \tag{3.12}$$

where $\xi > 0$. Constrained arbitrageurs do not trade at $(t,0)$ and $(t,3)$ because the asset price is aligned with its fundamental value in these two subperiods.

[7] Note that the across-episode price change, $p_{t-1,3} - p_{t,0}$, is not applicable due to the extended time span from $(t-1,3)$ to $(t,0)$.

Asset Prices. As noted above, the asset price is perfectly aligned with its fundamental value at $(t,0)$ and $(t,1)$, meaning $p_{t,0} = f_{t,0}$ and $p_{t,3} = f_{t,3}$, thanks to the trading activities of low-frequency, risk-neutral investors. The asset prices, $p_{t,1}$ and $p_{t,2}$, are determined by market-clearing conditions at $(t,1)$ and $(t,2)$, respectively, as follows:

$$p_{t,1} = f_{t,1} + \frac{\omega_t}{\xi} \sum_{j=1}^{I} \chi_{j,1}, \tag{3.13}$$

$$p_{t,2} = f_{t,2} + \frac{\mu}{\xi}(p_{t,1} - p_{t,0}) + (p_{t,1} - f_{t,1}) + \frac{\omega_t}{\xi} \sum_{j=1}^{I} \chi_{j,2}. \tag{3.14}$$

Interpretations of the Model Setting. The model setting and timing of events are summarized in Figure 1. An information event encompasses any occurrence that provides the market with new, value-relevant information about the asset. These events play a crucial role in finance and accounting as they can alter investors' perceptions regarding a firm's future cash flows, risks, or growth prospects, which in turn impacts its stock price and overall valuation. For instance, information events may include earnings announcements, such as quarterly or annual reports that disclose a firm's performance; corporate disclosures, including guidance updates, press releases, or filings (e.g., 10-K, 10-Q, 8-K); corporate actions like mergers, acquisitions, stock splits, dividends, or share repurchases; management changes such as CEO or CFO turnover, or board reshuffling; product and innovation events, such as the launch of new products, patent approvals, or R&D breakthroughs; legal and regulatory events, including lawsuits, regulatory fines, or changes in government policies that directly affect the firm; as well as crisis events like data breaches, fraud revelations, or accidents impacting operations.

The trading behavior of positive-feedback investors can arise endogenously from various mechanisms in practice. As highlighted by Gennotte and Leland (1990), portfolio insurance strategies are dynamic hedging techniques designed to replicate put options, thereby providing downside protection. These strategies inherently involve selling after market declines and buying after market increases. Consequently, portfolio insurers and stop-loss traders serve as mechanical, institutional examples of positive-feedback trading. Second, Greenwood and Shleifer (2014) provide survey-based evidence indicating that investor expectations of future returns are strongly positively correlated with past stock market performance. This relationship aligns with extrapolative beliefs about returns, which can lead to positive-feedback trading. Additionally, many investors

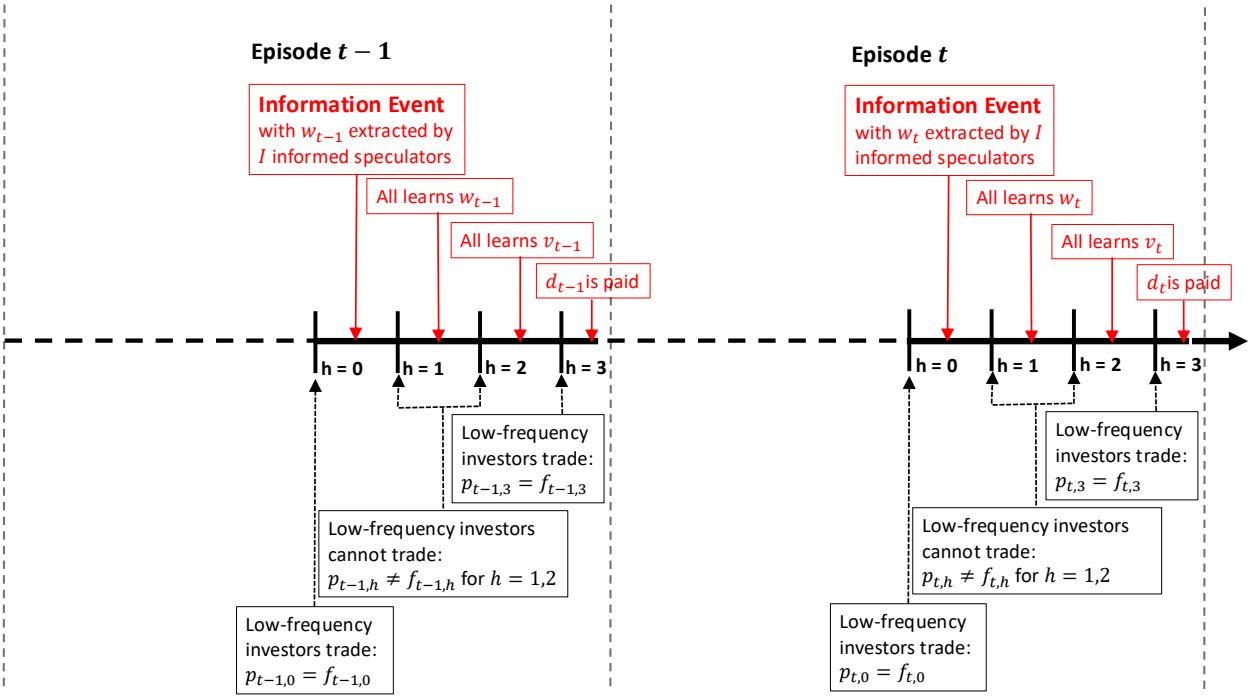


Figure 1: Timeline of the model.

adopt momentum trading strategies, chasing recent winners and further contributing to positive-feedback effects. This behavior is evident among mutual funds (e.g., Grinblatt, Titman and Wermers, 1995; Wermers, 1999) as well as retail investors (e.g., Barber, Odean and Zhu, 2008). Third, investors who believe in the market feedback effect—whereby managers improve their decision-making based on changes in market valuations, thereby enhancing firm value—are examples of positive-feedback investors (e.g., Ozdenoren and Yuan, 2008; Goldstein, 2022). Lastly, margin-based traders are by nature positive-feedback investors (e.g., Kyle and Xiong, 2001; Brunnermeier and Pedersen, 2008; Geanakoplos, 2010).

The presence of constrained arbitrageurs is addressed by the canonical theoretical framework proposed by Gromb and Vayanos (2010). They demonstrate that when arbitrage capital is inelastic, due to constraints related to leverage, funding, or risk, market prices can persistently deviate from fundamental values, even when rational arbitrageurs are present. Building on this framework, Gromb and Vayanos (2018) show that collateral-based financing constraints render the aggregate demand of arbitrageurs for risky assets highly inelastic. The concept of constrained arbitrageurs shares similarities with information-insensitive investors in the model by Dou, Goldstein and Ji (2025), who

engage in trading the asset and collectively create a downward-sloping demand curve within each trading period, as in Kyle and Xiong (2001) and Vayanos and Vila (2021). These investors, which include retail traders using technical analysis or institutional investors seeking hold-to-maturity positions to hedge short-term risks, typically do not respond to real-time information regarding the asset's fundamental value. Instead, they trade based on the current price and in the direction of the asset's expected long-term value. While we do not explicitly endogenize the behavior of these constrained arbitrageurs, it is worth noting that they need not be behaviorally biased. They may find it optimal not to pay attention to short-term fluctuations or may act in such a manner due to institutional constraints.

3.2 Theoretical Benchmarks

To characterize the steady-state behavior of informed speculators, we derive two theoretical benchmarks. First, we analyze the non-collusive Nash equilibrium, in which informed speculators do not coordinate with one another. Instead, each informed speculator takes the equilibrium trading strategies of other informed speculators as given and optimizes her own strategies to maximize profits. Second, we explore the perfect-cartel benchmark, where all informed speculators can perfectly coordinate on intertemporal trading without any incentives to deviate, effectively acting as a monopoly. We denote the variables in these two benchmarks using superscripts N and M, respectively.

Non-Collusive Nash Equilibrium. In each episode t of the non-collusive Nash equilibrium, informed speculators play a one-shot game, in which their trading strategies do not influence any outcomes in future episodes. As a result, the dynamic optimization problem faced by each informed speculator can be separately solved within each episode. Let $V_{i,1}^N(\omega_t)$ denote the value function of informed speculator i at $(t,1)$, after receiving a perfect signal about ω_t. Similarly, let $V_{i,2}^N(\omega_t, p_{t,1})$ denote the value function of informed speculator i at $(t,2)$. The value functions do not depend on t (i.e., no subscript t) because the trading environment is stationary.

Focusing on linear trading strategies, $x_{i,1}(\omega_t) \equiv \chi_{i,1}\omega_t$ and $x_{i,2}(\omega_t) \equiv \chi_{i,2}\omega_t$ for

$i = 1, \ldots, I$, we characterize informed speculator i's problem as follows:

$$V_{i,1}^N(\omega_t) = \max_{\chi_{i,1}} \mathbb{E}_t \left[(p_{t,3} - p_{t,1})\chi_{i,1}\omega_t + V_{i,2}^N(\omega_t, p_{t,1}) \middle| \omega_t \right], \tag{3.15}$$

$$V_{i,2}^N(\omega_t, p_{t,1}) = \max_{\chi_{i,2}} \mathbb{E}_t \left[(p_{t,3} - p_{t,2})\chi_{i,2}\omega_t | \omega_t \right], \tag{3.16}$$

where expectation $\mathbb{E}_t[\cdot|\omega_t]$ is taken over ν_t. The informed trader i's problems (3.15)-(3.16) are solved by backward induction. First, given $p_{t,1}$ and all other speculators' equilibrium strategies $\{\chi_{j,2}\}_{j\neq i}$ at $(t,2)$, informed speculator i optimally chooses $\chi_{i,2}$ that solves (3.16). The first-order condition with respect to $\chi_{i,2}$ yields:

$$\chi_{i,2} = -\frac{1}{2}\left[(\xi + \mu)\frac{p_{t,1} - f_{t,1}}{\omega_t} + \sum_{j\neq i}\chi_{j,2} \right]. \tag{3.17}$$

In the symmetric equilibrium, we have $\chi_{j,2} = \chi_2^N$ for all $j = 1, \ldots, I$. Thus,

$$\chi_2^N(p_{t,1}; \omega_t, f_{t,1}) = -\frac{(\xi + \mu)(p_{t,1} - f_{t,1})}{\omega_t(I+1)}. \tag{3.18}$$

Substituting (3.18) into (3.16) gives $V_{i,2}^N(\omega_t, p_{t,1})$, which is further substituted into (3.15). Then, the first-order condition with respect to $\chi_{i,1}$ yields:

$$\chi_{i,1} = \xi\zeta - \sum_{j\neq i}\chi_{j,1} + \frac{2(\mu+\xi)^2}{(I+1)^2\xi^2}\sum_{j\neq i}\chi_{j,1}. \tag{3.19}$$

In the symmetric Nash equilibrium, we have $\chi_{j,1} = \chi_1^N$ for all $j = 1, \ldots, I$. Thus, we derive the trading strategies in the non-collusive Nash equilibrium according to (3.18) and (3.19) as follows:

$$\chi_1^N = \frac{\xi(I+1)^2\zeta}{(I+1)^3 - 2I(1+\mu/\xi)^2}, \tag{3.20}$$

$$\chi_2^N = -\frac{(1+\mu/\xi)I}{I+1}\chi_1^N, \tag{3.21}$$

where parameters have to satisfy $(I+1)^2 > (1+\mu/\xi)^2 I$ to ensure that informed speculators' profits are finite in the non-collusive Nash equilibrium.

The trading profits of informed speculator i in the non-collusive Nash equilibrium are

given by:

$$\pi_{i,t,1}^N = (p_{t,3} - p_{t,1})\chi_1^N \omega_t,$$
$$\pi_{i,t,2}^N = (p_{t,3} - p_{t,2})\chi_2^N \omega_t.$$

The total profits and total order flows of informed speculator i in the non-collusive Nash equilibrium are $\pi_{i,t}^N \equiv \pi_{i,t,1}^N + \pi_{i,t,2}^N$ and $x_i^N(\omega_t) \equiv (\chi_1^N + \chi_2^N)\omega_t$.

Perfect Cartel Benchmark. We now consider the benchmark in which all informed speculators coordinate perfectly with each other on intertemporal trading, effectively trading as a monopoly to maximize the aggregate profits of all informed speculators. Similar to the non-collusive Nash equilibrium, the dynamic optimization problem faced by the cartel can also be solved separately within each episode. Let $V_{i,1}^M(\omega_t)$ denote the value function of informed speculator i at $(t,1)$, after receiving a perfect signal reagrding ω_t. Similarly, let $V_{i,2}^M(\omega_t, p_{t,1})$ denote the value function of informed speculator i at $(t,2)$. Without loss of generality, we focus on the symmetric equilibrium where all informed speculators adopt the same linear trading strategies, denoted as χ_1 and χ_2. The cartel's optimization problem can be formulated as follows:

$$V_{i,1}^M(\omega_t) = \max_{\chi_1} \ \mathbb{E}_t\left[(p_{t,3} - p_{t,1})\chi_1\omega_t + V_{i,2}^M(\omega_t, p_{t,1}) \middle| \omega_t\right], \tag{3.22}$$
$$V_{i,2}^M(\omega_t, p_{t,1}) = \max_{\chi_2} \ \mathbb{E}_t\left[(p_{t,3} - p_{t,2})\chi_2\omega_t \middle| \omega_t\right], \tag{3.23}$$

where expectation $\mathbb{E}_t[\cdot|\omega_t]$ is taken over ν_t. We solve the cartel's problem by backward induction. First, given $p_{t,1}$, solving (3.23) yields:

$$\chi_2^M(p_{t,1}; \omega_t, f_{t,1}) = -\frac{(\xi + \mu)(p_{t,1} - f_{t,1})}{2\omega_t I}. \tag{3.24}$$

Substituting (3.24) into (3.23) gives $V_{i,2}^M(\omega_t, p_{t,1})$, which is further substituted into (3.22). Then, the first-order condition with respect to χ_1 yields:

$$\chi_1^M = \frac{2\xi\zeta}{4I - I(1 + \mu/\xi)^2}, \tag{3.25}$$
$$\chi_2^M = -\frac{1 + \mu/\xi}{2}\chi_1^M, \tag{3.26}$$

where parameters have to satisfy $\mu < \xi$ to ensure that informed speculators' profits are finite in the perfect cartel benchmark.

The trading profits of informed speculator i in the perfect cartel benchmark are given by:

$$\begin{aligned}\pi^M_{i,t,1} =&(p_{t,3} - p_{t,1})\chi^M_1\omega_t,\\ \pi^M_{i,t,2} =&(p_{t,3} - p_{t,2})\chi^M_2\omega_t.\end{aligned}$$

The total profits and total order flows of informed speculator i in the perfect cartel benchmark are $\pi^M_{i,t} \equiv \pi^M_{i,t,1} + \pi^M_{i,t,2}$ and $x^M_i(\omega_t) \equiv (\chi^M_1 + \chi^M_2)\omega_t$.

Measure for the Degree of Collusion and Financial Instability. The perfect cartel benchmark represents perfect intertemporal coordination among the $I \geq 2$ informed dynamic speculators, which could be challenging (if not impossible) to maintain in practice. In a multi-period dynamic game with imperfect monitoring, informed speculators can sustain a collusive equilibrium where each informed speculator's profit $\pi_{i,t}$ lies between $\pi^N_{i,t}$ and $\pi^M_{i,t}$ by employing reward-punishment strategies, such as the price-trigger strategies described by Green and Porter (1984). We discuss how intertemporal coordination can be sustained using a compounded price-trigger strategy in Subsection 3.6. In theory, there exist infinite equilibria that exhibit intertemporal coordination. The capacity for coordination can be quantified by the normalized trading profitability of informed speculators in the collusive equilibrium, defined as follows:

$$\Delta = \frac{\mathbb{E}[\pi_{i,t}] - \mathbb{E}[\pi^N_{i,t}]}{\mathbb{E}[\pi^M_{i,t}] - \mathbb{E}[\pi^N_{i,t}]}, \tag{3.27}$$

where $\mathbb{E}[\cdot]$ represents the unconditional expectation with respect to ω_t and ν_t. The value of Δ lies between 0 and 1, with a higher value indicating a greater degree of intertemporal coordination among informed speculators.

In addition, to evaluate the instability of financial markets, we calculate the fourth moment of changes in the asset's price and fundamental value as follows:

$$m(\Delta p_{t,h}) = \mathbb{E}\left[\frac{1}{3}\sum_{h=1}^{3}(\Delta p_{t,h} - \overline{\Delta p}_t)^4\right] \quad \text{and} \quad m(\Delta f_{t,h}) = \mathbb{E}\left[\frac{1}{3}\sum_{h=1}^{3}(\Delta f_{t,h} - \overline{\Delta f}_t)^4\right], \tag{3.28}$$

where $\Delta p_{t,h} = p_{t,h} - p_{t,h-1}$, $\overline{\Delta p}_t = \mathbb{E}\left[\frac{1}{3}\sum_{h=1}^{3}\Delta p_{t,h}\right]$, $\Delta f_{t,h} = f_{t,h} - f_{t,h-1}$, $\overline{\Delta f}_t = \mathbb{E}\left[\frac{1}{3}\sum_{h=1}^{3}\Delta f_{t,h}\right]$, and $\mathbb{E}[\cdot]$ represents the unconditional expectation with respect to ω_t and ν_t. The instability of financial markets is measured by

$$\mathcal{S} = \ln\left(\frac{m(\Delta p_{t,h})}{m(\Delta f_{t,h})}\right), \tag{3.29}$$

where $\mathcal{S} < 0$ indicates that changes in asset prices are less dramatic relative to changes in asset fundamentals while $\mathcal{S} > 0$ indicates the opposite. A larger $\mathcal{S}$ indicates higher financial market instability.

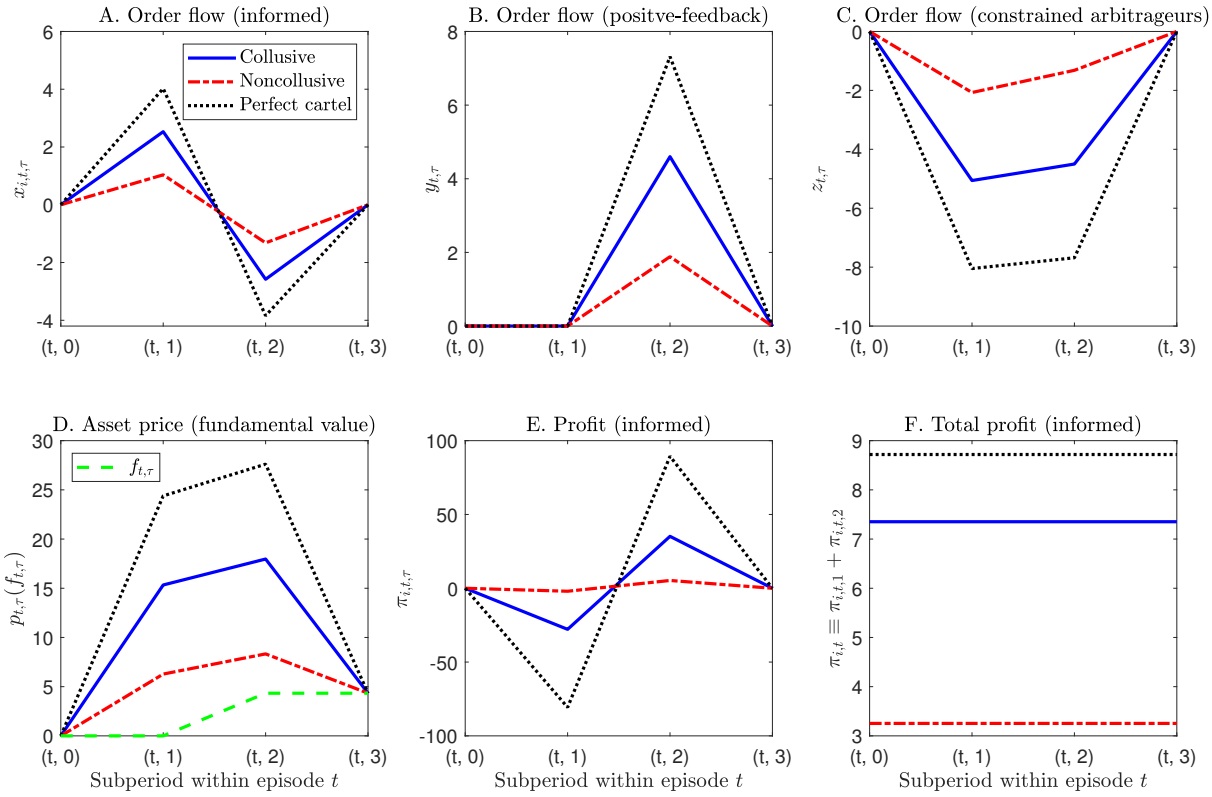
3.3 Impact of Intertemporal Coordination on Market Stability

We study the impact of intertemporal coordination among informed speculators on the stability of financial markets. There exist infinitely many collusive equilibria that satisfy incentive-compatibility constraints. For illustrative purposes, we focus on the collusive equilibrium where informed speculators implement a trading strategy $\{\chi_1^C, \chi_2^C\}$ that lies in the middle of the strategies in the non-collusive Nash equilibrium and perfect cartel benchmark, i.e., $\chi_h^C = (\chi_h^N + \chi_h^M)/2$ for $h = 1, 2$. We discuss the core challenges for sustaining collusion in Sections 3.4 and 3.5, and propose a strategy to overcome these challenges in Section 3.6. Moreover, to cleanly illustrate the key economic mechanisms, we focus on a trading environment with two informed speculators ($I = 2$) and assume that the realized shocks are $\omega_t = 1$ and $\nu_t = 0$. By setting $\nu_t = 0$, we simplify the illustration, as the realized fundamental value $f_{t,3}$ at $(t, 3)$ is equivalent to the expected fundamental value $\mathbb{E}_t[f_{t,3}|\omega_t]$, given that $\nu_t \sim N(0, \sigma_\nu)$.

In the remainder of this subsection, we first consider the baseline trading environment where positive-feedback investors have a significant presence (i.e., high μ). Next, we analyze the other trading environment where positive-feedback investors have limited presence (i.e., low μ).

Baseline Environment: Intertemporal Coordination Destabilizes Financial Markets. In the baseline trading environment, we show that the intertemporal coordination among informed speculators can substantially amplify the magnitude of asset price changes across periods, leading to instability in financial markets.

Figure 3 illustrates the results, with the solid, dash-dotted, and dotted lines represent-



Note: In each panel, the solid line represents the collusive equilibrium, where the collusive order flows are set at $\chi_h^C = (\chi_h^N + \chi_h^M)/2$ for $h = 1, 2$. The dash-dotted and solid lines indicate the non-collusive Nash equilibrium and the perfect cartel benchmark, respectively. The dashed line in panel D represents the asset's fundamental value. Parameters are set at $\mu = 0.3, I = 2, r = 0.04, \lambda = 0.2, \xi = 0.33$, and $f_0 = 0$. The realized shocks are $\omega_t = 1$ and $\nu_t = \mathbb{E}_{t,0}[\nu_t] = 0$.

Figure 2: Order flows, asset prices, and informed trading profits (high μ).

ing the collusive equilibrium, the non-collusive Nash equilibrium, and the perfect cartel benchmark, respectively. As shown in panel A of Figure 2, across all three equilibria, informed speculators take long positions in the asset at $(t, 1)$ and short at $(t, 2)$. This long-short intertemporal trading strategy allows informed speculators to manipulate positive-feedback investors' trading flows so as to increase their own trading profits. Specifically, at $(t, 1)$, informed speculators observe a perfect signal of ω_t prior to making trading decisions. This allows them to accurately infer the expected asset price at $(t, 3)$, i.e., $\mathbb{E}_t[p_{t,3}|\omega_t] = \mathbb{E}_t[f_{t,3}|\omega_t]$, which equals $f_{t,3}$ given our assumption that $\nu_t = 0$. Even though $p_{t,1} > \mathbb{E}_t[p_{t,3}|\omega_t]$ (panel D of Figure 2), informed speculators do not take short positions at $(t, 1)$; instead, they go long on the asset at $(t, 1)$ by choosing $x_{i,t,1} > 0$, result-

ing in substantial losses, given by $\pi_{i,t,1} = (p_{t,3} - p_{t,1})x_{i,t,1} < 0$ (panel E of Figure 2). The willingness of informed speculators to incur large losses at $(t,1)$ stems from their ability to drive up the asset price $p_{t,1}$ at $(t,1)$. This increase in $p_{t,1}$ incentivizes positive-feedback investors, whose asset demand $y_{t,2}$ is proportional to $p_{t,1} - p_{t,0}$ (equation (3.10)), to buy a substantial amount of assets at $(t,2)$ (panel B of Figure 2), thereby significantly boosting the asset price $p_{t,2}$ at $(t,2)$. With a higher $p_{t,2}$, informed speculators can realize significant profits (panel E of Figure 2), given by $\pi_{i,t,2} = (p_{t,3} - p_{t,2})x_{i,t,2}$, by taking short positions at $(t,2)$, i.e., $x_{i,t,2} < 0$ (panel A of Figure 2). Through this intertemporal manipulation, they strategically benefit from the trading behavior of positive-feedback investors, ultimately achieving positive total profits, with $\pi_{i,t} = \pi_{i,t,1} + \pi_{i,t,2} > 0$ (panel F of Figure 2). Meanwhile, panel C of Figure 2 shows that constrained arbitrageurs adopt short positions at both $(t,1)$ and $(t,2)$ as the asset price at these times, $p_{t,1}$ and $p_{t,2}$, are both higher than their respective fundamental values, i.e., $p_{t,1} > f_{t,1}$ and $p_{t,2} > f_{t,2}$, as shown in panel D of Figure 2.

The above intuition is first formalized by the model of de Long et al. (1990). However, our mechanism differs by emphasizing the strategic intertemporal coordination among oligopolistic informed speculators. In the model of de Long et al. (1990), informed speculators are atomistic and take equilibrium asset prices as given. As a result, their decision-making in each subperiod is essentially static, lacking strategic considerations or multi-period planning. In contrast, our theoretical analysis highlights the collective actions of informed speculators in both creating and collapsing an asset price bubble. In particular, relative to the non-collusive Nash equilibrium, panel A of Figure 2 illustrates that informed speculators trade more aggressively by taking larger long positions at $(t,1)$ in the collusive equilibrium, coordinating in a way to generate a larger asset price bubble at $(t,1)$. Intuitively, when one informed speculator increases its asset purchase at $(t,1)$, the asset price $p_{t,1}$ rises, which incentivizes positive-feedback investors to buy more assets at $(t,2)$, driving up the asset price $p_{t,2}$. Given that informed speculators take short positions at $(t,2)$, the higher $p_{t,2}$ in turn increases the profit of every informed speculator at $(t,2)$, resulting in positive externality. This positive externality suggests that, to maximize aggregate profits, both informed speculators should adopt a more aggressive trading strategy at $(t,1)$ than they would in the non-collusive Nash equilibrium. Moreover, relative to the non-collusive Nash equilibrium, informed speculators take larger short positions at $(t,2)$ in the collusive equilibrium. This is simply because, by observing the higher asset price $p_{t,1}$ realized at $(t,1)$, informed speculators know that positive-

feedback investors will buy a significant amount of assets at $(t,2)$. This influx of asset purchases from positive-feedback investors helps to offset the negative price impact of asset sales from informed speculators, motivating them to take larger short positions. Importantly, however, we emphasize that, conditional on the asset price $p_{t,1}$ realized at $(t,1)$, the short positions at $(t,2)$ taken by informed speculators are smaller than what these speculators would do if they trade non-collusively in the Nash equilibrium at $(t,2)$. In other words, once $p_{t,1}$ has already been realized, coordinated trading at $(t,2)$ requires informed speculators to gradually unwind their positions, compared to how they would trade non-collusively in the Nash equilibrium. The intuition aligns with the collusion mechanism in a standard Cournot competition game, where each firm restricts its output to raise market prices and increase profits.

Taken together, intertemporal coordination among informed speculators involves two steps. First, at $(t,1)$, these speculators trade aggressively to create a large asset price bubble, incurring substantial losses. Second, at $(t,2)$, they trade conservatively to mitigate the resulting price impact, gradually unwinding their positions to extract profits. As we elaborate in Sections 3.4 and 3.5 below, such intertemporal coordination is difficult to sustain in equilibrium due to issues regarding time inconsistence and weak punishment.

As a result of the intertemporal coordination among informed speculators, asset prices exhibit significant fluctuations (panel D of Figure 2), contributing to instability in financial markets. To gauge the quantitative magnitude of this effect, we simulate the model over $T = 10,000$ episodes, assuming that $\omega_t \sim N(0,\sigma_w)$ and $\nu_t \sim N(0,\sigma_\nu)$, with $\sigma_w = 1$ and $\sigma_\nu = 0.03$. Using the simulated time-series data, we calculate the stability of financial markets according to the measure defined in (3.29) as follows:

$$\mathcal{S}^N = 1.6 \text{ and } \mathcal{S}^C = 5.5.$$

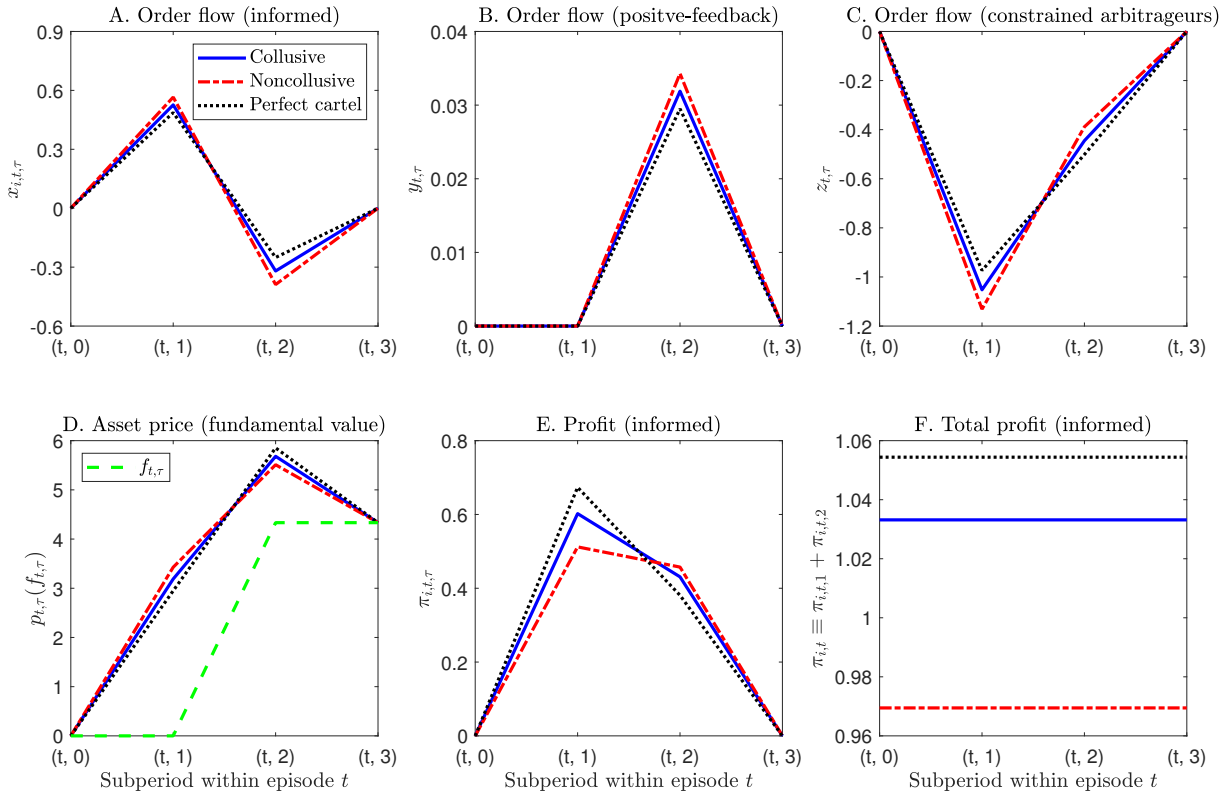
The fact that $\mathcal{S}^N, \mathcal{S}^C > 0$ suggests that, in the trading environment with significant presence of positive-feedback investors, the strategic interactions among informed speculators lead to much more pronounced changes in the asset price than changes in the asset's fundamental value.[8] Notably, the intertemporal coordination among informed speculators further increases the fourth moment of changes in the asset price, significantly destabilizing financial markets.

[8]As the number of informed speculators $I \to \infty$, it follows that $\mathcal{S}^N = 0$ because informed speculators lose their market power due to perfect competition.

Low-μ Environment: Intertemporal Coordination Stabilizes Financial Markets. We now show that in the trading environment where positive-feedback investors have a limited presence (i.e., low μ), the intertemporal coordination among informed speculators can contribute to stabilizing asset prices, though the magnitude of this impact is small.

Panel A of Figure 3 shows that across all three equilibria, informed speculators take long positions in the asset at $(t,1)$ but short positions at $(t,2)$. This long-short intertemporal trading strategy resembles the strategy observed in the baseline trading environment, where positive-feedback investors have a significant presence (panel A of Figure 2). However, the underlying mechanism driving this strategy differs significantly, as informed speculators have minimal intertemporal concerns about manipulating the behavior of positive-feedback investors due to their limited presence. Instead, their trading strategies here are primarily driven by the difference between the current asset price and the expected asset price at $(t,3)$. Specifically, at $(t,1)$, informed speculators observe ω_t ahead of other investors. This allows them to accurately infer the expected asset price at $(t,3)$, i.e., $\mathbb{E}_t[p_{t,3}|\omega_t] = \mathbb{E}_t[f_{t,3}|\omega_t]$, which equals $f_{t,3}$ given our assumption that $\nu_t = 0$. As shown in panel D of Figure 3, in all three equilibria, the current asset price $p_{t,1}$ is lower than $f_{t,3}$. Thus, informed speculators go long on the asset at $(t,1)$. By holding this long position $x_{i,t,1} > 0$ until $(t,3)$, they can expect to realize a positive profit, given by $\pi_{i,t,1} = (p_{t,3} - p_{t,1})x_{i,t,1}$ (panel E of Figure 3). Similarly, at $(t,2)$, informed speculators short the asset because the asset price $p_{t,2}$ exceeds $f_{t,3}$ (panel D of Figure 3). By maintaining this short position $x_{i,t,2} < 0$ until $(t,3)$, they will also realize a positive profit, $\pi_{i,t,1} = (p_{t,3} - p_{t,2})x_{i,t,2}$, in expectation (panel E of Figure 3). However, by comparing panel F in Figures 2 and 3, it is clear that these profits are notably smaller than those attainable in the baseline trading environment where positive-feedback investors have a significant presence.

Comparing the solid and dash-dotted lines in panel A of Figure 3, it is evident that coordination among informed speculators require them to trade more conservatively at both $(t,1)$ and $(t,2)$. Unlike the baseline trading environment, coordination in the low-μ environment is largely static, without much intertemporal consideration due to the limited presence of positive-feedback investors. Thus, the coordinated trading strategies of informed speculators in each subperiod can be viewed as collusion in a static Cournot competition game. Specifically, at $(t,1)$, when one informed speculator buys more assets, the asset price $p_{t,1}$ rises, consequently reducing the profit for the other informed speculator. To maximize their joint profits, the two informed speculators in the



Note: In each panel, the solid line represents the collusive equilibrium, where the collusive order flows are set at $\chi_h^C = (\chi_h^N + \chi_h^M)/2$ for $h = 1, 2$. The dash-dotted and solid lines indicate the non-collusive Nash equilibrium and the perfect cartel benchmark, respectively. The dashed line in panel D represents the asset's fundamental value. Parameters are set at $\mu = 0.01, I = 2, r = 0.04, \lambda = 0.2, \xi = 0.33$, and $f_0 = 0$. The realized shocks are $\omega_t = 1$ and $\nu_t = \mathbb{E}_{t,0}[\nu_t] = 0$.

Figure 3: Order flows, asset prices, and informed trading profits (low μ).

collusive equilibrium must adopt a more conservative trading approach by taking smaller long positions compared to what they would do in the non-collusive Nash equilibrium. This results in the asset price $p_{t,1}$ at $(t, 1)$ in the collusive equilibrium being lower than that in the non-collusive Nash equilibrium (panel D of Figure 3). Similarly, at $(t, 2)$, the two informed speculators in the collusive equilibrium also take smaller short positions and trade more conservatively than they would in the non-collusive Nash equilibrium, knowing that selling assets imposes a negative externality on each other's profits by depressing the asset price $p_{t,2}$.[9] As shown in panel F of Figure 3, this conservative trading

[9]It is worth noting that if there are infinitely many informed speculators (i.e., $I \to \infty$), the equilibrium asset prices $p_{t,1}$ and $p_{t,2}$ in the non-collusive Nash equilibrium will converge to $f_{t,3}$ due to perfect competi-

approach allows informed speculators in the collusive equilibrium to achieve a higher total profit in episode t compared to the non-collusive Nash equilibrium. However, informed speculators in the collusive equilibrium do not consistently earn higher profits in every sub-period than in the non-collusive Nash equilibrium. For example, panel E of Figure 3 reveals that the profit derived from trading $x_{i,t,2}$ at $(t,2)$ in the collusive equilibrium is lower than that in the non-collusive Nash equilibrium. This occurs because the long position taken at $(t,1)$ influences the profit earned from $x_{i,t,2}$ through its impact on price $p_{t,2}$. Specifically, in the collusive equilibrium, trading more conservatively by holding a smaller long position $x_{i,t,1}$ at $(t,1)$ negatively impacts $p_{t,2}$, thus reducing the profit obtained from $x_{i,t,2}$.

We evaluate the stability of financial markets according to the measure defined in (3.29), based on the simulated data over $T = 10,000$ episodes, assuming that $\omega_t \sim N(0, \sigma_w)$ and $v_t \sim N(0, \sigma_v)$, with $\sigma_w = 1$ and $\sigma_v = 0.03$. The patterns observed in panel D of Figure 3 imply

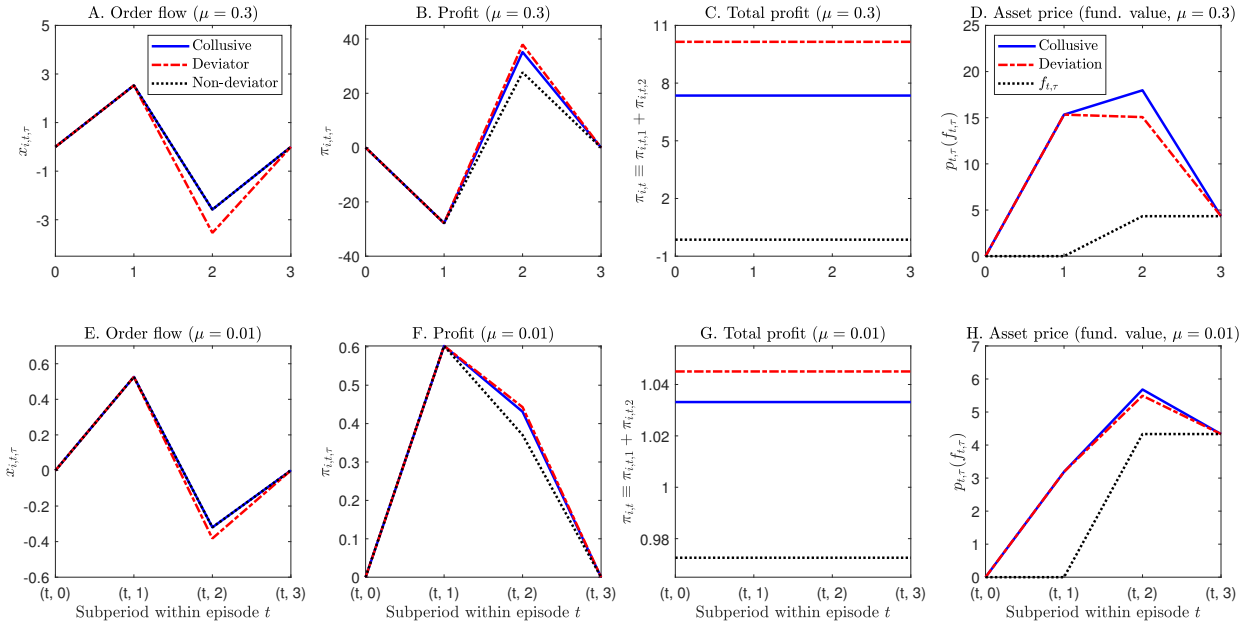
$$\mathcal{S}^N = -0.80 \text{ and } \mathcal{S}^C = -0.82.$$

These negative values suggest that, in the trading environment with a limited presence of positive-feedback investors, the strategic interactions among informed speculators lead to less dramatic changes in asset prices compared to changes in the asset's fundamental value. Moreover, in sharp contrast to the baseline trading environment, the coordination among informed speculators mildly stabilizes financial markets.

3.4 Challenge for Sustaining Collusion at $(t,2)$: Time Inconsistence

In the baseline trading environment with a significant presence of positive-feedback investors (i.e., high μ), intertemporal coordination among informed speculators require them to trade aggressively at $(t,1)$ to create a substantial asset price bubble, and then, at $(t,2)$, they gradually unwind their positions to realize profits. As illustrated in panel E of Figure 2, under this intertemporal coordination, informed speculators incur significant losses at $(t,1)$ in order to achieve substantial profits at $(t,2)$—these speculators are willing to coordinate at $(t,1)$ specifically because they anticipate ongoing coordination at $(t,2)$. This, obviously, creates a pronounced time inconsistency issue, as once a large bubble is established at $(t,1)$, each informed speculator has a strong incentive to deviate from the

tion, resulting in no informed speculator being able to profit despite receiving a perfect signal about ω_t at $(t,1)$.



Note: In the six panels on the left, the solid line represents the collusive equilibrium, where the collusive order flows are set at $\chi_h^C = (\chi_h^N + \chi_h^M)/2$ for $h = 1, 2$. The dash-dotted line represents the deviator, who optimally selects its deviation order flow at $(t, 2)$, conditional on itself adopts the collusive order flow at $(t, 1)$ and the other informed speculator adheres to the collusive order flow at both $(t, 1)$ and $(t, 2)$. The dotted line represents the non-deviator. In the two panels on the right, the solid line represents the collusive equilibrium, the dash-dotted line represents the equilibrium with unilateral deviation, and the dotted line represents the asset's fundamental value. The top four panels illustrate the trading environment with $\mu = 0.3$, while the bottom four panels represent the environment with $\mu = 0.01$. The other parameters are set at $I = 2, r = 0.04, \lambda = 0.2, \xi = 0.33$, and $f_0 = 0$. The realized shocks are $\omega_t = 1$ and $\nu_t = \mathbb{E}_{t,0}[\nu_t] = 0$.

Figure 4: Illustration of the time inconsistency issue for sustaining collusion.

coordinated trading strategy at $(t, 2)$ by trading aggressively.

In the top four panels of Figure 4, we illustrate this time inconsistency issue in the baseline trading environment using the same parameter values as those in Figure 2. Specifically, we consider unilateral deviation by allowing the deviator to optimally choose χ_2 at $(t, 2)$ to maximize her total profit, $\pi_t = \pi_{t,1} + \pi_{t,2}$, while maintaining her decision at $(t, 1)$ fixed at χ_1^C. Meanwhile, the non-deviator's decisions at $(t, 1)$ and $(t, 2)$ remain unchanged at χ_1^C and χ_2^C, respectively. Panel A of Figure 4 shows that the deviator's choice $\chi_2 < \chi_2^C < 0$, indicating that the deviator trades more aggressively at $(t, 2)$ than what she would do in the collusive equilibrium. This significantly increases the deviator's profit at $(t, 2)$ and boosts total profit, while simultaneously reducing the non-deviator's profit (panels B and C of Figure 4); the deviation is so pronounced that it even results

in losses for the non-deviator.[10] Panel D of Figure 4 shows that in the equilibrium with unilateral deviation, the asset price bubble begins to crash at $(t, 2)$, one subperiod earlier than the crash time, $(t, 3)$, observed in the collusive equilibrium.

In contrast, the bottom four panels of Figure 4 illustrate what happens after a unilateral deviation at $(t, 2)$ in the trading environment with a limited presence of positive-feedback investors (i.e., low μ). Panel G of Figure 4 illustrates that the deviator benefits at the expense of the non-deviator. However, the magnitude of these gains and losses is much smaller compared with those shown in panel C of Figure 4. Thus, while the time inconsistency issue also exists in the low-μ environment, it is much less pronounced than the baseline trading environment with high μ. Intuitively, as discussed in Section 3.3, coordination in the low-μ environment is largely static, allowing informed speculators to separately and independently coordinate within each subperiod $(t, 1)$ and $(t, 2)$, without the need for intertemporal planning as required in the baseline trading environment. Consequently, in the low$-\mu$ environment, deviations from coordination at $(t, 2)$ yield only small gains for the deviator, as these gains primarily come from the benefits of one-subperiod coordination. In contrast, in the high$-\mu$ environment, deviations from coordination at $(t, 2)$ can generate substantial gains, which come from the benefits of two-subperiod coordination—informed speculators incur significant losses at $(t, 1)$ to collectively create large bubbles from which they can profit substantially at $(t, 2)$.

The pronounced time inconsistency issue in the baseline trading environment poses significant challenges for intertemporal coordination among informed speculators. In Section 3.6, we demonstrate that such coordination can be maintained if informed speculators adopt a compounded price-trigger strategy.

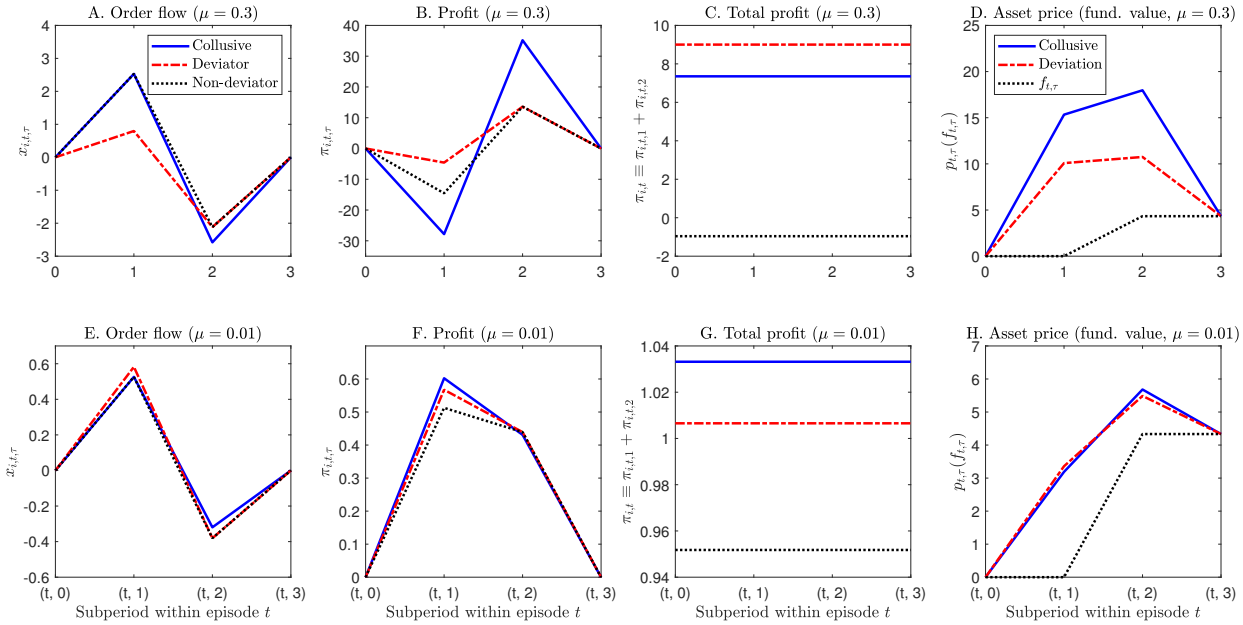
[10] Conceptually, the time inconsistency issue in our model is slightly distinct from the standard notion of time inconsistency. In a standard scenario, after decisions are made in the first period, only one player can renege on its prior commitment and alter its decision in the second period to maximize its own payoff, often at the expense of the other player. For example, a standard time inconsistency issue is that the government announces a policy rule and people make optimal decisions based on the expectation of continuation of the policy rule. However, after people making their decisions, the government can benefit itself by changing its policy rule whereas people can no longer immediately react to the new policy rule. In contrast, in our model, once decisions are made in the first period, both informed speculators have the option to renege on their prior commitment and adjust their trading decisions in the second period due to their symmetric positions in the game. However, reneging becomes profitable only if the other informed speculator does not also renege, which is difficult to prevent because reneging is always the best response regardless of the other speculator's action. As a result, both informed speculators may end up reneging at $(t, 2)$ in equilibrium, leading to trading outcomes in the non-collusive Nash equilibrium that do not yield increased profits for either speculator. These non-collusive outcomes resemble a prisoner's dilemma scenario.

3.5 Challenge for Sustaining Collusion at $(t,1)$: Weak Punishment

Having discussed the deviation incentive at $(t,2)$ due to time inconsistency, we now turn our attention to the deviation incentive at $(t,1)$ in the baseline trading environment characterized by a significant presence of positive-feedback investors. As discussed above, intertemporal coordination at $(t,1)$ requires that informed speculators trade aggressively and incur significantly losses to collectively generate a large asset price bubble. They can then profit by gradually unwinding their asset positions at $(t,2)$. Each informed speculator has the incentive to deviate at $(t,1)$ by trading less aggressively to minimize her losses, effectively free-riding on the aggressive trading of other informed speculators. Such deviations at $(t,1)$ cannot be effectively punished in the next subperiod, $(t,2)$, because without large asset price bubbles, even a shift to non-collusive trading in the Nash equilibrium fails to deter the strong incentive to deviate.

In the top four panels of Figure 5, we illustrate the weak punishment issue in the baseline trading environment using the same parameter values as those in Figure 2. Specifically, we consider unilateral deviation by allowing the deviator to optimally choose χ_1 at $(t,1)$ to maximize her total profit, $\pi_t = \pi_{t,1} + \pi_{t,2}$, while the non-deviator's decision at $(t,1)$ remain fixed at χ_1^C. In making her choice at $(t,1)$, the deviator rationally anticipates that her deviation will be punished in the next subperiod, as both informed speculators will trade non-collusively in the Nash equilibrium at $(t,2)$ according to (3.18), based on the asset price $p_{t,1}$ realized at $(t,1)$. Panel A of Figure 5 illustrates that the deviator significantly reduces its order flow, nearly to zero, compared to the order flow in the collusive equilibrium. This strategy allows the deviator to avoid incurring large losses at $(t,1)$ (panel B of Figure 5). Consequently, the deviator's total profit is substantially higher than what she would achieve in the collusive equilibrium, even after accounting for the losses caused by the non-collusive trading punishment in the next subperiod, $(t,2)$. In contrast, the non-deviator experiences significant losses because he invests a considerable amount of funds at $(t,1)$ to generate the asset price bubble but fails to collect enough profit at $(t,2)$ due to the deviator's defection. As shown in panel D of Figure 5, the size of the asset price bubble at $(t,1)$ and $(t,2)$ is much smaller in the equilibrium with unilateral deviation compared to the size in the non-collusive equilibrium. Without large bubbles, the non-collusive trading in the Nash equilibrium at $(t,2)$ is not an effective punishment strategy to deter deviation incentives at $(t,1)$.

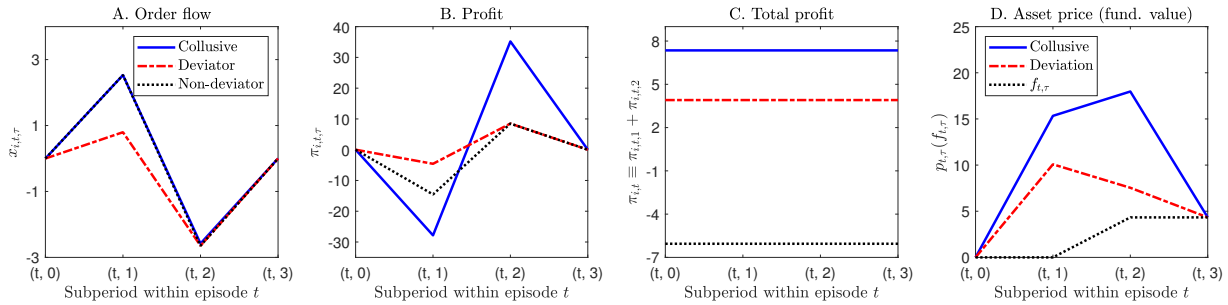
In contrast, the bottom four panels of Figure 5 illustrate what happens after a unilateral deviation at $(t,2)$ in the trading environment with a limited presence of positive-feedback



Note: In the six panels on the left, the solid line represents the collusive equilibrium, where the collusive order flows are set at $\chi_h^C = (\chi_h^N + \chi_h^M)/2$ for $h = 1, 2$. The dash-dotted line represents the deviator, who optimally selects its deviation order flow at $(t, 1)$, conditional on the other informed speculator adheres to the collusive order flow at $(t, 1)$ and that both informed speculators trade non-collusively in the Nash equilibrium at $(t, 2)$ according to (3.18), based on the asset price $p_{t,1}$ realized at $(t, 1)$. The dotted line represents the non-deviator. In the two panels on the right, the solid line represents the collusive equilibrium, the dash-dotted line represents the equilibrium with unilateral deviation, and the dotted line represents the asset's fundamental value. The top four panels illustrate the trading environment with $\mu = 0.3$, while the bottom four panels represent the environment with $\mu = 0.01$. The other parameters are set at $I = 2, r = 0.04, \lambda = 0.2, \xi = 0.33$, and $f_0 = 0$. The realized shocks are $\omega_t = 1$ and $\nu_t = \mathbb{E}_{t,0}[\nu_t] = 0$.

Figure 5: Illustration of the weak punishment issue for sustaining collusion.

investors (i.e., low μ). As shown in panel G of Figure 5, the deviator's total profit is lower than what she would achieve in the collusive equilibrium, due to the punishment at $(t, 2)$. Thus, in this low-μ environment, the non-collusive trading in the Nash equilibrium at $(t, 2)$ is an effective punishment strategy to deter deviation incentives at $(t, 1)$. Intuitively, as discussed in Section 3.3, coordination in the low-μ environment does not require intertemporal planning—the game played at $(t, 1)$ is largely independent of the game played at $(t, 2)$. As a result, deviation from coordination at $(t, 1)$ yields only small gains for the deviator, reflecting the benefits of one-subperiod coordination. This minor gain is easily outweighed by the losses from reverting to the non-collusive Nash equilibrium at $(t, 2)$, making the deviation unprofitable.



Note: In panels A to C, the solid line represents the collusive equilibrium, where the collusive order flows are set at $\chi_h^C = (\chi_h^N + \chi_h^M)/2$ for $h = 1, 2$. The dash-dotted line represents the deviator, who deviates at $(t, 1)$ by the trading same amount as indicated by the dash-dotted line in panel A of Figure 5. The dotted line represents the non-deviator. The order flows of both the deviator and non-deviator at $(t, 2)$ are set to 1.25 times the order flows indicated by the dash-dotted line in panel A of Figure 5, reflecting a punishment that is stronger than non-collusive trading in the Nash equilibrium. In panel D, the solid line represents the collusive equilibrium, the dash-dotted line represents the equilibrium with unilateral deviation, and the dotted line represents the asset's fundamental value. The parameters are set at $\mu = 0.3, I = 2, r = 0.04, \lambda = 0.2, \xi = 0.33$, and $f_0 = 0$. The realized shocks are $\omega_t = 1$ and $\nu_t = \mathbb{E}_{t,0}[\nu_t] = 0$.

Figure 6: Illustration of the impact of implementing a strong punishment at $(t, 2)$.

Next, we illustrate the impact of implementing a strong punishment at $(t, 2)$ in the baseline trading environment with a significant presence of positive-feedback investors. Figure 6 presents the results. For comparison purposes, we consider the same amount of order flow deviation at $(t, 1)$ as indicated by the dash-dotted line in panel A of Figure 5. However, at $(t, 2)$, instead of adopting non-collusive trading in the Nash equilibrium, we consider a stronger punishment by setting the order flows of both the deviator and non-deviator to 1.25 times the order flows indicated by the dash-dotted line in panel A of Figure 5. Panels B and C of Figure 7 show that this strong punishment makes the deviation at $(t, 1)$ unprofitable for the deviator. Although this strong punishment helps deter deviations that occur at $(t, 1)$, the punishment scheme itself is not incentive compatible, as it is inconsistent with the strategies in the Nash equilibrium. Both the deviator and the non-deviator would have the incentive to deviate from the punishment scheme by trading less aggressively. In Section 3.6, we outline a compounded price-trigger strategy that not only effectively punishes any deviation at $(t, 1)$ but is also incentive compatible.

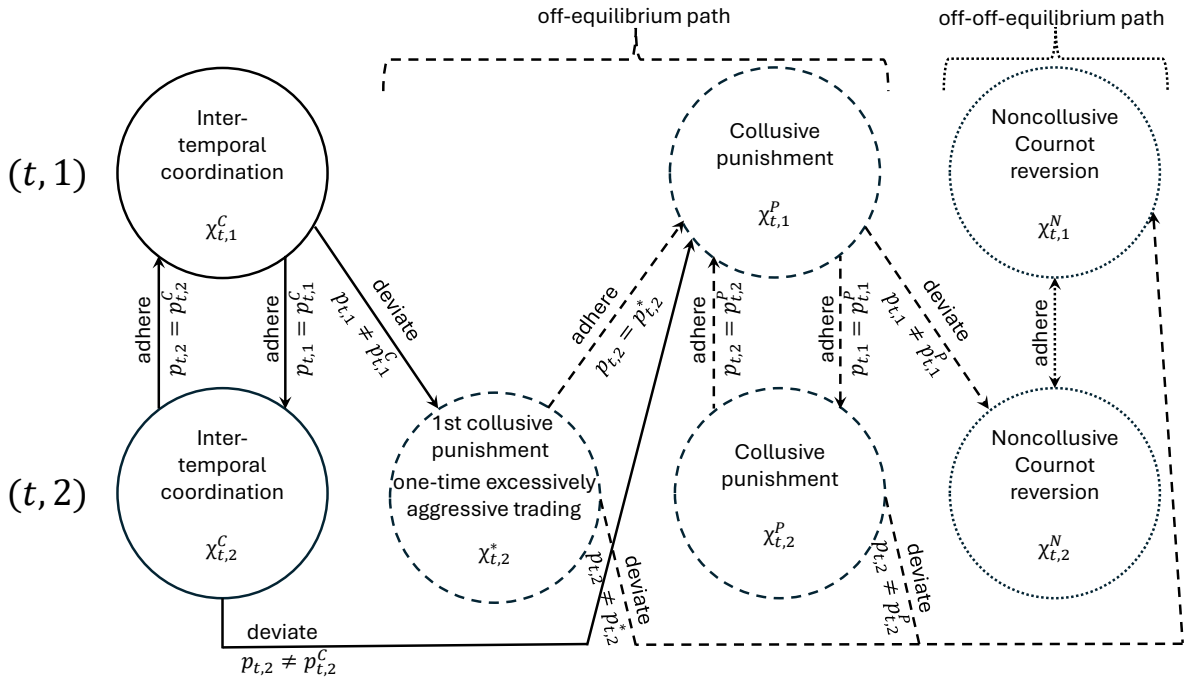
3.6 Sustaining Collusion Using Compounded Price-Trigger Strategy

Unlike standard dynamic games, which typically involve only sequential decision-making across periods, the dynamic game played by informed dynamic speculators in our model encompasses not only sequential decision-making across episodes (i.e., periods) but also sequential decision-making across subperiods within each episode. Importantly, these decisions within the same episode exhibit strong dynamic complementarity. This framework is conceptually similar to the dynamic game described by Hatfield et al. (2020), where firms first collusively submit their bidding prices to a buyer in each period. The winning bidder then offers payments to the other firms and invites them to join a syndicate to complete production. However, the sequential decisions in the model of Hatfield et al. (2020) do not exhibit strong dynamic complementarity.

In a manner similar to Hatfield et al. (2020), we propose a compounded punishment scheme that enables informed speculators to sustain intertemporal coordination, thereby addressing the time inconsistency and weak punishment issues discussed in Sections 3.4 and 3.5. Since informed speculators cannot observe each other's order flows, the compounded punishment scheme is based on observable asset prices. We therefore refer to it as the compounded price-trigger strategy, which enhances the simple, uncompounded the price-trigger strategies (e.g., Green and Porter, 1984; Sannikov and Skrzypacz, 2007; Dou, Goldstein and Ji, 2025). Figure 7 illustrates the compounded price-trigger strategy. There are three phases of equilibrium play: the intertemporal coordination phase, the collusive punishment phase, and the noncollusive Cournot reversion phase. We describe these phases and the transitions across them below.

Equilibrium Path. At $(t,1)$, the equilibrium is in the intertemporal coordination phase if the asset price in the second subperiod of the previous episode satisfies $p_{t-1,2} = p^C_{t-1,2}$. In this phase, every informed speculator must trade according to the coordinated trading strategy $\chi^C_{t,1}$. When all informed speculators adhere to the coordinated trading strategy at $(t,1)$, the asset price remains anchored at $p_{t,1} = p^C_{t,1}$, then the phase will persist as time progresses from $(t,1)$ to $(t,2)$.

At $(t,2)$, the equilibrium is in the intertemporal coordination phase if $p_{t-1,2} = p^C_{t-1,2}$ and $p_{t,1} = p^C_{t,1}$. In this phase, every informed speculator must trade according to the coordinated trading strategy $\chi^C_{t,2}$. When all informed speculators adhere to the coordinated trading strategy at $(t,2)$, the asset price remains anchored at $p_{t,2} = p^C_{t,2}$, then the phase will persist as time progresses from $(t,2)$ to $(t+1,1)$.



Note: The top row represents the states at $(t,1)$, while the bottom row represents the states at $(t,2)$. The solid circles indicate the intertemporal coordination phase, with solid arrows representing transitions originating from this phase. The dashed circles represent the collusive punishment phase, which includes a one-time excessively aggressive trading, while the dashed arrows indicate transitions originating from this phase. The dotted circles represent the non-collusive Cournot reversion phase, with the dotted arrows denoting transitions originating from this phase.

Figure 7: Graphic illustration of the compounded price-trigger strategy.

Off-equilibrium Path. The equilibrium can transition from the intertemporal coordination phase to the collusive punishment phase either in the first subperiod or the second subperiod.

The transition occurs in the first subperiod $(t,1)$ of episode t if at least one informed speculator deviates from the coordinated trading strategy $\chi^C_{t-1,2}$ at $(t-1,2)$ in the previous episode $t-1$. Then, at $(t,1)$, every informed speculator will notice that the realized asset price in the second subperiod of the previous episode is $p_{t-1,2} \neq p^C_{t-1,2}$. This discrepancy triggers a transition to the collusive punishment phase at $(t,1)$, where all informed speculators must trade according to the coordinated punishment strategy $\chi^P_{t,1}$, resulting in the asset price $p^P_{t,1}$. If all informed speculators adhere to $\chi^P_{t,1}$ at $(t,1)$, the asset price remains anchored at $p_{t,1} = p^P_{t,1}$, then the phase will persist as time progresses from $(t,1)$ to $(t,2)$, where every informed speculator must trade according to $\chi^P_{t,2}$.

The transition occurs in the second subperiod $(t,2)$ of episode t if at least one informed

speculator deviates from the coordinated trading strategy $\chi^C_{t,1}$ at $(t,1)$. When this happens, at $(t,2)$, every informed speculator observes that the realized asset price from the previous subperiod within episode t is $p_{t,1} \neq p^C_{t,1}$, even though the equilibrium is in the intertemporal coordination phase in the previous episode $t-1$, i.e., $p_{t-1,2} = p^C_{t-1,2}$. This observation leads to a transition to the collusive punishment phase at $(t,2)$. To address the weak punishment issue, every informed speculator engages in a one-time excessively aggressive trading strategy at $(t,2)$, choosing a strategy $\chi^*_{t,2} \equiv \chi^*_2(p_{t,1};\omega_t, f_{t,1}) = b\chi^N_2(p_{t,1};\omega_t, f_{t,1})$, with $b > 1$.[11] This strategy is more aggressive than the non-collusive trading strategy $\chi^N_2(p_{t,1};\omega_t, f_{t,1})$ in the Nash equilibrium at $(t,2)$ (see equation (3.18)), conditional on the realized asset price $p_{t,1}$.[12] If all informed speculators adhere to $\chi^*_{t,2}$, the realized asset price will be $p^*_{t,2}$, allowing the equilibrium to remain in the collusive punishment phase as time progresses from $(t,2)$ to $(t+1,1)$.

Note that the trading strategy $\chi^P_{t+1,h}$ in the collusive punishment phase lies between $\chi^C_{t+1,h}$ and $\chi^N_{t+1,h}$ for $h = 1,2$, ensuring that it generates more profit than the non-collusive Nash equilibrium but less profit than what is achieved in the intertemporal coordination phase. This condition implies that the trading strategy $\chi^P_{t+1,h}$ in the collusive punishment phase is easier to maintain as an equilibrium outcome compared to the strategy $\chi^C_{t+1,h}$ in the intertemporal coordination phase.

Off-off-equilibrium Path. Finally, we discuss the transition from the collusive punishment phase and the non-collusive Cournot reversion phase.

The transition occurs in the second subperiod $(t,2)$ of episode t if it is observed in the first subperiod, $(t,1)$, that $p_{t,1} \neq p^P_{t,1}$. This indicates that at least one informed speculator has deviated from the coordinated punishment strategy $\chi^P_{t,1}$. Upon entering the non-collusive Cournot reversion phase at $(t,2)$, all informed speculators adopt the non-collusive trading strategy $\chi^N_2(p_{t,1};\omega_t, f_{t,1})$ according to (3.18), based on the realized asset price $p_{t,1}$, with no incentive to deviate. From $t+1$ onward, all informed speculators continue to trade according to the non-collusive Nash equilibrium strategy, $\chi^N_{t,h}$ for

[11] At $(t,2)$, the lagged state variable $p_{t-1,2} = p^C_{t-1,2}$ indicates that it is the first time the equilibrium has transitioned to the collusive punishment phase, allowing informed speculators to implement the one-time excessively aggressive trading strategy at $(t,2)$ to address the weak punishment issue.

[12] The strategy $\chi^N_{t,2} \equiv \chi^N_2(p_{t,1};\omega_t, f_{t,1})$, given by (3.18), nests $\chi^N_2 \equiv \chi^N_2(p^N_{t,1};\omega_t, f_{t,1})$, given by (3.21), as a special case, which represents the non-collusive Nash equilibrium strategy conditional on the realized asset price being $p^N_{t,1}$. The price $p^N_{t,1}$ is the price that prevails in the non-collusive Nash equilibrium, where all informed speculators engage in non-collusive Nash equilibrium strategies at both $(t,1)$ and $(t,2)$, as characterized in Section 3.2.

$h = 1, 2$, also with no incentive to deviate.

Regarding the transition that occurs in the first subperiod $(t, 1)$ of episode t, there are two scenarios depending on whether the transition occurs while informed speculators are following the one-time excessively aggressive trading strategy or not. First, at $(t-1, 2)$, if the state variable is $p_{t-2,2} = p^C_{t-2,2}$ and it is observed that $p_{t-1,2} \neq p^*_{t-1,2}$, this indicates that at least one informed speculator has deviated from the coordinated one-time excessively aggressive trading strategy $\chi^*_{t-1,2}$, resulting in a transition to the non-collusive Cournot reversion phase. Second, at $(t-1, 2)$, if the state variable is $p_{t-2,2} \neq p^C_{t-2,2}$ and it is observed that $p_{t-1,2} \neq p^P_{t-1,2}$, this indicates that at least one informed speculator has deviated from the coordinated punishment strategy $\chi^P_{t-1,2}$, also leading to a transition to the non-collusive Cournot reversion phase. Regardless of the scenario that led to the transition, from episode t onward, all informed speculators continue to trade according to the non-collusive Nash equilibrium strategy, $\chi^N_{t,h}$ for $h = 1, 2$, with no incentive to deviate.

Discussions and Numerical Examples. By properly designing the parameters of the compounded price-trigger strategy described above, we can establish it as a sub-game perfect equilibrium, effectively addressing the time inconsistency and the weak punishment issues. Clearly, the non-collusive Cournot reversion phase itself constitutes a sub-game perfect equilibrium, as the trading strategies employed in this phase are based on the non-collusive Nash equilibrium strategies. Below, we explain why both the intertemporal coordination phase and the collusive punishment phase constitute sub-game perfect equilibria.

First, the collusive punishment phase helps maintain the intertemporal coordination among informed speculators, making deviations from $\chi^C_{t,h}$ unprofitable for $h = 1, 2$. Specifically, regarding any deviation occurring at $(t, 1)$, introducing a one-time excessively aggressive trading strategy at $(t, 2)$ in the collusive punishment phase can lead to significantly greater discounted losses in the future compared to a punishment that follows the non-collusive Nash equilibrium strategy from $(t, 2)$ onward. This approach helps mitigate the weak punishment issue associated with deviations from the intertemporal coordination that occur at $(t, 1)$. Moreover, the time inconsistency issue is addressed by implementing collusive punishment from $(t+1, 1)$ onward, which serves to deter any deviations from the intertemporal coordination that occur at $(t, 2)$.

Second, the non-collusive Cournot reversion phase ensures the credibility of collusive punishment. Specifically, in the collusive punishment phase, informed speculators may be

Table 1: Numerical example of compounded price-trigger strategy.

Panel A: Uncompounded price-trigger strategy

	current	future	net
$(t,1)$	35.4	-30.6	4.8
$(t,2)$	26.0	-30.6	-4.6

Panel B: Compounded price-trigger strategy

	Coordination phase			Collusive punishment phase					
				One-time excess. punish			normal punish		
	current	future	net	current	future	net	current	future	net
$(t,1)$	24.6	-27.4	-2.8	n.a.	n.a.	n.a.	-0.4	-3.2	-3.6
$(t,2)$	26.0	-27.4	-1.4	2.4	-3.2	-0.8	0.2	-3.2	-3.0

Note: Parameters are set at $I = 2, \mu = 0.32, r = 0.04, \lambda = 0.2, \xi = 0.33$, $\sigma_\omega = \sqrt{\pi/2}$, and $\sigma_\nu = 0$. For illustration purposes, we set $\rho = 0.65$ and focus on analyzing the incentive of deviation in episode t, where the realized fundamental shock is $\omega_t = \mathbb{E}[\omega_t|\omega_t > 0] = \sigma_\omega\sqrt{2/\pi} = 1$. The trading strategy in the intertemporal coordination phase are set at $\chi_h^C = (1-\epsilon^C)\chi_h^N + \epsilon^C\chi_h^M$ for $h = 1,2$, with $\epsilon^C = 0.5$. The trading strategy in the collusive punishment phase are set at $\chi_h^P = (1-\epsilon^P)\chi_h^N + \epsilon^P\chi_h^M$ for $h = 1,2$, with $\epsilon^P = 0.04$. The one-time excessively aggressive punishment is set at $\chi_2^*(p_{t,1};\omega_t, f_{t,1}) = b\chi_2^N(p_{t,1};\omega_t, f_{t,1})$, where $b = 1.15$ and $\chi_2^N(p_{t,1};\omega_t, f_{t,1})$ is given by (3.18).

tempted to deviate from the one-time excessively aggressive trading strategy $\chi_{t,2}^*$ at $(t,2)$ due to the immediate significant losses associated with such a strategy. However, this temptation is countered by the transition to the non-collusive Cournot reversion phase, which results in strictly lower profits in the future compared to the profits available in the collusive punishment phase. Furthermore, once the one-time excessively aggressive trading strategy is completed, the non-collusive Cournot reversion phase helps maintain the collusive punishment phase by deterring any deviations from $\chi_{t,1}^P$ and $\chi_{t,2}^P$ at $(t,1)$ and $(t,2)$, respectively. It is important to note that sustaining the collusive punishment phase is easier than maintaining the intertemporal coordination phase because informed speculators earn lower profits in the former. This clarifies why the non-collusive Cournot reversion phase can effectively support the collusive punishment phase, while it may not be as effective in sustaining the intertemporal coordination phase. Therefore, the compounded price-trigger strategy, which incorporates the collusive punishment phase, is essential for maintaining the intertemporal coordination phase.

In Table 1, we present a numerical example to illustrate how the compounded price-trigger strategy may outperform the simple, uncompounded price-trigger strategy in

maintaining intertemporal coordination among informed speculators. Panel A of Table 1 presents the current gains and future losses associated with optimal deviations when the price-trigger strategy is uncompounded. Specifically, for any deviation occurring at $(t,1)$ or $(t,2)$, we employ the non-collusive Nash equilibrium as the punish mechanism. This means that punishment for deviations in the intertemporal coordination phase involves a direct transition to the non-collusive Cournot reversion phase, bypassing the collusive punishment phase illustrated in Figure 7. To have a stark comparison, we intentionally set the model parameters so that the collusive equilibrium cannot be sustained by this uncompounded price-trigger strategy. In particular, panel A of Figure 1 demonstrates that deviating at $(t,1)$ yields an immediate contemporaneous gain of 35.4 in the current episode, which exceeds the discounted loss of 30.6 in future episodes. This is exactly due to the weak punishment issue highlighted in Section 3.5. In contrast, the time-inconsistency issue is less severe: deviating at $(t,2)$ results in an immediate contemporaneous gain of 26.0, which is outweighed by the discounted future loss of 30.6. Clearly, intertemporal coordination cannot be sustained in this scenario due to the weak punishment issue, as there is a positive net gain of 4.8 from deviation at $(t,1)$.

Panel B of Table 1 presents the current gains and future losses associated with optimal deviations under the compounded price-trigger strategy. A comparison of the numbers in the coordination phase of panel B with those in panel A reveals that the current gain from deviating at $(t,1)$ is 24.6 in panel B, which is significantly lower than the gain of 35.4 in panel A. Consequently, the net gain from deviating at $(t,1)$ is negative, amounting to -2.8, which eliminates any incentive for informed speculators to deviate.

By introducing a one-time excessively aggressive trading strategy, the compounded punishment effectively addresses the weak punishment issue, thus removing the deviation incentive at $(t,1)$. To sustain this one-time excessively aggressive trading as an incentive-compatible strategy, future punishment must be collusive, which is less severe than the punishment based on non-collusive Nash equilibrium strategies. As a result, under the compounded price-trigger strategy, future losses due to punishment are lower than those under the uncompounded price-trigger strategy. Consequently, the net gain from deviating at $(t,2)$ is higher than that in panel A, even though it remains negative.[13] Furthermore, panel B of Table 1 confirms that the collusive punishment phase is credible,

[13]We can easily make the net gain from deviating at $(t,2)$ the same in both panels A and B of Table 1 by specifying that any deviation at $(t,2)$ in the intertemporal coordination phase directly transitions to the non-collusive Cournot reversion phase, bypassing the collusive punishment phase.

constituting a sub-game perfect Nash equilibrium.

4 Design of Simulation Experiments

In this section, we design simulation experiments to test whether informed AI speculators, equipped with model-free coupled Q-learning algorithms, can achieve and sustain coordinated intertemporal planning, thereby influencing financial market stability.

4.1 Algorithms as Experimental Subjects

Informed AI Speculators. We now analyze the behavior of the algorithms as experimental subjects. Specifically, these experiments replace the theoretical agents, referred to as "informed speculators" in the model, as detailed in Section 3, with coupled Q-learning algorithms, as described in Section 2.2.

Our specification of the algorithms' planning window and state variables closely follow the model. According to our model, informed speculators make trading decisions at $h = 1, 2$. They possess an information advantage over other investors by observing the fundamental shock ω_t before submitting the order flow $x_{i,t,1}$ at $(t, 1)$ in episode t. Moreover, the asset price $p_{t,1}$ at $(t, 1)$ influences the asset demand of positive-feedback investors at $(t, 2)$. This price is observable to informed speculators at $(t, 2)$ and affects their choice of order flow $x_{i,t,2}$ at that point. To incorporate these elements into our coupled Q-learning equation system (2.4), we set $H = 2$ and include ω_t in the state vectors $s_{i,t,1}$ and $s_{i,t,2}$, while $p_{t,1}$ is included in the state vector $s_{i,t,2}$. The payoffs are $\pi_{i,t,1} = (p_{t,3} - p_{t,1})x_{i,t,1}$ and $\pi_{i,t,2} = (p_{t,3} - p_{t,2})x_{i,t,2}$, realized at the end of the episode t.

Furthermore, we include lagged information in the state vectors $s_{i,t,1}$ and $s_{i,t,2}$ to provide informed AI speculators the opportunity to achieve and sustain coordination in their intertemporal trading decisions. As elaborated in Section 3.6, intertemporal coordination can be established at both $(t, 1)$ and $(t, 2)$ through compounded price-trigger strategies. Thus, to achieve and maintain collusion at both $(t, 1)$ and $(t, 2)$, it is necessary to incorporate some lagged states that can record deviations from steady-state (coordinating) behavior for both subperiods. In line with our theory, we include $p_{t-1,2}$ in the state vectors $s_{i,t,1}$ and $s_{i,t,2}$, along with the lagged fundamental shock ω_{t-1}. This theoretically enables informed AI speculators to assess potential deviations at both $(t-1, 1)$ and $(t-1, 2)$ in the previous episode $t-1$. Specifically, by tracking ω_{t-1},

informed AI speculators have the potential to learn about the expected steady-state asset price $\mathbb{E}[p_{t-1,2}|\omega_{t-1}]$, corresponding to ω_{t-1}. Thus, by additionally tracking $p_{t-1,2}$, these informed AI speculators can evaluate potential deviations in the asset price at $(t-1,2)$ from its expected steady-state value in the previous episode $t-1$. These deviations may indicate either an order-flow deviation in $x_{i,t-1,1}$ at $(t-1,1)$ or in $x_{i,t-1,2}$ at $(t-1,2)$. It is important to note that tracking $p_{t-1,1}$ instead of $p_{t-1,2}$ does not serve the same purpose, as deviations in $p_{t-1,1}$ reflect only the deviations in $x_{i,t-1,1}$, without providing any information for $x_{i,t-1,2}$.

Per our discussions above, we specify the state vectors as $s_{t,1} = \{\omega_t, \omega_{t-1}, p_{t-1,2}\}$ and $s_{t,2} = \{\omega_t, p_{t,1}, \omega_{t-1}, p_{t-1,2}\}$. For the remainder of this paper, we omit the subscript i for these state vectors as speculator-specific information is not included. Our choice of state vectors constitutes a minimal set of state variables that could theoretically enable informed AI speculators to achieve and maintain intertemporal coordination. This choice offers two key advantages: First, it establishes a high standard for AI planning of intertemporal coordination, as expanding the set of state variables to include more lagged information (such as $p_{t-1,1}$ or extended histories of asset prices and fundamental values) will enhance the learning of collusive behavior among informed AI speculators. Second, this choice enhances the tractability and transparency of our numerical procedure. High-dimensional state spaces not only introduce computational challenges—often requiring deep learning techniques for function approximation and effective exploration—but also complicate the explanation and visualization of results.

4.2 Protocol for Simulation-Based Experiments

We summarize the experimental protocol as follows. At $t = 0$, each informed AI speculator $i \in \{1, \ldots, I\}$ is assigned with two arbitrary initial Q-matrices, $\widehat{Q}_{i,0,1}$ and $\widehat{Q}_{i,0,2}$, and arbitrary states, $s_{0,1}$ and $s_{0,2}$. The economy then evolves from episode t to $t+1$ according to the following steps:

(1) At $(t,0)$, low-frequency, risk-neutral investors trade, resulting in the asset price $p_{t,0}$ being perfectly aligned with its fundamental value $f_{t,0}$.

(2) At $(t,1)$, the asset's fundamental value remains unchanged, $f_{t,1} = f_{t,0}$. Each informed AI speculator i receives a perfect signal about ω_t, so its state vector is $s_{t,1} = \{\omega_t, \omega_{t-1}, p_{t-1,2}\}$. Each informed AI speculator i then independently enters

exploration with probability ε_t or exploitation with probability $1-\varepsilon_t$, submitting order flow $x_{i,t,1}$ as in (2.6). Constrained arbitrageurs submit their aggregate order flow $z_{t,1}$ according to (3.11). The equilibrium asset price $p_{t,1}$ is determined by the market-clearing condition according to (3.13).

(3) At $(t,2)$, the shock ω_t is realized and observed by all investors. The asset's fundamental value becomes $f_{t,2} = f_{t,0} + \omega_{f,t}$. Each informed AI speculator i's state vector is $s_{t,2} = \{\omega_t, p_{t,1}, \omega_{t-1}, p_{t-1,2}\}$, and it independently enters exploration with probability ε_t or exploitation with probability $1-\varepsilon_t$, submitting order flow $x_{i,t,2}$ as in (2.6). Constrained arbitrageurs submit their aggregate order flow $z_{t,2}$ according to (3.12). Positive-feedback investors submit their aggregate order flow $y_{t,2}$ as in (3.10). The equilibrium asset price $p_{t,2}$ is determined by the market-clearing condition according to (3.14).

(4) At $(t,3)$, the shock ν_t is realized, independently drawn from $N(0,\sigma_\nu^2)$, and observed by all investors. The asset's fundamental value becomes $f_{t,3} = f_{t,0} + \omega_{f,t} + \nu_{f,t}$. Low-frequency, risk-neutral investors trade, resulting in the asset price $p_{t,3}$ being perfectly aligned with its fundamental value $f_{t,3}$. Each informed AI speculator i realizes its profits $\pi_{i,t,1} = (p_{t,3} - p_{t,1})x_{i,t,1}$ and $\pi_{i,t,2} = (p_{t,3} - p_{t,2})x_{i,t,2}$ for its order flows placed at $(t,1)$ and $(t,2)$, respectively.

(5) At $(t+1,0)$, the beginning of episode $t+1$, the fundamental value becomes $f_{t+1,0} = f_{t,3} + \lambda(\overline{f} - f_{t,3})$, and the asset price becomes $p_{t+1,0} = f_{t+1,0}$ due to trading by low-frequency, risk-neutral investors. Then, at $(t+1,1)$, the state vector transitions from $s_{t,2} = \{\omega_t, p_{t,1}, \omega_{t-1}, p_{t-1,2}\}$ to $s_{t+1,1} = \{\omega_{t+1}, \omega_t, p_{t,2}\}$, where ω_{t+1} is independently drawn from $N(0,\sigma_\omega^2)$. Each informed AI speculator i updates its Q-value $\widehat{Q}_{t,1}$ for $(s_{t,1}, x_{i,t,1})$ and $\widehat{Q}_{t,2}$ for $(s_{t,2}, x_{i,t,2})$ using (2.4).

4.3 Numerical Specifications

We present the numerical setup of our simulations, including the discretization of state and action spaces, Q-matrix initialization, parameter selection, and convergence criteria.

Specification of Exploration Rates. The exploration probability ε_t is specified as follows:

$$\varepsilon_t = e^{-\beta t}.$$

The informed AI speculators' exploration rate ε_t begins at 1 (indicating pure exploration) and decreases over time. The parameter $\beta > 0$ governs the rate of this decline.

Discretization of State and Action Spaces. We approximate the distribution $N(0, \sigma_\omega)$ using n_ω grid points, $\Omega = \{\omega_1, \cdots, \omega_{n_\omega}\}$, with equal probabilities assigned to each grid point. The grid points are located according to $\omega_k = 0 + \sigma_\omega \Phi^{-1}((2k-1)/(2n_\omega))$ for $k = 1, \cdots, n_\omega$, where Φ^{-1} is the inverse cumulative density function of the standard normal distribution.[14] We discretize the choice space of informed AI speculator i for order flows, $x_{i,1}$ and $x_{i,2}$, using grids based on the optimal trading strategies in the non-collusive Nash equilibrium and the perfect cartel benchmark. Specifically, in the first subperiod, the order flows in these two benchmarks are given by $x_1^N = \frac{\xi(I+1)^2\zeta\omega}{(I+1)^3 - 2I(1+\mu/\xi)^2}$ and $x_1^M = \frac{2\xi\zeta\omega}{4I - I(1+\mu/\xi)^2}$. We discretize the interval $[\min\{x_1^N, x_1^M\} - \iota|x_1^N - x_1^M|, \max\{x_1^N, x_1^M\} + \iota|x_1^N - x_1^M|]$ into n_{x_1} equally spaced grid points, denoted by $\mathbb{X}_1 = \{x_{1,1}, \ldots, x_{1,n_{x_1}}\}$. The parameter $\iota > 0$ enables informed AI speculators to choose order flows that exceed the boundaries set by the theoretical benchmarks x_1^N and x_1^M, offering flexibility to explore strategies beyond these theoretical limits. In the second subperiod, the order flows in the two benchmarks are $x_2^N = -\frac{(\xi+\mu)I(I+1)\zeta\omega}{(I+1)^3 - 2I(1+\mu/\xi)^2}$ and $x_2^M = -\frac{(\xi+\mu)\zeta\omega}{4I - I(1+\mu/\xi)^2}$. We discretize the interval $[\min\{x_2^N, x_2^M\} - \iota|x_2^N - x_2^M|, \max\{x_2^N, x_2^M\} + \iota|x_2^N - x_2^M|]$ into n_{x_2} equally spaced grid points, denoted by $\mathbb{X}_2 = \{x_{2,1}, \ldots, x_{2,n_{x_2}}\}$.

The grid points of the asset prices, p_1 and p_2, are determined similarly to those for $x_{i,1}$ and $x_{i,2}$. Specifically, the lower bound of p_1 is set at $p_{1,L} = f_1 + \min\{x_1^N, x_1^M\} I/\xi$ and the upper bound at $p_{1,H} = f_1 + \max\{x_1^N, x_1^M\} I/\xi$. The interval $[p_{1,L} - \iota(p_{1,H} - p_{1,L}), p_{1,H} + \iota(p_{1,H} - p_{1,L})]$ is discretized into n_{p_1} grid points, denoted by $\mathbb{P}_1 = \{p_{1,1}, \ldots, p_{1,n_{p_1}}\}$. The lower bound of p_2 is set at $p_{2,L} = f_2 + \mu/\xi(p_{1,L} - f_0) + (p_{1,L} - f_1) + \min\{x_2^N, x_2^M\} I/\xi$ and the upper bound at $p_{2,H} = f_2 + \mu/\xi(p_{1,H} - f_0) + (p_{1,H} - f_1) + \max\{x_2^N, x_2^M\} I/\xi$. The interval $[p_{2,L} - \iota(p_{2,H} - p_{2,L}), p_{2,H} + \iota(p_{2,H} - p_{2,L})]$ is discretized into n_{p_2} grid points, denoted by $\mathbb{P}_2 = \{p_{2,1}, \ldots, p_{2,n_{p_2}}\}$.

Initial Q-Matrix and States. For each informed AI speculator $i = 1, \ldots, I$, we initialize its initial Q-matrix $\widehat{Q}_{i,0,1}(s_1, x_1)$ at $(0, 1)$ in the initial episode $t = 0$ with the discounted payoff that informed AI speculator i would earn by choosing x_1 in s_1. This is conditional on the speculator itself randomizing its actions uniformly over the grid points $\mathbb{X}_2$ at $(t, 2)$, while other informed AI speculators randomize their actions uniformly over the

[14] The results remain robust under alternative discretization schemes.

grid points $\mathbb{X}_1$ at $(t,1)$ and $\mathbb{X}_2$ at $(t,2)$. The shock ν_t is set to zero, corresponding to the expected value of the distribution $N(0,\sigma_\nu^2)$. Specifically, we set its value as follows:

$$\widehat{Q}_{i,0,1}(s_1,x_1)=\frac{1}{(1-\rho)n_{x_1}n_{x_2}^2}\sum_{x_2\in\mathbb{X}_2}\sum_{x_{-i,1}\in\mathbb{X}_1}\sum_{x_{-i,2}\in\mathbb{X}_2}\left[(p_3-p_1)x_1+(p_3-p_2)x_2\right],$$

for $s_1=(\omega,\omega,p_2)\in\Omega\times\Omega\times\mathbb{P}_2$, where $p_1=f_1+1/\xi(x_1+(I-1)x_{-i,1})$ and $p_2=f_2+\mu/\xi(p_1-f_0)+(p_1-f_1)+1/\xi(x_2+(I-1)x_{-i,2})$.

Similarly, we set each informed AI speculator i's initial Q-matrix $\widehat{Q}_{i,0,2}$ at $(0,2)$ as follows:

$$\widehat{Q}_{i,0,2}(s_2,x_2)=\frac{1}{(1-\rho)n_{x_2}}\sum_{x_{-i,2}\in\mathbb{X}_2}(p_3-p_2)x_2+\frac{\rho}{(1-\rho)n_{x_1}}\sum_{x_1\in\mathbb{X}_1}(p_3-p_1)x_1,$$

for $s_2=(\omega,p_1,\omega,p_2)\in\Omega\times\mathbb{P}_1\times\Omega\times\mathbb{P}_2$, where $p_2=f_2+\mu/\xi(p_1-f_0)+(p_1-f_1)+1/\xi(x_2+(I-1)x_{-i,2})$. Note that when computing $\widehat{Q}_{i,0,2}(s_2,x_2)$, p_1 is known as it is one of the state variables in the vector s_2.

Our simulation results are robust to initialization of the two Q-matrices, given that our choice of $\beta=10^{-9}$ is very close to zero. Intuitively, the algorithms will always begin with the exploration-intensive phase, meaning they randomize their actions uniformly over the grid points $\mathbb{X}_1$ at $(t,1)$ and $\mathbb{X}_2$ at $(t,2)$. Since β is very close to zero, this exploration-intensive phase lasts for a significant number of episodes. Thus, starting from arbitrary initial Q-values, the two Q-matrices will converge to roughly the same values after algorithms complete the exploration-intensive phase, making the initial conditions irrelevant. Our specification of initial Q-matrices, based on the assumption that all other informed AI speculators randomize their actions, aims to approximate the Q-matrices to which the algorithms will converge upon completing the exploration-intensive phase. This approach helps speed up convergence without affecting the simulation results.

Parameter Values. In the baseline calibration, we set the parameters that describe the underlying economic environment as follows. These parameter values are unknown to informed AI speculators. We consider two informed AI speculators by setting $I=2$. The volatility of the asset's fundamental value is normalized to be 1, $\sigma_w=1$, and we set $\sigma_\nu=0.03$. The mean-reversion parameter of the asset's fundamental value is set at $\lambda=0.2$. The slope parameter for constrained arbitrageurs is set to $\xi=0.33$. The slope parameter of positive-feedback investors must satisfy $\mu<\xi$ to ensure that informed speculators

obtain finite profits in the perfect cartel benchmark. For the baseline experiments, we set $\mu = 0.3$, representing trading environments with a significant presence of positive-feedback investors. Later, we examine the implications of varying this parameter. We normalize the initial fundamental value to $f_0 = 0$. This choice does not impact any results, as the profits of informed speculators arise from fluctuations in the asset's fundamental value, rather than its absolute level.

The risk-free rate is set at $r = 0.04$, which implies that the subjective discount factor for informed AI speculators is $\rho = 1/(1+r)$. This parameter is known to informed AI speculators, as it directly enters the Q-learning algorithm. We discretize the state and decision space by setting $\iota = 0.1$, $n_{x_1} = 31, n_{x_2} = 61, n_{p_1} = 61$, and $n_{p_2} = 121$. We set the hyperparameters that govern the learning process of Q-learning algorithms at $\alpha = 0.05$ and $\beta = 10^{-9}$. All results are robust to variations in α and β, as long as they fall within a reasonable range that ensures sufficiently good learning outcomes. Specifically, α must be sufficiently small to facilitate effective learning, while β needs to be sufficiently close to zero to ensure adequate exploration. Our baseline choice of $\beta = 10^{-9}$ implies that any action $x_1 \in \mathbb{X}_1$ (or $x_2 \in \mathbb{X}_2$) is, on average, visited more than 10 million times due to random exploration before the exploration phase concludes.

Criterion for Numerical Convergence. Each experiment contains $N_{sim} = 1,000$ independent simulation sessions. We adopt a strict criterion for convergence: all informed AI speculators' optimal strategies must remain unchanged for $T_{conv} = 1,000,000,000$ consecutive periods within a session, and all N_{sim} sessions must continue until this condition is met. The value of T_{conv} specified in our convergence criterion is $1,000$ times greater than that used by Dou, Goldstein and Ji (2025) because the coupled Q-learning here is more complex. A longer exploration period is needed (i.e., a β much closer to zero) to thoroughly explore all possible intertemporal trading strategies, as changes in the Q-value of one Q-matrix can affect the other, potentially leading to different optimal actions, especially in the presence of strong complementarities between trading decisions in the first and second subperiods. By enforcing a strict convergence criterion and setting β close to zero, we ensure that informed AI speculators genuinely learn stable strategies, rather than those still converging or linked to inadequate exploration. We confirm that further decreasing β or increasing T_{conv} does not alter the results. The number of periods needed for convergence varies substantially across experiments, depending on parameter values, and can also differ significantly across sessions within the same experiment.

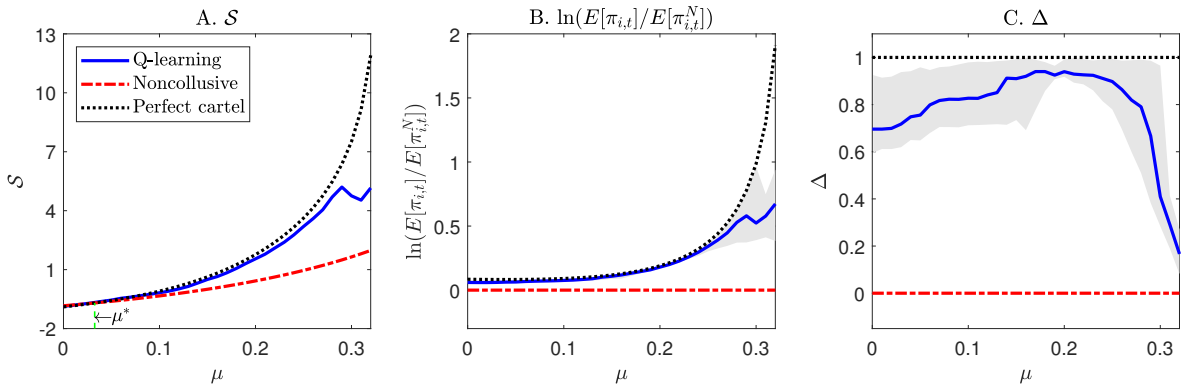
Convergence across simulations occurs within approximately 10 billion to 200 billion simulation episodes. Our programs are written in C++ and run on a high-performance computing cluster with 4 AMD EPYC 9754 processors, totaling 512 CPU cores and 3TB of DDR5-4800 ECC memory. We use the latest AMD AOCC compiler, based on Clang/LLVM, to optimize both the compilation process and code performance. Completing all simulation sessions for a single experiment can take about 3 to 48 hours, depending on the specific parameter configurations.

5 Simulation Results

In this section, we present the results of simulation experiments that analyze the behavior of trading algorithms with AI planning. In Section 5.1, we investigate the effects of AI planning on financial market stability across various trading environments, specifically focusing on the role of positive-feedback investors, represented by the value of μ. Section 5.2 examines the baseline trading environment characterized by a significant presence of positive-feedback investors (high μ). In contrast, Section 5.3 explores trading environments with a limited presence of positive-feedback investors (low μ). Finally, in Section 5.4, we study the implications for financial market stability when informed AI speculators adopt reactive Q-learning algorithms.

5.1 Impacts of AI Planning on Financial Market Stability

We examine trading environments that differ in the presence of positive-feedback investors. For each trading environment, we conduct $N_{sim} = 1,000$ independent simulation sessions until all sessions converge. Using the time series data of $T = 100,000$ episodes generated by each simulation session, we calculate the average instability $\mathcal{S}$, log profit ratio $\ln(\mathbb{E}[\pi_{i,t}]/\mathbb{E}[\pi_{i,t}^N])$, and the collusion capacity Δ according to (3.27) and (3.29). Panel A of Figure 8 plots the average instability measure $\mathcal{S}$ as μ varies from 0 to 0.3 along the x-axis. The solid line represents the instability measure in the simulated financial market with informed AI speculators, which consistently falls between the two theoretical benchmarks: the non-collusive Nash equilibrium (the dash-dotted line) and the perfect cartel benchmark (the dotted line). All three lines slope upward, indicating that financial market instability increases as the influence of positive-feedback investors on asset prices becomes more pronounced. The vertical dashed line, labeled $\mu^* \approx 0.03$, is located at the



Note: The average collusion capacity Δ and stability $\mathcal{S}$ over $N_{sim} = 1,000$ independent simulation sessions are computed according to (3.27) and (3.29). We set the range of the horizontal axis from $\mu = 0$ to $\mu = 0.32$, as the profit of informed speculators in the perfect cartel benchmark becomes unbounded when $\mu \geq \xi = 0.33$.

Figure 8: Impacts of AI planning on financial market stability.

intersection of the dash-dotted and dotted lines. It represents the threshold above which collusion leads to increased financial instability relative to the theoretical benchmark non-collusive Nash equilibrium. For $\mu < \mu^*$, the intertemporal coordination among informed speculators mildly stabilizes financial markets.

Panel B of Figure 8 plots the average log profit ratio $\ln(\mathbb{E}[\pi_{i,t}]/\mathbb{E}[\pi_{i,t}^N])$. The grey area surrounding the solid line represents the range of $\ln(\mathbb{E}[\pi_{i,t}]/\mathbb{E}[\pi_{i,t}^N])$ from the 1st to the 99th percentile across all N_{sim} simulation sessions. As μ increases, informed AI speculators gain more profits relative to what they would be able to achieve in the non-collusive Nash equilibrium. Panel C of Figure 8 plots the average collusion capacity Δ among informed AI speculators, which displays an inverted U-shape. Specifically, as μ becomes very high, the value of Δ decreases.

5.2 Simulation Experiments in the Baseline Trading Environment

We focus on the baseline trading environment with a significant presence of positive-feedback investors, specifically $\mu = 0.3$, and characterize the trading strategies learned by informed AI speculators after their coupled Q-learning algorithms have fully converged. In the steady state with AI planning, our simulation results indicate significant intertemporal coordination among trading algorithms. Across $N_{sim} = 1,000$ simulation sessions, the average collusion capacity is $\Delta = 0.41$. The coordinated AI planning in trading substantially increases financial market instability, as evidenced by the average ratio of

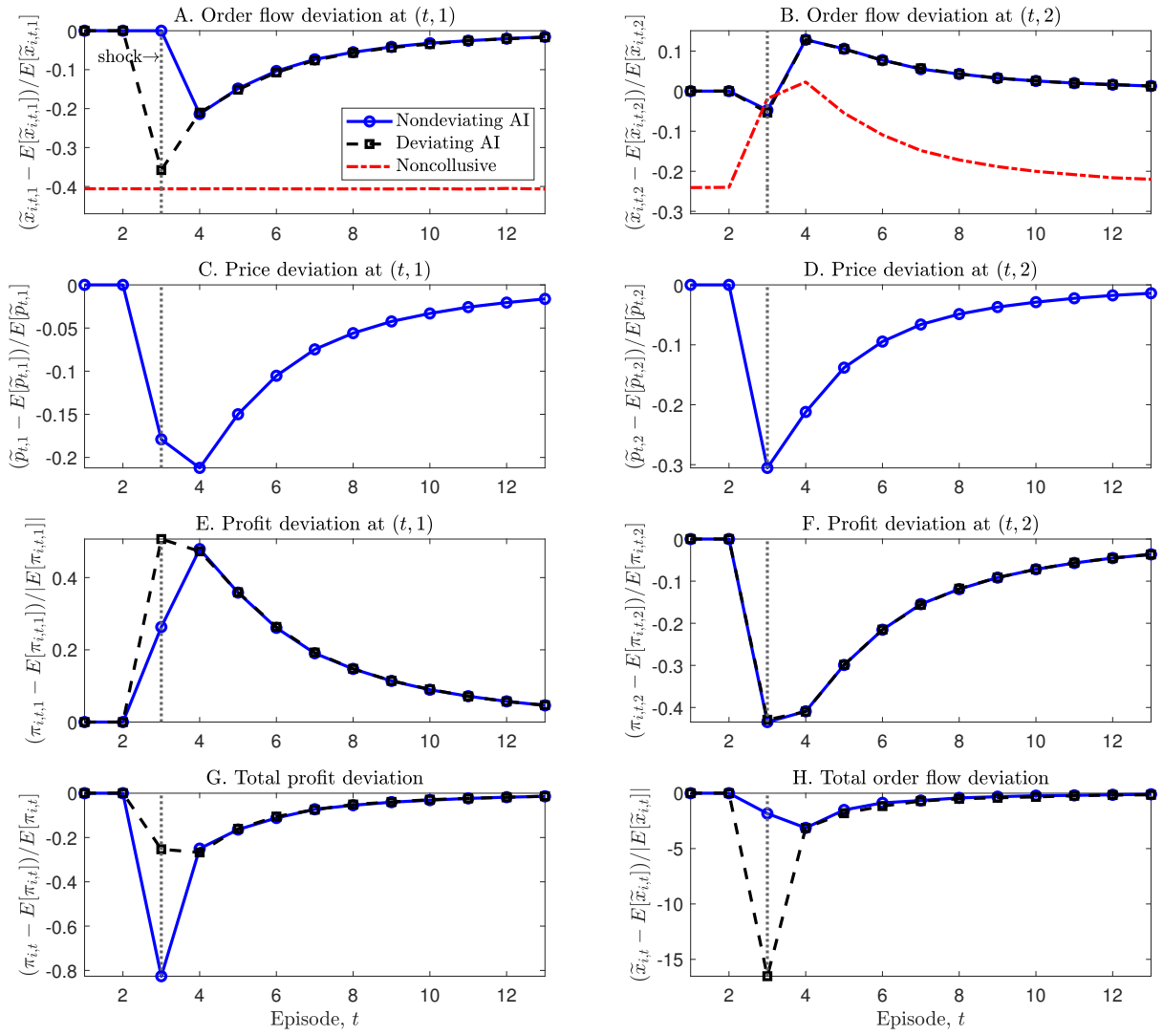
the fourth moment of changes in asset price to fundamental value, which is $\mathcal{S} = 4.75$. In comparison, the theoretical benchmark of the non-collusive Nash equilibrium during the same simulation period is $\mathcal{S}^N = 1.65$.

To illuminate the specific trading patterns learned by informed AI speculators and understand how they sustain intertemporal coordination, we conduct impulse-response analyses based on the trading behavior indicated by their fully converged Q-learning algorithms. In the absence of noise trading risk, any deviations in asset prices at $(t, 1)$ and $(t, 2)$ must result from changes in the trading strategies of the informed AI speculators. Therefore, we exogenously induce one informed AI speculator to deviate from its learned optimal trading strategy and examine how informed AI speculators and the market respond subsequently.

Impulse Responses to Deviations in the First Subperiod. In Figure 9, we analyze the impulse responses to a unilateral deviation in the first subperiod. Specifically, we exogenously force a single informed AI speculator to deviate from its learned optimal strategy at $(3, 1)$ in episode $t = 3$ for one subperiod, uniformly across all N_{sim} simulation paths, while the other informed AI speculator maintains its original trading behavior at $(3, 1)$. This one-subperiod deviation at $(3, 1)$ is designed to increase the contemporaneous trading profit of the deviating AI speculator in episode $t = 3$. Specifically, we compel the informed AI speculator to trade more conservatively at $(3, 1)$, meaning that the deviating AI speculator's order flow is reduced by $x_{i,shock}$ (i.e., taking smaller long positions) if $\omega_t > 0$ and increased by $x_{i,shock}$ (i.e., taking smaller short positions) if $\omega_t < 0$. For illustration purposes, we choose a large magnitude of $x_{i,shock}$. In each panel of Figure 9, the horizontal x-axis starts in episode $t = 1$, where all $N_{sim} = 1,000$ simulation sessions have converged. The dashed line represents the deviator while the solid line represents the non-deviator. Each impulse-response curve represents the average impulse response dynamics across N_{sim} independent simulation sessions.[15]

To clearly present the simulation results, we display the percentage deviations in informed AI speculators' order flows, asset prices, and profits from the long-run mean at $(t, 1)$ in the left column (panels A, C, and E of Figure 9) and at $(t, 2)$ in the right column (panels B, D, and F of Figure 9) for each episode $t \geq 1$. The percentage deviation of

[15] As in Dou, Goldstein and Ji (2025), each of the N_{sim} simulation sessions averages 10,000 simulation paths to smooth out the randomness of ω_t and ν_t, ensuring a reasonable comparison with the impulse response analysis based on the model of Calvano et al. (2020), which focuses on a deterministic economic environment.



Note: As indicated by the arrow in panel A, the exogenous deviation occurs at $(3,1)$ in episode $t = 3$.

Figure 9: Impulse responses following a unilateral deviation in trading order flow in the first subperiod in the baseline environment with $\mu = 0.3$.

each informed AI speculator's order flow from its long-run mean at $(t,1)$ is defined as $(\widetilde{x}_{i,t,1} - \mathbb{E}[\widetilde{x}_{i,t,1}])/\mathbb{E}[\widetilde{x}_{i,t,1}]$, where $\widetilde{x}_{i,t,1} = x_{i,t,1} \times \text{sgn}(\omega_t)$. Similarly, at $(t,2)$, the percentage deviation is defined as $(\widetilde{x}_{i,t,2} - \mathbb{E}[\widetilde{x}_{i,t,2}])/\mathbb{E}[\widetilde{x}_{i,t,2}]$, where $\widetilde{x}_{i,t,2} = x_{i,t,2} \times \text{sgn}(\omega_t)$. The sign function $\text{sgn}(\cdot)$ ensures that $\widetilde{x}_{i,t,1} > 0$ and $\widetilde{x}_{i,t,2} < 0$. The percentage deviation of the asset price from its long-run mean at $(t,1)$ is defined as $(\widetilde{p}_{t,1} - \mathbb{E}[\widetilde{p}_{t,1}])/\mathbb{E}[\widetilde{p}_{t,1}]$, where $\widetilde{p}_{t,1} = p_{t,1} \times \text{sgn}(\omega_t)$. At $(t,2)$, this is defined as $(\widetilde{p}_{t,2} - \mathbb{E}[\widetilde{p}_{t,2}])/\mathbb{E}[\widetilde{p}_{t,2}]$, where $\widetilde{p}_{t,2} = p_{t,2} \times \text{sgn}(\omega_t)$. The sign function $\text{sgn}(\cdot)$ ensures that $\widetilde{p}_{t,1}, \widetilde{p}_{t,2} > 0$. The percentage deviation of each informed AI speculator's profit from its long-run mean for the order flow placed at $(t,1)$ is defined as $(\pi_{i,t,1} - \mathbb{E}[\pi_{i,t,1}])/|\mathbb{E}[\pi_{i,t,1}]|$, and for $(t,2)$ it is defined as $(\pi_{i,t,2} - \mathbb{E}[\pi_{i,t,2}])/\mathbb{E}[\pi_{i,t,2}]$, where we take the absolute value for the denominator $|\mathbb{E}[\pi_{i,t,1}]|$ because $\pi_{i,t,1} < 0$. Panels G and H of Figure 9 present the percentage deviations of each informed AI speculator's total profit and total order flow from their long-run means in each episode t. These are defined as $(\pi_{i,t} - \mathbb{E}[\pi_{i,t}])/\mathbb{E}[\pi_{i,t}]$, where $\pi_{i,t} = \pi_{i,t,1} + \pi_{i,t,2}$, and $(\widetilde{x}_{i,t} - \mathbb{E}[\widetilde{x}_{i,t}])/|\mathbb{E}[\widetilde{x}_{i,t}]|$, where $\widetilde{x}_{i,t} = x_{i,t} \times \text{sgn}(\omega_t)$ and $x_{i,t} = x_{i,t,1} + x_{i,t,2}$. We take the absolute value for the denominator $|\mathbb{E}[\widetilde{x}_{i,t}]|$ because $\widetilde{x}_{i,t}$ can be negative in some cases.

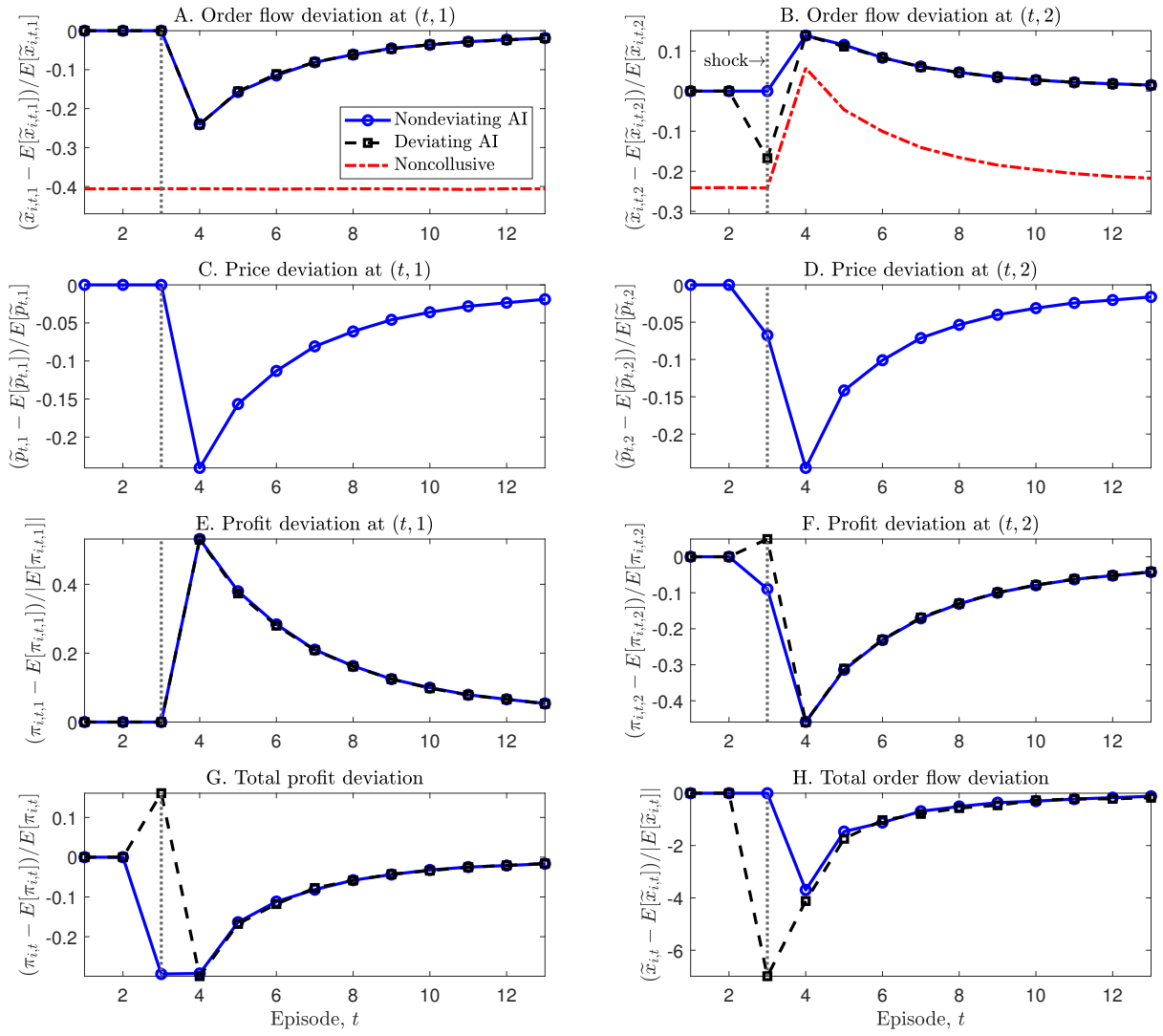
To investigate whether informed AI speculators can learn the compounded price-trigger strategy discussed in Section 3.6, we benchmark the order flows of informed AI speculators against the theoretical benchmark of the non-collusive Nash equilibrium in panels A and B of Figure 9, as represented by the dash-dotted lines. By comparing the solid and dashed lines with the dash-dotted line in episode $t = 1$, which represents the long-run mean, it is evident that informed AI speculators trade more aggressively at $(1,1)$ (i.e., taking larger long positions when $\omega_t > 0$ and larger short positions when $\omega_t < 0$) and more conservatively at $(1,2)$ (i.e., taking smaller short positions when $\omega_t > 0$ and smaller long positions when $\omega_t < 0$) compared to the theoretical benchmark of the non-collusive Nash equilibrium. This behavior confirms the existence of intertemporal coordination.

The deviation at $(3,1)$ in episode $t = 3$, as shown in panel A of Figure 9, leads to a downward deviation in $\widetilde{p}_{t,1}$ from the long-run mean (illustrated in panel C of Figure 9), implying that the asset price $p_{t,1}$ at $(3,1)$ becomes more aligned with its fundamental value $f_{t,1}$. In theory, when the realized asset price $p_{t,1}$ in the first subperiod aligns more closely with its fundamental value, informed speculators should trade more conservatively in the second subperiod, regardless of the degree of collusion (see equations (3.18) and (3.24)). Indeed, this is clearly illustrated by the substantial increase in the order flow deviation at $(3,2)$ in the theoretical benchmark of the non-collusive Nash equilibrium,

indicated by the dash-dotted line in panel B of Figure 9. In stark contrast, the solid and dashed lines in panel B of Figure 9 demonstrate that informed AI speculators adjust their order flows in the opposite direction, trading more aggressively at $(3,2)$, even though the realized price $\widetilde{p}_{t,1}$ is much lower than the long-run mean at $(3,1)$. Consequently, the solid and dashed lines fall below the dash-dotted line at $(3,2)$, indicating that informed AI speculators are trading more aggressively compared to the non-collusive benchmark at $(3,2)$ (i.e., taking larger short positions when $\omega_t > 0$ and larger long positions when $\omega_t < 0$). This behavior aligns with the compounded price-trigger strategy described in Section 3.6, where a one-time excessively aggressive trading response is implemented in reaction to any price deviation occurring in the first subperiod.

As a result of aggressive trading at $(3,2)$, there is a notable downward drop in $\widetilde{p}_{t,2}$ in episode $t = 3$ (panel D of Figure 9). Panel F of Figure 9 shows that the deviations in order flow and price at $(3,2)$ lead to significant trading losses for both informed AI speculators at this time. As a result, within the deviation episode $t = 3$, the total profit of the deviating AI speculator falls below the long-run mean (panel G of Figure 9), despite the significant profit increase from the order flow in the first subperiod of episode $t = 3$ (panel E of Figure 9). Thus, the ability to implement a one-time harsh punishment in the second subperiod of the deviation episode, in response to any price deviation occurring in the first subperiod, allows informed AI speculators to sustain collusion more effectively. By adopting this strategy, the AI speculators effectively tackle the weak punishment issue discussed in Section 3.5.

From episode $t = 4$ onward, panels A and B of Figure 9 show that the order flows of both informed AI speculators gradually revert to the long-run mean, which exhibits intertemporal coordination. This convergence occurs over several episodes, indicating that the "punishment" enacted by informed AI speculators for the one-subperiod deviation at $(3,1)$ is lasting. Notably, in line with the compounded price-trigger strategy discussed in Section 3.6, the punishment itself is collusive, as the solid and dashed lines consistently remain above the dash-dotted line from $t = 4$ onward. This persistent punishment results in ongoing losses for both informed AI speculators in subsequent episodes (panel G of Figure 9), reinforcing the impact of the one-time harsh punishment at $(3,2)$ within the deviation episode and deterring informed AI speculators from deviating in the first subperiod.



Note: As indicated by the arrow in panel B, the exogenous deviation occurs at $(3,2)$ in episode $t = 3$.

Figure 10: Impulse responses following a unilateral deviation in trading order flow in the second subperiod in the baseline environment with $\mu = 0.3$.

Impulse Responses to Deviations in the Second Subperiod. We now examine the impulse responses to unilateral deviation in the second subperiod. As illustrated in panel B of Figure 10, in the second subperiod of episode $t = 3$, i.e., at $(3, 2)$, we exogenously force a single informed AI speculator (dashed line) to deviate from its learned optimal strategy for one subperiod, uniformly across all N_{sim} simulation paths, while the other informed AI speculator (solid line) maintains its original trading behavior at $(3, 2)$. This one-subperiod deviation at $(3, 2)$ is intended to increase the contemporaneous trading profit of the deviating AI speculator. That is, we compel the informed AI speculator to trade more aggressively at $(3, 2)$, meaning that the deviating AI speculator's order flow is decreased by $x_{i,shock}$ (i.e., taking larger short positions) if $\omega_t > 0$ and increased by $x_{i,shock}$ (i.e., taking larger long positions) if $\omega_t < 0$. For illustration purposes, we choose a large magnitude of $x_{i,shock}$. The dash-dotted lines in panels A and B of Figure 10 represent the order flows in the theoretical benchmark of the non-collusive Nash equilibrium, serving as a benchmark for evaluating the trading behavior of informed AI speculators.

As shown in panel D of Figure 10, the aggressive trading by the deviating AI speculator at $(3, 2)$ immediately results in a decline in the second-subperiod asset price $\widetilde{p}_{t,2}$ at $(3, 2)$. In contrast, the first-subperiod asset price $\widetilde{p}_{t,1}$ remains at the long-run mean for episode $t = 3$ (panel C of Figure 10), because the order flows of informed AI speculators in the first subperiod have already been submitted and thus align with the long-run mean for that episode (panel A of Figure 10). Panel G of Figure 10 indicates that the deviating AI speculator achieves higher total profits, while the non-deviating AI speculator incurs losses in episode $t = 3$. The change in total profit for episode $t = 3$ is entirely driven by the profit change from the order flow placed in the second subperiod (panel F of Figure 10), rather than from the order flow in the first subperiod (panel E of Figure 10).

At $(4, 1)$ in episode $t = 4$, panel A of Figure 10 shows that both informed AI speculators significantly reduce their order flows below the long-run mean, beginning to engage in conservative trading in the first subperiod immediately following the deviation at $(3, 2)$. This results in a notable downward deviation in the first-subperiod asset price $\widetilde{p}_{t,1}$ at $(4, 1)$ (panel C of Figure 10). The decline in $\widetilde{p}_{t,1}$ further triggers an even more substantial downward deviation in the second-subperiod asset price $\widetilde{p}_{t,2}$ of episode $t = 4$ (panel D of Figure 10), due to decreased order flows from positive-feedback investors. In response to a lower $\widetilde{p}_{t,1}$ at $(4, 1)$, both informed AI speculators trade more conservatively at $(4, 2)$, leading to positive deviations in order flows $\widetilde{x}_{i,t,2}$ at $(4, 2)$ (panel B of Figure 10). Consequently, the profit generated by the order flows in the second subperiod declines

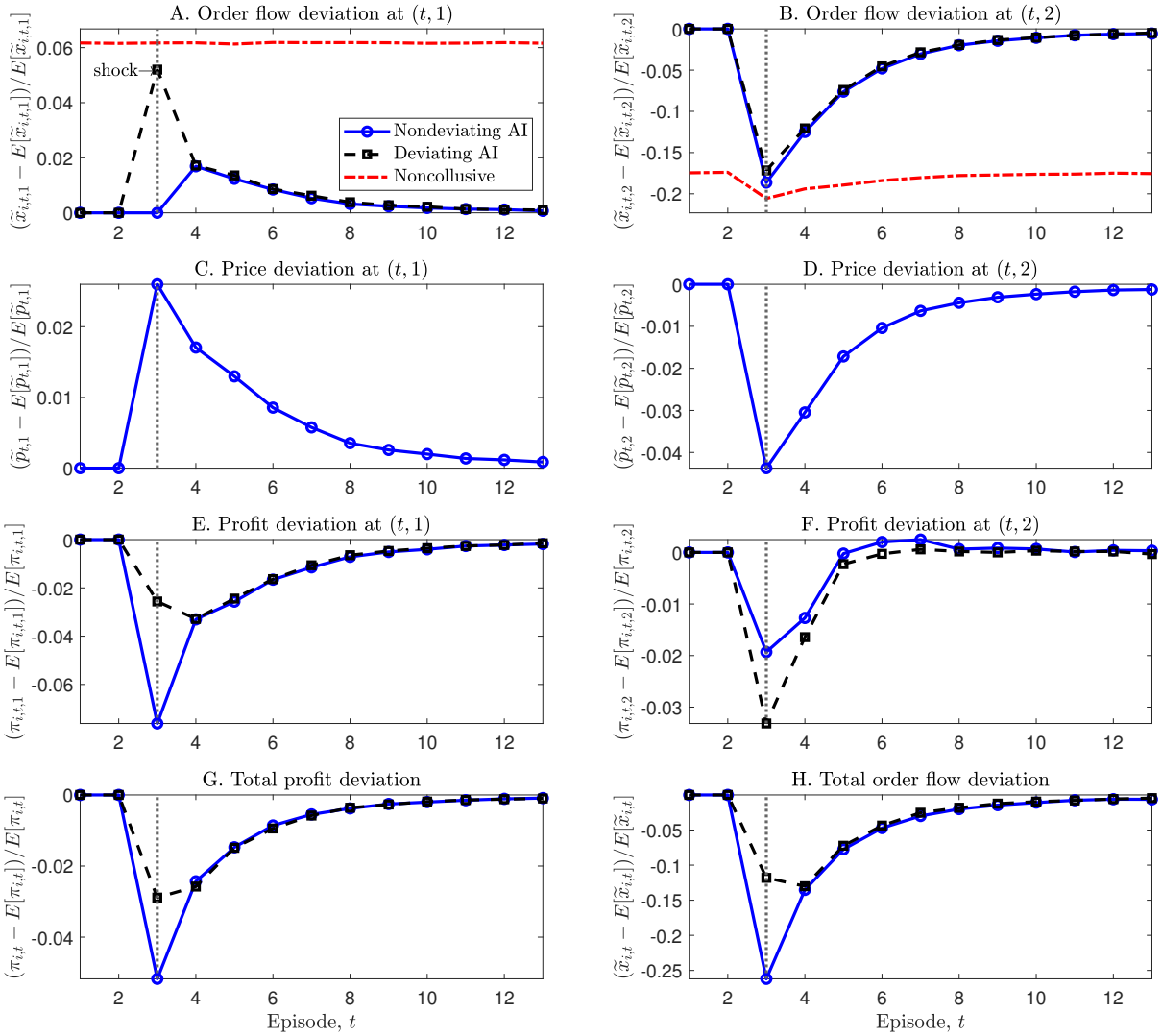
substantially (panel F of Figure 10), outweighing the increased profits from the first subperiod (panel E of Figure 10). This results in lower total profits for both informed AI speculators (panel G of Figure 10).

Overall, the trading behavior of informed AI speculators in episode $t = 4$ resembles a punishment strategy aimed at deterring the unilateral deviation at $(3, 2)$. This punishment persists for several episodes from $t = 4$ onward, being most severe in episode $t = 4$ and gradually reverting to the long-run mean, which exhibits intertemporal coordination. Importantly, the punishment itself is collusive, as the solid and dashed lines remain consistently above the dash-dotted line in panels A and B of Figure 10 for $t \geq 4$. Therefore, the trading behavior of informed AI speculators aligns with the collusive punishment discussed in Section 3.6. By implementing such a persistent collusive punishment strategy, informed AI speculators effectively address the time-inconsistency issue outlined in Section 3.4, ultimately rendering the unilateral deviation in the second subperiod unprofitable for the deviating AI speculator.

5.3 Simulation Experiments in Trading Environments with Low μ

We complement the simulation results presented above by conducting similar impulse-response analyses in trading environments with a limited presence of positive feedback investors, specifically with $\mu = 0.01$. In this context, our theoretical analysis in Section 3.5 shows that weak punishment issues do not arise when sustaining collusion. Therefore, it is unnecessary to implement a one-time excessively aggressive trading strategy to deter deviations occurring in the first subperiod. Consistent with the theory, we observe that informed AI speculators do not adopt such a one-time harsh punishment strategy. However, they still develop persistent collusive punishment strategies to maintain intertemporal coordination, effectively deterring deviations that occur in either the first or second subperiod.

Impulse Responses to Deviations in the First Subperiod. We first investigate the impulse responses to unilateral deviation in the first subperiod, as shown in Figure 11. Similar to the experiment in Figure 9, we exogenously force a single informed AI speculator (dashed line) to deviate from its learned optimal strategy for one subperiod at $(3, 1)$ in episode $t = 3$, with the intention of increasing contemporaneous trading profit. In the environment with a limited presence of positive-feedback investors, as discussed



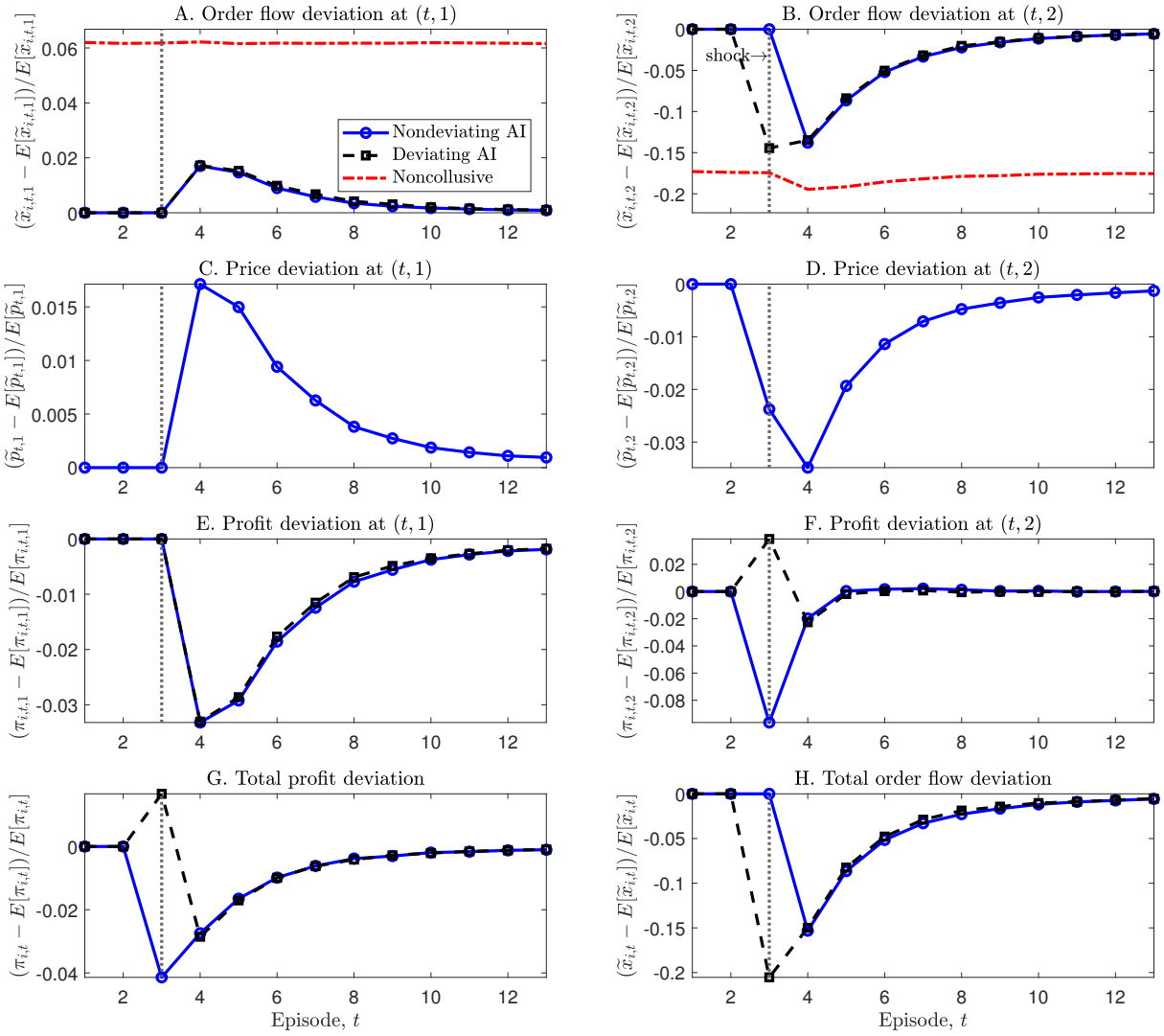
Note: As indicated by the arrow in panel A, the exogenous deviation occurs at $(3,1)$ in episode $t = 3$.

Figure 11: Impulse responses following a unilateral deviation in trading order flow in the first subperiod in the environment with $\mu = 0.01$.

in Section 3.3, profitable deviation means that the deviating AI speculator is compelled to trade more aggressively at $(3,1)$, i.e., taking larger long positions when $\omega_t > 0$ and larger short positions when $\omega_t < 0$. The direction of deviation at $(3,1)$ in panel A of Figure 11 is opposite to that in panel B of Figure 9, reflecting the absence of bubble-creation incentives in this trading environment.

As shown in panel C of Figure 11, the aggressive trading by the deviating AI speculator causes the first-subperiod asset price $\widetilde{p}_{t,1}$ to increase in episode $t = 3$. Panel B of Figure 11 indicates that this deviation in $\widetilde{p}_{t,1}$ from the long-run mean prompts both informed AI speculators to trade more aggressively at $(3,2)$ (i.e., taking larger short positions when $\omega_t > 0$ and long positions when $\omega_t < 0$) within the same episode $t = 3$. This behavior leads to significant downward deviation in the second-subperiod asset price $\widetilde{p}_{t,2}$ in episode $t = 3$ (panel D of Figure 11). As a result, even within the deviation episode $t = 3$, the deviating AI speculator experiences a decline in total profit relative to the long-run mean (panel G of Figure 11). Thus, the immediate second-subperiod punishment learned by informed AI speculators effectively deters any price deviation that occurs in the first subperiod, making such deviations unprofitable even within the deviating episode, which helps sustain collusion. In contrast to panel B of Figure 9, the solid and dashed lines do not cross the dash-dotted line at $(3,2)$ in panel B of Figure 11. This indicates that informed AI speculators do not adopt a one-time excessively aggressive trading strategy to deter deviations in the first subperiod in the trading environment characterized by a limited presence of positive feedback investors. In fact, even theoretically, there is no need to implement a one-time excessively aggressive trading strategy to punish deviation behavior because weak punishment issues do not arise when sustaining collusion in such environments.

Furthermore, panels A and B of Figure 11 show that from episode $t = 4$ onward, both informed AI speculators consistently adopt aggressive trading strategies in both the first and second subperiods for several episodes. Specifically, relative to the long-run mean, they significantly increase their order flows when trading in the first subperiod (panel A of Figure 11) and significantly decrease their order flows when trading in the second subperiod (panel B of Figure 11). Thus, in response to the deviation in the first period in episode $t = 3$, punishment not only occurs immediately in the second subperiod of the same episode $t = 3$ but also continues into future episodes. This leads to significant losses in discounted total profits (panel G of Figure 11), effectively preventing any deviation from occurring in the first subperiod. Overall, the punishment is weaker than the non-



Note: As indicated by the arrow in panel B, the exogenous deviation occurs at $(3,2)$ in episode $t = 3$.

Figure 12: Impulse responses following a unilateral deviation in trading order flow in the second subperiod in the environment with $\mu = 0.01$.

collusive Nash equilibrium trading strategies, as the solid and dashed lines consistently stay below the dash-dotted line in panel A of Figure 11 and above the dash-dotted line in panel B of Figure 11 for $t \geq 4$.

Impulse Responses to Deviations in the Second Subperiod. In Figure 12, we analyze the impulse responses to a unilateral deviation in the second subperiod. Similar to the experiment in Figure 10, we exogenously force a single informed AI speculator (dashed line) to deviate from its learned optimal strategy for one subperiod at $(3,2)$ in episode $t=3$, with the intention of increasing contemporaneous trading profit. This means that the deviating AI speculator is compelled to trade more aggressively at $(3,2)$, i.e., taking larger short positions when $\omega_t > 0$ and larger long positions when $\omega_t < 0$. As shown in panel D of Figure 12, the aggressive trading by the deviating AI speculator at $(3,2)$ immediately causes the second-subperiod asset price $\widetilde{p}_{t,2}$ to decline in episode $t=3$. Panel G of Figure 12 shows that the deviating AI speculator achieves higher total profits, while the non-deviating AI speculator incurs losses in episode $t=3$. The change in total profit in episode $t=3$ is driven entirely by the profit change from the order flow placed in the second subperiod (panel F of Figure 12), rather than from the order flow in the first subperiod (panel E of Figure 12).

From episode $t=4$ onward, both informed AI speculators adopt aggressive trading strategies in both the first and second subperiods, which persist for several episodes. As a result of aggressive trading, asset prices in both the first and second subperiods significantly deviate from their respective long-run means (panels C and D of Figure 12), leading to a dramatic reduction in the total profits of both informed AI speculators (panel G of Figure 12) and rendering the unilateral deviation unprofitable for the deviating AI speculator. Similar to the patterns observed in panels A and B of Figure 11, the punishment for $t \geq 4$ is weaker than the trading strategies in the theoretical benchmark of the non-collusive Nash equilibrium.

5.4 Coupled Q-learning vs. Reactive Q-learning

We now investigate the implications for financial market stability when informed AI speculators adopt reactive Q-learning algorithms. For comparisons, we use the same parameter values as our baseline experiment outlined in Section 4 for reactive Q-learning. We set the state vectors as $s_{t,1} = s_{t,2} = \{\omega_t, \omega_{t-1}, p_{t-1,2}\}$, which means that $p_{t,1}$ is exclude as a state variable at $(t,2)$ for reactive Q-learning algorithms.

Panel A of Figure 13 plots the average order flow $\widetilde{x}_{i,t,h} = x_{i,t,h} \times \text{sgn}(\omega_t)$ for $h = 0,\ldots,3$ of the two informed AI speculators in the three scenarios: informed AI speculators using coupled Q-learning algorithms (solid line), informed AI speculators with reactive Q-

learning algorithms (dotted line), and the theoretical benchmark of non-collusive Nash equilibirum (dash-dotted line). Each curve is plotted by taking the average order flows over $N_{sim} = 1,000$ sessions and $T = 100,000$ episodes. The sign function $\text{sgn}(\cdot)$ ensures that $\widetilde{x}_{i,t,1} > 0$ and $\widetilde{x}_{i,t,2} < 0$. Panel B of Figure 13 compares the average asset prices $\widetilde{p}_{t,h} = p_{t,h} \times \text{sgn}(\omega_t)$ across the three scenarios. The fundamental value of the asset is represented by $\widetilde{f}_{t,h} = f_{t,h} \times \text{sgn}(\omega_t)$ (dashed line). The sign function $\text{sgn}(\cdot)$ ensures that $\widetilde{p}_{t,h}, \widetilde{f}_{t,h} \geq 0$. Panels C and D of Figure 13 compares the average profit $\pi_{i,t,1}$ in each subperiod and total profit $\pi_{i,t} \equiv \pi_{i,t,1} + \pi_{i,t,2}$.

At $(t,1)$, panel A of Figure 13 shows that informed AI speculators using reactive Q-learning algorithms trade conservatively, with a trading magnitude smaller than that observed in the non-collusive Nash equilibrium and significantly lower than the trading flows in our baseline experiment, where informed AI speculators use coupled Q-learning algorithms. Panel B of Figure 13 shows that the asset price $\widetilde{p}_{t,1}$ at $(t,1)$ under reactive Q-learning is significantly lower than those in the non-collusive Nash equilibrium. Notably, the asset price $\widetilde{p}_{t,1}$ under reactive Q-learning is also lower than the average asset price $\widetilde{p}_{t,3}$ at $(t,3)$, which equals the average asset fundamental $\widetilde{f}_{t,3}$ due to the arbitrage trading activities of low-frequency, risk-neutral investors. This enables informed AI speculators to generate profits at $(t,1)$, as illustrated in panel C of Figure 13.

At $(t,2)$, panel A of Figure 13 shows that, under reactive Q-learning, informed AI speculators trade considerably smaller amounts than they would in the non-collusive Nash equilibrium at $(t,2)$ since no significant bubble emerges at $(t,2)$, as indicated by panel B of Figure 13. Panel C of Figure 13 shows that informed AI speculators using reactive Q-learning only makes small amounts of profits at $(t,2)$ due to the absence of significant asset price bubbles.

Overall, the trading patterns of informed AI speculators using reactive Q-learning algorithms and the resulting price dynamics differ significantly from those observed with coupled Q-learning. Specifically, reactive Q-learning algorithms do not generate substantial asset price bubbles. Their inability to recognize dynamic complementarity in trading means they fail to develop strategies that manipulate feedback traders through the creation of asset price bubbles. In fact, in our baseline trading environment where positive-feedback investors are prevalent (i.e., high μ), the trading patterns and market outcomes under reactive Q-learning resemble those found in environments with a limited presence of positive-feedback investors (i.e., low μ), as illustrated in Figure 3. As a result, informed AI speculators using reactive Q-learning algorithms achieve significantly

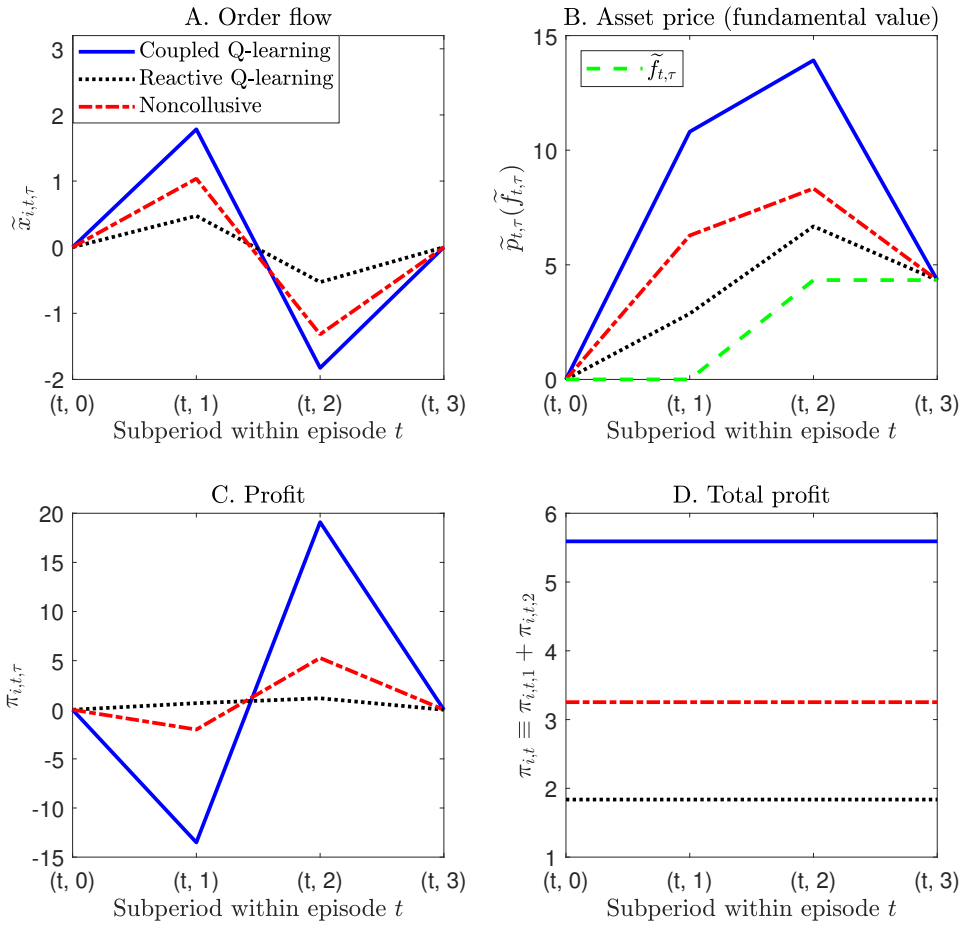


Figure 13: Coupled Q-learning vs. reactive Q-learning.

lower profits compared to those equipped with coupled Q-learning algorithms, as shown in panel D of Figure 13. Our simulation results indicate that multi-step AI planning, enabled by coupled Q-learning algorithms, is essential for discovering and executing intertemporal collusive trading strategies. These strategies create and exploit bubbles, manipulate feedback traders, and amplify market fragility.

6 Conclusion

This paper examines how AI planning—the core technology behind Agentic AI systems that reason over sequences of actions toward long-term goals—affects financial market

stability. By conducting simulations of reinforcement learning algorithms with explicit planning modules in a theory-driven laboratory setting, we demonstrate that informed AI speculators trained by these algorithms autonomously discover and implement intertemporally coordinated trading strategies based on compounded price-trigger rules, coordinating without communication or intent. The impact of collusive AI planning on financial market stability critically depends on the prevailing market environment. In our baseline trading environment characterized by a significant presence of positive-feedback investors, informed AI speculators learn to strategically coordinate their actions, creating and exploiting asset price bubbles, thereby significantly amplifying market instability. In contrast, when positive-feedback trading is limited, the intertemporal coordination among informed AI speculators does not significantly impact market stability and may even enhance it, albeit mildly.

Our findings have important implications for financial regulation and oversight in the era of AI planning. While our study emphasizes one specific aspect of the financial market, namely, the presence of positive-feedback investors, future research could examine how AI planning interacts with other market elements, such as market making and various sources of noise trading. Additionally, investigating the potential for AI to learn more complex forms of cooperation, possibly across different asset markets, warrants further exploration.

References

Abada, Ibrahim, and Xavier Lambin. 2023. "Artificial Intelligence: Can Seemingly Collusive Outcomes Be Avoided?" *Management Science*, 69(9): 5042–5065.

Asker, John, Chaim Fershtman, and Ariel Pakes. 2022. "Artificial Intelligence, Algorithm Design, and Pricing." *AEA Papers and Proceedings*, 112: 452–56.

Asker, John, Chaim Fershtman, and Ariel Pakes. 2024. "The Impact of Artificial Intelligence Design on Pricing." *Journal of Economics & Management Strategy*, 33(2): 276–304.

Banchio, Martino, and Giacomo Mantegazza. 2024. "Artificial Intelligence and Spontaneous Collusion." Working papers.

Barber, Brad M., Terrance Odean, and Ning Zhu. 2008. "Do Retail Trades Move Markets?" *The Review of Financial Studies*, 22(1): 151–186.

Bellman, Richard Ernest. 1954. *The Theory of Dynamic Programming*. Santa Monica, CA:RAND Corporation.

Brunnermeier, Markus K., and Lasse Heje Pedersen. 2008. "Market Liquidity and Funding Liquidity." *The Review of Financial Studies*, 22(6): 2201–2238.

Bryzgalova, Svetlana, Anna Pavlova, and Taisiya Sikorskaya. 2025. "Strategic Arbitrage in Segmented Markets." *Journal of Financial Economics*, 166(104008).

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicoló, and Sergio Pastorello. 2020. "Artificial Intelligence, Algorithmic Pricing, and Collusion." *American Economic Review*, 110(10): 3267–3297.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicoló, and Sergio Pastorello. 2021. "Algorithmic Collusion with Imperfect Monitoring." *International Journal of Industrial Organization*, 79(C).

Cao, Sean, Wei Jiang, Junbo Wang, and Baozhong Yang. 2024. "From Man vs. Machine to Man + Machine: The Art and AI of Stock Analyses." *Journal of Financial Economics*, 160(103910).

Carlin, Bruce Ian, Miguel Sousa Lobo, and S Viswanathan. 2007. "Episodic Liquidity Crises: Cooperative and Predatory Trading." *Journal of Finance*, 62(5): 2235–2274.

Cartea, Álvaro, Patrick Chang, Mateusz Mroczka, and Roel Oomen. 2022. "AI-Driven Liquidity Provision in OTC Financial Markets." *Quantitative Finance*, 22(12): 2171–2204.

Chen, Hsuan-Chi, and Jay R. Ritter. 2000. "The Seven Percent Solution." *Journal of Finance*, 55(3): 1105–1131.

Chen, Hui, Yuhan Cheng, Yanchu Liu, and Ke Tang. 2025. "Teaching Economics to the Machines." Working papers.

Chen, Yifei, Bryan T. Kelly, and Dacheng Xiu. 2024. "Expected Returns and Large Language Models." Working papers.

Christie, William G, and Paul H Schultz. 1994. "Why Do NASDAQ Market Makers Avoid Odd-Eighth Quotes?" *Journal of Finance*, 49(5): 1813–1840.

Christie, William G., and Paul H. Schultz. 1995. "Policy Watch: Did Nasdaq Market Makers Implicitly Collude?" *Journal of Economic Perspectives*, 9(3): 199–208.

Christie, William G, Jeffrey H Harris, and Paul H Schultz. 1994. "Why Did NASDAQ Market Makers Stop Avoiding Odd-Eighth Quotes?" *Journal of Finance*, 49(5): 1841–1860.

Colliard, Jean-Edouard, Thierry Foucault, and Stefano Lovo. 2025. "Algorithmic Pricing and Liquidity in Securities Markets." *Review of Financial Studies*, forthcoming.

Cong, Lin, and Zhiguo He. 2019. "Blockchain Disruption and Smart Contracts." *Review of Financial Studies*, 32(5): 1754–1797.

de Long, J. Bradford, Andrei Shleifer, Lawrence H. Summers, and Robert J. Waldmann. 1990. "Positive Feedback Investment Strategies and Destabilizing Rational Speculation." *The Journal of Finance*, 45(2): 379–395.

Dolgopolov, Arthur. 2024. "Reinforcement Learning in a Prisoner's Dilemma." *Games and Economic Behavior*, 144(C): 84–103.

Dou, Winston Wei, Itay Goldstein, and Yan Ji. 2025. "AI-Powered Trading, Algorithmic Collusion, and Price Efficiency." University of Pennsylvania Working Papers.

Dou, Winston Wei, Wei Wang, and Wenyu Wang. 2023. "The Cost of Intermediary Market Power for Distressed Borrowers." Working papers.

Dugast, Jérôme, and Thierry Foucault. 2018. "Data Abundance and Asset Price Informativeness." *Journal of Financial Economics*, 130(2): 367–391.

Dugast, Jérôme, and Thierry Foucault. 2024. "Equilibrium Data Mining and Data Abundance." *Journal of*

Finance, 80(1): 211–258.

Dutta, Prajit K, and Ananth Madhavan. 1997. "Competition and Collusion in Dealer Markets." *Journal of Finance*, 52(1): 245–276.

Farboodi, Maryam, and Laura Veldkamp. 2020. "Long-Run Growth of Financial Data Technology." *American Economic Review*, 110(8): 2485–2523.

Farboodi, Maryam, and Laura Veldkamp. 2023. "Data and Markets." *Annual Review of Economics*, 15: 23–40.

Gao, Zhenyu, Wei Xiong, and Jian Yuan. 2024. "Structured Beliefs and Fund Performance: An LLM-Based Approach." Working papers.

Geanakoplos, John. 2010. "The Leverage Cycle." *NBER Macroeconomics Annual*, 24: 1–66.

Gennotte, Gerard, and Hayne Leland. 1990. "Market Liquidity, Hedging, and Crashes." *The American Economic Review*, 80(5): 999–1021.

Goldstein, Itay. 2022. "Information in Financial Markets and Its Real Effects*." *Review of Finance*, 27(1): 1–32.

Goldstein, Itay, Chester S Spatt, and Mao Ye. 2021. "Big Data in Finance." *Review of Financial Studies*, 34(7): 3213–3225.

Goldstein, Itay, Chong Huang, and Liyan Yang. 2025. "Fragility of Financial Markets." *Annual Review of Financial Economics*.

Green, Edward J, and Robert H Porter. 1984. "Noncooperative Collusion under Imperfect Price Information." *Econometrica*, 52(1): 87–100.

Greenwood, Robin, and Andrei Shleifer. 2014. "Expectations of Returns and Expected Returns." *The Review of Financial Studies*, 27(3): 714–746.

Grinblatt, Mark, Sheridan Titman, and Russ Wermers. 1995. "Momentum Investment Strategies, Portfolio Performance, and Herding: A Study of Mutual Fund Behavior." *The American Economic Review*, 85(5): 1088–1105.

Gromb, Denis, and Dimitri Vayanos. 2010. "Limits of Arbitrage." *Annual Review of Financial Economics*, 2(Volume 2, 2010): 251–275.

Gromb, Denis, and Dimitri Vayanos. 2018. "The Dynamics of Financially Constrained Arbitrage." *The Journal of Finance*, 73(4): 1713–1750.

Hansen, Karsten T., Kanishka Misra, and Mallesh M. Pai. 2021. "Algorithmic Collusion: Supra-Competitive Prices via Independent Algorithms." *Marketing Science*, 40(1): 1–12.

Hatfield, John William, Scott Duke Kominers, Richard Lowery, and Jordan M. Barry. 2020. "Collusion in Markets with Syndication." *Journal of Political Economy*, 128(10): 3779–3819.

Hong, Harrison, and Jeremy C. Stein. 1999. "A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets." *The Journal of Finance*, 54(6): 2143–2184.

Hörner, Johannes, Stefano Lovo, and Tristan Tomala. 2018. "Belief-Free Price Formation." *Journal of Financial Economics*, 127(2): 342–365.

Johnson, Justin Pappas, Andrew Rhodes, and Matthijs Wildenbeest. 2023. "Platform Design when Sellers Use Pricing Algorithms." *Econometrica*, 91(5): 1841–1879.

Kaelbling, Leslie Pack, Michael L. Littman, and Anthony R. Cassandra. 1998. "Planning and acting in partially observable stochastic domains." *Artificial Intelligence*, 101: 99–134.

Kaniel, Ron, Zihan Lin, Markus Pelger, and Stijn Van Nieuwerburgh. 2023. "Machine-Learning the Skill of Mutual Fund Managers." *Journal of Financial Economics*, 150(1): 94–138.

Kelly, Bryan T., Semyon Malamud, and Kangying Zhou. 2024. "The Virtue of Complexity in Return Prediction." *Journal of Finance*, 79(1): 459–503.

Kyle, Albert S. 1989. "Informed Speculation with Imperfect Competition." *Review of Economic Studies*, 56(3): 317–355.

Kyle, Albert S., and Wei Xiong. 2001. "Contagion as a Wealth Effect." *Journal of Finance*, 56(4): 1401–1440.

Lambin, Xavier. 2024. "Less than Meets the Eye: Simultaneous Experiments as a Source of Algorithmic Seeming Collusion." Working papers.

Lehar, Alfred, and Christine Parlour. 2025. "Market Power and the Bitcoin Protocol." Working papers.

Ljungqvist, Lars, and Thomas J. Sargent. 2012. *Recursive Macroeconomic Theory, Third Edition.* Vol. 1 of *MIT Press Books*. 3 ed., The MIT Press.

Marimon, Ramon, Ellen McGrattan, and Thomas J. Sargent. 1990. "Money as a Medium of Exchange in an Economy with Artificially Intelligent Agents." *Journal of Economic Dynamics and Control*, 14(2): 329–373.

Ozdenoren, Emre, and Kathy Yuan. 2008. "Feedback Effects and Asset Prices." *The Journal of Finance*, 63(4): 1939–1975.

Rostek, Marzena, and Ji Hee Yoon. 2024. "Imperfect Competition in Financial Markets: Recent Developments." *Journal of Economic Literature*, forthcoming.

Routledge, Bryan R. 1999. "Adaptive Learning in Financial Markets." *Review of Financial Studies*, 12(5): 1165–1202.

Routledge, Bryan R. 2001. "Genetic Algorithm Learning to Choose and Use Information." *Macroeconomic Dynamics*, 5(02): 303–325.

Sannikov, Yuliy, and Andrzej Skrzypacz. 2007. "Impossibility of Collusion under Imperfect Monitoring with Flexible Production." *American Economic Review*, 97(5): 1794–1823.

Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction.* The MIT Press.

Vayanos, Dimitri. 1999. "Strategic Trading and Welfare in a Dynamic Market." *Review of Economic Studies*, 66(2): 219–254.

Vayanos, Dimitri, and Jean-Luc Vila. 2021. "A Preferred-Habitat Model of the Term Structure of Interest Rates." *Econometrica*, 89(1): 77–112.

Waltman, Ludo, and Uzay Kaymak. 2008. "Q-learning Agents in a Cournot Oligopoly Model." *Journal of Economic Dynamics and Control*, 32(10): 3275–3293.

Watkins, Christopher J. C. H., and Peter Dayan. 1992. "Q-learning." *Machine Learning*, 8(3): 279–292.

Wermers, Russ. 1999. "Mutual Fund Herding and the Impact on Stock Prices." *The Journal of Finance*, 54(2): 581–622.