# AI-Powered Trading, Algorithmic Collusion, and Price Efficiency

Winston Wei Dou Itay Goldstein Yan Ji *

March 10, 2024

**Abstract**

The integration of algorithmic trading and reinforcement learning, known as AI-powered trading, has significantly impacted capital markets. This study utilizes a model of imperfect competition among informed speculators with asymmetric information to explore the implications of AI-powered trading strategies on speculators' market power, information rents, price informativeness, market liquidity, and mispricing. Our results demonstrate that informed AI speculators, even though they are "unaware" of collusion, can autonomously learn to employ collusive trading strategies. These collusive strategies allow them to achieve supra-competitive trading profits by strategically under-reacting to information, even without any form of agreement or communication, let alone interactions that might violate traditional antitrust regulations. Algorithmic collusion emerges from two distinct mechanisms. The first mechanism is through the adoption of price-trigger strategies ("artificial intelligence"), while the second stems from homogenized learning biases ("artificial stupidity"). The former mechanism is evident only in scenarios with limited price efficiency and noise trading risk. In contrast, the latter persists even under conditions of high price efficiency or large noise trading risk. As a result, in a market with prevalent AI-powered trading, both price informativeness and market liquidity can suffer, reflecting the influence of both artificial intelligence and stupidity.

**Keywords:** Reinforcement learning, AI collusion, Homogenization, Self-confirming equilibrium, Asymmetric information, Price informativeness, Market liquidity.
**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 Snehal Banerjee, Hui Chen, Antoine Didisheim, Itamar Drechsler, Slava Fos, Cary Frydman, Paolo Fulghieri, Vincent Glode, Joao Gomes, Mark Grinblatt, Tim Johnson, Chris Jones, Scott Joslin, Larry Harris, Zhiguo He, David Hirshleifer, Jerry Hoberg, Leonid Kogan, Pete Kyle, Tse-Chun Lin, Deborah Lucas, Ye Luo, Semyon Malamud, Andrey Malenko, George Malikov, Albert Menkveld, Jonathan Parker, Lasse Pedersen, Josh Pollet, Paul Romer, Nick Roussanov, Tom Sargent, Antoinette Schoar, Daniel Sokol, Rob Stambaugh, Eric Talley, Anton Tsoy, Liyan Yang, Jiang Wang, Neng Wang, Xiaoyan Zhang, and seminar and conference participants at ASU Sonoran Winter Finance Conference, Boston College, CUFE, Fudan, George Mason, HKU, HKUST, Jackson Hole Finance Conference, Johns Hopkins Carey Finance Conference, Melbourne Asset Pricing Meeting, MIT, Nordic Fintech Symposium, NYU/Penn Law and Finance Conference, Olin Finance Conference at WashU, PKU/PHBS Sargent Institute Macro-Finance Workshop, QES Global Quant and Macro Investing Conference, QRFE Workshop on Market Microstructure, Fintech and AI, SHUFE, Toronto Macro/Finance Conference, Tsinghua PBCSF, UIUC, University of Macau, University of Toronto, USC, and Wharton for their comments. We are grateful for the insightful discussions with Jacob Yunger and other colleagues at the Financial Industry Regulatory Authority (FINRA). Dou is grateful for the financial supports from the Golub Faculty Scholar Award at Wharton.

# 1 Introduction

The integration of algorithmic trading with reinforcement learning (RL) algorithms, often termed AI-powered trading, poses new regulatory challenges and has the potential to fundamentally reshape capital markets.[1] With Nasdaq receiving SEC approval for an RL-based, AI-driven order type, the momentum for AI integration in trading continues to build. Leading digital trading platforms like MetaTrader are endorsing RL-based AI trading bots, and major hedge funds such as Two Sigma, along with investment powerhouses like Blackrock and J.P. Morgan, are adopting AI technologies. This trend has led policymakers, regulators, and financial market supervisors worldwide to make AI a regulatory priority. Their focus is now on understanding how AI is applied in financial markets, its potential implications, and the risks of unintended systemic impacts.[2]

In particular, the U.S. Security and Exchange Commission (SEC) has recently cautioned against the possibility of AI destabilizing the global financial market if big tech-based trading companies monopolize AI development and applications within the financial sector. The SEC points out that the real challenge is fostering competitive and efficient markets amidst the swift adoption of AI technologies, as AI might be optimized to benefit sophisticated speculators at the expense of other investors, potentially compromising competition and market efficiency. Notably, SEC Chair Gary Gensler has emphasized this concern, noting that there is evidence of machines in high-frequency trading starting to exhibit cooperative behavior independently of human intervention or interaction.

Promoting competition in financial markets is a primary objective of the SEC and similar regulatory bodies worldwide. As such, the potential for collusion among AI trading algorithms is a significant concern for these organizations. However, the underlying scientific and economic principles of such "cooperation" among autonomous AI algorithms remain unclear, not to mention how it might affect competition, price formation, and overall market efficiency. In this paper, we demonstrate that "AI collusion" – where autonomous, self-interested algorithms independently learn to coordinate without any explicit agreement, communication, or intention – can robustly occur via one of two distinct mechanisms. These mechanisms are collusion through price-trigger strategies or homogenized learning biases, and their emergence is contingent on the condition of the trading environment. We find that AI collusion impairs competition and thereby market efficiency, leading to reduced liquidity, less informative pricing, and increased mispricing.

The economics of AI collusion in trading can be intuitively understood as follows. On one hand, consider a trading environment where subgame perfect collusive Nash equilibria theoretically exist for rational-expectations agents, supported by price-trigger strategies as introduced by Green and Porter (1984). In this environment, even without direct monitoring of trading behaviors, agents can develop collusive incentives. This is achieved by allowing non-collusive competition to

[1] Traditional algorithmic trading is based on rigid, human-defined trading protocols that are hardcoded.

[2] For example, the SEC proposed novel rules concerning the application of AI technologies (SEC, 2023). Additionally, the European Securities and Markets Authority (ESMA) published a report on AI utilization within EU securities markets (Bagattini, Benetti and Guagliano, 2023).

occur when market prices diverge from the expected collusive level beyond a certain threshold. If the trading environment is not overly disrupted by noise trading flows, AI algorithms have the capacity to interact and learn, ultimately achieving a steady state, within which they engage in collusive trading based on a price-trigger strategy, even though they might not achieve the most profitable collusive equilibrium, due to a learning bias. On the other hand, in a trading environment where subgame perfect collusive Nash equilibria do not theoretically exist, AI algorithms cannot learn to sustain collusion through price-trigger strategies. Instead, they may converge to a steady state characterized by a self-conforming equilibrium, as introduced by Fudenberg and Levine (1993). This equilibrium concept, weaker than Nash equilibrium, allows for potentially incorrect or biased off-equilibrium beliefs, tightly aligned with the learning and trading behaviors of AI algorithms. Beliefs may be accurate along the equilibrium path, as this is more commonly observed, but can be inaccurate off the equilibrium path, unless there is sufficient exploration of non-optimal actions (e.g., Fudenberg and Kreps, 1988, 1995; Cho and Sargent, 2008). Crucially, these incorrect off-equilibrium beliefs are not necessarily inconsistent with observed outcomes along the equilibrium path.

Notably, AI algorithms are distinct from human traders in that they do not simply mimic human behavior. Traditional theories and experimental studies about human behavior are insufficient for understanding AI traders' behavior and the equilibria they might form. This is because AI possesses a fundamentally different form of intelligence. Unlike humans, AI decision-making is not influenced by emotions or logical thinking; rather, it is driven primarily by pattern recognition and is not affected by higher-order beliefs. Therefore, understanding the dynamics of capital markets with the prevalence of AI-powered trading algorithms requires insights into algorithmic behavior akin to the "psychology" of machines (Goldstein, Spatt and Ye, 2021), in a similar vein to how decision theory and psychology literature have provided insights into modeling human behavior in an economic context. In this paper, we conduct an experimental study to examine the behavior of AI algorithms endowed with private information. Following the tradition of experimental research, our study is qualitative and intended as a proof-of-concept demonstration.

In this paper, we adopt a streamlined theoretical framework as our laboratory. Building upon the seminal work of Kyle (1985), we extend this framework in three novel ways. First, our model incorporates multiple informed speculators within a repeated-trading context. Second, we introduce a continuum of atomistic long-term preferred-habitat investors, who together create a collective downward-sloping demand curve. Third, we expand the role of the market maker to consider both inventory costs and pricing errors, thereby extending beyond the original model's focus on pricing errors alone, as in Kyle (1985). Within each trading period, agents execute a single transaction. The sequence of events for each period unfolds as follows: Initially, the fundamental value of the asset is determined. Subsequently, a continuum of noise traders collectively places an order flow, which is independent of the asset's fundamental value. The variance of such an aggregate noise trading flow encapsulates the noise trading risk (Long et al., 1990). This noise trading risk is a crucial characteristic of the trading environment. Each oligopolistic informed

speculator is aware of the fundamental value but remains uninformed about the noise trading flow when determining his or her optimal trading strategy. The market maker, in turn, sets the market price with the goal of minimizing the weighted average of inventory costs and pricing errors. In doing so, the market maker also takes into account the price elasticity of the preferred-habitat investors' demand. This price elasticity represents another critical characteristic of the trading environment.

In our experimental study, we position our subjects – AI algorithms – within the laboratory framework we have established. Specifically, we substitute the rational-expectations informed speculators and market maker as in Kyle (1985)'s model with Q-learning algorithms. These algorithms are tasked with learning and guiding the real-time trading decisions. Known for their simplicity, transparency, and economic interpretability, Q-learning algorithms provide a foundational basis for various RL procedures that have significantly advanced the AI domain. Our theoretical framework, coupled with simulation-based experiments that blend theoretical rigor with practical relevance, serves as a laboratory for examining the impact of AI-powered trading strategies. Specifically, it allows us to investigate their influence on the market power of informed AI speculators, as well as on the price formation process, including implications for market liquidity, price informativeness, and mispricing within financial markets.

To ascertain whether informed AI speculators' behavior exhibits collusion sustained by price-trigger strategies due to the intelligence of the algorithms, our analysis starts with examining the theoretical properties of tacit collusion that can be maintained through price-trigger strategies. This analysis is based on the assumption that both the informed speculators and the market maker operate under rational expectations and have a thorough understanding of the preferred-habitat demand curve. We examine how tacit collusion varies across different trading environments. This includes variations in the price elasticity of preferred-habitat investors and noise trading risk levels, as well as variations in the number of informed speculators and their time discount rates. This theoretical investigation enables us to establish a baseline understanding of collusive behavior in the presence of asymmetric information and the endogenous strategic pricing rules of the market maker. Importantly, it lays the groundwork for our experimental study on the AI trading behavior, wherein we assess whether the observed collusion of informed AI speculators aligns with the theoretical predictions under the assumption of rational expectations and perfect knowledge of the preferred-habitat demand curve.

As a noteworthy theoretical contribution, we establish a novel result on the impossibility of collusion under information asymmetry. We demonstrate that informed speculators are unable to achieve collusive outcomes through price-trigger strategies in certain conditions. This includes scenarios where market prices are already efficient, accurately reflecting the asset's fundamental value, especially when the preferred-habitat investor has high price elasticity of demand, thereby playing a minimal role in price formation. Another scenario precluding collusion is when the noise trading risk is excessively high. This novel result illuminates a mechanism distinct from existing theories on the impossibility of collusion under information asymmetry in the context of product market competition (Abreu, Milgrom and Pearce, 1991; Sannikov and Skrzypacz, 2007).

Intuitively, sustaining price-trigger collusion requires two conditions: first, monitoring necessitates high price informativeness, and second, maintaining informational rents requires a low price impact of informed trading. These two conditions cannot be simultaneously met when price efficiency or noise trading risk is high.

Furthermore, as an additional theoretical contribution, we illustrate that in scenarios where the preferred-habitat investor, exhibiting low price elasticity of demand, significantly influences price formation, market prices can become inefficient. In such cases, tacit collusion among informed speculators can be sustained through price-trigger strategies. The success of these strategies is contingent on the number of informed speculators and the level of noise trading risk in the market. We find that price-trigger strategies can only sustain collusion in markets with a low level of noise trading risk and a few informed speculators. Additionally, we show that collusion capacity increases, market liquidity decreases, price informativeness decreases, and mispricing increases, when the number of informed speculators drops, the level of noise trading risk decreases, or the subjective rate of time preference (i.e., "impatience") declines.

Having established the baseline theoretical results, we now turn back to our simulation experiments, which involve informed AI speculators using Q-learning algorithms. These simulations provide compelling evidence that these AI speculators can robustly collude and secure supra-competitive profits by strategically manipulating excessively low order flows relative to their information about the asset's fundamental value. This occurs without any form of agreement or communication that would typically be seen as an antitrust infringement. The cruciality, and even necessity, of communication in collusion among humans is well-documented in the literature of experimental economics. To underscore the concept of AI collusion in our simulations, we deliberately employ relatively simple Q-learning algorithms that base their decisions solely on one-period-lagged asset prices as state variables. This approach is intentional, omitting more extensive lagged data, such as information on lagged self-order flows or multiple-period-lagged asset prices. Although the trading environment is excessively complex relative to the simple AI algorithms used, our simulation results remarkably indicate that informed AI speculators can intelligently form collusion across diverse trading environments. Specifically, in environments characterized by low price efficiency and low noise trading risk, the behavior of algorithmic collusion aligns with the predictions of our rational-expectations model, where informed AI speculators are capable of learning price-trigger strategies to sustain collusion. Conversely, in environments with high price efficiency or high noise trading risk, informed AI speculators are unable to learn price-trigger strategies, consistent with our rational-expectations model predictions. However, strikingly, going beyond the rational-expectations model, our simulation results demonstrate that informed AI speculators can still collude and achieve supra-competitive profits by manipulating excessively low order flows, even without relying on traditional price-trigger strategies, provided they use equally naive algorithms. These findings suggest the existence of two distinct mechanisms underpinning algorithmic collusion, depending on the trading environment.

Finally, we elaborate further on the two distinct mechanisms behind AI collusion across various trading environments. The first mechanism, known as "algorithmic collusion through

price-trigger strategies," involves a form of collusion driven by "artificial intelligence." In this scenario, informed AI speculators have the capability to learn and implement price-trigger strategies effectively. This price-trigger strategy enables the AI speculators to sustain collusion and reach a steady state closely resembling a subgame perfect Nash equilibrium. Such a scenario can only occur if both price efficiency and noise trading risk are low. Leveraging simulation experiments, we provide direct evidence that sizable price deviations trigger aggressive trading flows similar to those in a non-collusive Nash equilibrium, which diminishes the trading profits of all informed AI speculators. While the underlying mechanisms through which AI speculators learn to conduct the price-trigger trading strategy, thereby achieving algorithmic collusion, may differ from those behind how humans would learn to coordinate using price-trigger trading strategies, the resulting patterns exhibit notable similarities. At the heart of these mechanisms, whether involving AI or human speculators, the threat of punishment effectively acts as a deterrent, discouraging individual speculators from violating the collusive agreement. Closely aligned with the theoretical predictions of a collusive Nash equilibrium sustained by price-trigger strategies with rational-expectations agents, as the number or impatience of speculators decreases, the extent of achievable collusion increases. This leads to reduced market liquidity, diminished price informativeness, and increased mispricing.

Importantly, algorithmic collusion through price-trigger strategies introduces a paradoxical situation concerning price informativeness. This paradox arises because such collusion relies on the informativeness of prices – specifically, the ability of an informed AI speculator to infer the order flows of other informed AI speculators from observed prices. High price informativeness typically characterizes environments where prices are sensitive to new information about the fundamental value of the asset and are not predominantly driven by noise trading flows. However, in such environments, the heightened price informativeness actually facilitates informed AI speculators in discerning each other's order flows, thereby strengthening collusion among them. This stronger collusion, in turn, endogenously compromises price informativeness by distorting the information content of prices – specifically, it reduces the responsiveness of prices to new information about the fundamental value of the asset. Consequently, in a capital market dominated by AI-powered trading, where algorithmic collusion through price-trigger strategies is prevalent, achieving perfect price informativeness becomes unattainable.

The second mechanism, known as "algorithmic collusion through homogenized learning biases," involves a form of collusion driven by "artificial stupidity." Despite the learning biases originating from intrinsic imperfections in the algorithms, informed AI speculators might still achieve and sustain supra-competitive profits. This can occur when they use similar foundational models that have homogenized learning biases, effectively forming a kind of hub-and-spoke conspiracy.[3] Johnson and Sokol (2021) emphasize the prevalence of this type of AI collusion in the context of e-commerce platforms, observing that many retailers adopt similar or even identical AI

[3]In the context of product market competition, the term "hub-and-spoke conspiracy" is a metaphor used to describe a cartel that includes a firm at one level of a supply chain, typically a supplier, acting as the "hub" of a wheel. Vertical agreements down the supply chain represent the "spokes." This common supplier facilitates the implicit coordination among its customers.

pricing algorithms. Specifically, anti-competitive effects may emerge when multiple competitors use the same AI pricing algorithm supplied by a common service provider, who serves as the hub. In the financial markets, informed speculators often rely on similar foundational models for their AI-powered trading systems. This practice, whether intentional or not, can result in a significant degree of homogenization, a phenomenon documented by Bommasani et al. (2022), among others. In the context of RL learning, the emergence of a learning bias is directly linked to inconsistencies in statistical learning. These inconsistencies often stem from over-exploitation and insufficient exploration, especially when the noise trading risk is excessive. This inherently biased algorithm leads informed speculators to under-react to their private information in their learned trading strategies, compared to the optimal strategy in a non-collusive equilibrium setting. Consider a scenario in which an RL-based AI speculator explores a trading strategy that aggressively responds to private information and receives a positive signal about the asset's fundamental value. If a substantial and positive noise trading flow occurs, this could result in significant losses for the AI speculator. Consequently, the RL algorithm is unlikely to revisit and update its understanding of this state-strategy pair sufficiently, consistently deeming this strategy as suboptimal for the given state. This means the initial adverse effect on the Q function at the state-strategy pair due to such a shock is unlikely to be mitigated in subsequent iterations. Conversely, if a substantial and negative noise trading flow occurs, it could lead to significant gains for the AI speculator. In this fortunate case, the RL algorithm is more likely to revisit and thoroughly understand the performance of this state-strategy pair, adequately exploiting it, and thus, the initial beneficial effect on the Q function at this pair may be averaged out, which even leads to accurate estimations of Q function at this state-strategy pair. Such severe asymmetric learning outcomes from large positive and negative noise trading flows can lead AI speculators to generally under-react to their private information in their learned trading strategies.

Such under-reaction can lead to the realization of supra-competitive profits, a scenario that is more likely to occur with widespread homogenization in the algorithms adopted by AI speculators. This homogenized learning bias steers informed AI speculators toward a steady state where trading behaviors can be accurately characterized by a self-conforming equilibrium, as introduced by Fudenberg and Levine (1993). In contrast to the Nash equilibrium, the self-conforming equilibrium is weaker because it permits players to hold incorrect (or biased) off-equilibrium beliefs. This concept of equilibrium is motivated by the idea that noncooperative equilibria should be interpreted as outcomes of a learning process, where players form beliefs based on their past experiences. While beliefs can generally be correct along the equilibrium path of play due to its frequent observation, they are not necessarily correct off the equilibrium path. Correct beliefs off the equilibrium path require players to engage in sufficient experimentation with non-optimal actions, as suggested in works by Fudenberg and Kreps (1988), Fudenberg and Kreps (1995), and Cho and Sargent (2008).

Although adopting superior algorithms can disrupt the collusion created by homogenized learning biases, it is likely that no AI speculator would choose to gain an advantage by using superior algorithms due to the nature of AI collusion. Intuitively, if one speculator adopts a

superior algorithm, it could render the trading strategies of other AI speculators unprofitable, thereby compelling them to adopt equally or more advanced algorithms. This could spark a race towards algorithmic advancement, ultimately leading to an equilibrium where trading profitability is minimal for every AI speculator. Consequently, AI speculators autonomously learn to adopt similarly basic algorithms in equilibrium. To illustrate this point, we consider a simple extension of the baseline Q-learning algorithms, wherein informed AI speculators are able to learn both the key parameter that governs the sophistication of their Q-learning algorithms and their trading strategies based on the AI-chosen Q-learning algorithm. Our simulation experiments robustly demonstrate that informed AI speculators may collectively opt for less advanced algorithms. This occurs despite the potential for increased self-profit that could come from unilaterally choosing a more advanced algorithm while others' algorithms remain fixed.

These two types of AI collusion, while both generating supra-competitive trading profits, can exhibit opposite collusive behaviors as trading environments evolve. On one hand, akin to AI collusion through price-trigger strategies (referred to as "artificial intelligence"), a decrease in the number of speculators leads to increased potential for collusion. This, in turn, results in reduced market liquidity, diminished price informativeness, and increased mispricing. On the other hand, contrary to AI collusion through price-trigger strategies, an increase in speculator impatience, or an elevation in noise trading risk, enlarges the potential for collusion due to a more pronounced homogenized learning bias. (termed "AI collusion through artificial stupidity"). This also leads to reduced market liquidity, diminished price informativeness, and increased mispricing. Notably, unlike the scenario with price-trigger strategies, in the case of AI collusion through homogenized learning biases, an increase in noise trading risk leads to an increase, rather than a decrease, in trading profitability for AI speculators based on their private information.

***Related Literature.*** The topic of autonomous cooperation among multiple Q-learning agents in repeated games has garnered significant attention from researchers in the artificial intelligence and computer science community over the past decades (e.g., Sandholm and Crites, 1996; Tesauro and Kephart, 2002). Given the widespread adoption of AI technologies in pricing decisions across various marketplaces, Waltman and Kaymak (2008) demonstrate that Q-learning firms typically learn to attain supra-competitive profits in repeated Cournot oligopoly games with homogeneous products, even though a perfect cartel is usually unattainable. Klein (2021) also examines the strategies employed by algorithms in a context where firms selling homogeneous products alternate in adjusting prices to support supra-competitive profits. Recently, in a noteworthy contribution, Calvano et al. (2020) study collusion by AI algorithms in a logit model of differentiated products, not only uncovering the existence of supra-competitive profits but also pinpointing how algorithms might learn to sustain collusive outcomes through grim-trigger strategies. Expanding upon this, our paper extensively broadens the AI experimental framework, moving from a scenario of perfect information and a static demand curve to one imbued with asymmetric information and a strategically-determined demand scheme. We characterize the various types of AI algorithmic collusion, whether occurring through price-trigger strategies or

through learning biases and homogenization, across diverse market environments.

Inspired by the simulation-based studies on AI algorithmic collusion, empirical research has also emerged, demonstrating that the use of AI algorithms in setting product prices can lead to collusion, resulting in heightened supra-competitive prices (e.g., Assad et al., 2023). Additionally, recent studies have started to focus on policy interventions aiming to obstruct the ability of algorithms to collude, thereby ensuring the maintenance of competitive prices. Specially, based on simulation-based studies, Johnson, Rhodes and Wildenbeest (2023) show that platform design can benefit consumers and the platform. However, achieving these gains may require policies that condition on past behavior and treat sellers in a non-neutral fashion. Harrington (2018) delves into critical policy issues surrounding the definition of collusion, such as whether collusion should necessarily entail an explicit agreement among conspirators, or if it might be more aptly defined as the maintenance of elevated prices, sustained by a reward-and-punishment scheme.

Our paper is among the first to investigate how the widespread adoption of AI-powered trading strategies might affect capital markets. The work of Colliard, Foucault and Lovo (2022) is closely related to our research, as it also explores the implications of interactions among Q-learning algorithms in capital markets. However, there are notable differences in focus between their work and ours. Specifically, Colliard, Foucault and Lovo (2022) focuses on AI-powered oligopolistic market makers, while our study concentrates on AI-powered oligopolistic informed speculators who face perfectly competitive market makers. Their research illuminates the strategies that AI market makers would adopt by leveraging their market power. In contrast, our paper explores the dynamics and implications of algorithmic collusion among AI-powered informed speculators, particularly in the context of preferred-habitat long-term investors and perfectly competitive market makers. We provide novel insights into the strategies of informed AI speculators on how they leverage private information and maximize profits through autonomously forming collusion via distinct mechanisms.

## 2 AI-Powered Trading Algorithms

The traditional algorithmic trading system executes orders according to protocols predefined by human quantitative strategists. In contrast, AI-powered trading employs RL algorithms to dynamically adjust and optimize trading strategies in real time.

The RL algorithm, a pivotal technique in AI, forms the foundation of numerous successful AI algorithms, like "AlphaGo," demonstrating the superiority of RL-backed AI over human cognitive abilities in areas such as securities trading and other complex tasks. RL algorithms are model-free machine learning techniques that learn autonomously through trial-and-error experimentation, without relying on two common assumptions: first, that the multi-agent system is on an equilibrium path, and second, that agents have knowledge of the true state and payoff distributions at equilibrium. The fundamental rationale behind RL algorithms centers on the principle that actions yielding higher rewards historically are more likely to be selected in the future, compared to those that have led to lesser rewards. By interacting with its environment and

experimenting with different actions, the agent incrementally learns an optimal policy. Through continuous rounds of exploration and experimentation, it refines its strategy to prefer actions that offer the greatest long-term benefits, even without any knowledge of the environment beforehand. This iterative process enables the agent to progressively enhance its decision-making approach, consistently steering towards actions that maximize the cumulative rewards based on its gathered experiences.

While RL encompasses different 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 numerous RL algorithms, upon which many recent AI breakthroughs are built. However, it is important to note that AI trading algorithms currently in use may not exclusively rely on Q-learning principles. Second, Q-learning holds substantial popularity among computer scientists in practical applications. Third, Q-learning algorithms possess simplicity and transparency, offering clear economic interpretations, in contrast to the black-box nature of many machine learning and AI algorithms. Finally, Q-learning shares a common architecture with more sophisticated RL algorithms.

In the remainder of this section, we will concentrate on a multi-agent system, detailing the Bellman equation for each agent, and describe the Q-learning algorithm that an agent employs. This discussion will cover how each agent iteratively updates its Q-function and strategy based on the received rewards, thereby optimizing its long-term outcomes through the Q-learning algorithm.

### 2.1 Bellman Equation and Q-Function

In a multi-agent Markov decision process environment, there are $I$ agents, indexed by $i = 1, \cdots, I$. The state of the environment is represented by a Markov process, denoted by $s$. Each agent makes decisions based on the current state, which in turn evolves partly due to the collective actions of all agents within the system. Agent $i$'s intertemporal optimization is characterized by the Bellman equation and solved recursively via dynamic programming:

$$V_i(s) = \max_{x_i \in \mathcal{X}} \left\{ \mathbb{E}\left[\pi_i | s, x_i\right] + \rho \mathbb{E}\left[V_i(s') | s, x_i\right] \right\}, \tag{2.1}$$

where $x_i \in \mathcal{X}$ is action of agent $i$, with $\mathcal{X}$ denoting the set of available actions, $\pi_i$ is the payoff received by agent $i$, which may be influenced by the actions of other agents, and $s, s' \in S$ represent the states in the current and the next period, respectively, with $S$ denoting the set of states. In general, $s$ and $s'$ can depend on agent $i$'s individual characteristics and private information. However, for our purpose of illustration, it is sufficient to concentrate on the simple setting where the same state applies uniformly to all agents in the system. The first term on the right-hand side, $\mathbb{E}\left[\pi_i | s, x_i\right]$, is agent $i$'s expected payoff in the current period, and the second term, $\rho \mathbb{E}\left[V_i(s') | s, x_i\right]$, is agent $i$'s continuation value, with the parameter $\rho$ capturing the subjective rate of time preference.

The Bellman equation (2.1) represents the recursive formulation of dynamic control problems (e.g., Bellman, 1954; Ljungqvist and Sargent, 2012). It focuses on the equilibrium path, and thus the optimal value function $V_i(s)$ depends solely on the state variable $s$. In contrast to focusing solely on the equilibrium path, the Q function, denoted by $Q_i(s, x_i)$, extends the optimal value function to include the values of each state-action pair. This captures scenarios (or counterfactuals) that occur off the equilibrium path. By definition, the value of $Q_i(s, x_i)$ is the same as that in the curly brackets of the Bellman equation (2.1):

$$Q_i(s, x_i) = \mathbb{E}\left[\pi_i | s, x_i\right] + \rho \mathbb{E}\left[V_i(s') | s, x_i\right]. \tag{2.2}$$

Intuitively, the Q-function value, $Q_i(s, x_i)$, can be interpreted as the quality of action $x_i$ in state $s$. The optimal value of a state, $V_i(s)$, is the maximum of all the possible Q-function values of state $s$. That is, $V_i(s) \equiv \max_{x' \in \mathcal{X}} Q_i(s, x')$. By substituting $V_i(s')$ with $\max_{x' \in \mathcal{X}} Q_i(s', x')$ in equation (2.2), we can establish a recursive formula for the Q-function as follows:

$$Q_i(s, x_i) = \mathbb{E}\left[\pi_i | s, x_i\right] + \rho \mathbb{E}\left[\max_{x' \in \mathcal{X}} Q_i(s', x') \middle| s, x_i\right]. \tag{2.3}$$

When both $|S|$ and $|\mathcal{X}|$ are finite, the Q-function can be represented as an $|S| \times |\mathcal{X}|$ matrix, which is often referred to as the Q-matrix.

## 2.2 Q-Learning Algorithm

If agent $i$ possessed knowledge of its Q-matrix, determining the optimal actions for any given state $s$ would be straightforward. In essence, the Q-learning algorithm is a method to estimate the Q-matrix in environments where the underlying distribution $\mathbb{E}\left[\cdot | s, x_i\right]$ is unknown and there are limited observations for off-equilibrium pairs $(s, x_i)$ in the data. The Q-learning algorithm addresses both challenges concurrently: it employs Monte Carlo methods to estimate the underlying distribution $\mathbb{E}[\cdot | s, x_i]$ based on the law of large numbers, while at the same time, conducts trial-and-error experiments to produce off-equilibrium counterfactuals.

The iterative experimentation starts from an arbitrary initial Q-matrix of agent $i$, denoted by $\widehat{Q}_{i,0}$, and updates the estimated Q-matrix $\widehat{Q}_{i,t}$ recursively. The learning equation governing this update is as follows:

$$\widehat{Q}_{i,t+1}(s_t, x_{i,t}) = (1-\alpha) \underbrace{\widehat{Q}_{i,t}(s_t, x_{i,t})}_{\text{Past knowledge}} + \alpha \underbrace{\left[\pi_{i,t} + \rho \max_{x' \in \mathcal{X}} \widehat{Q}_{i,t}(s_{t+1}, x')\right]}_{\text{Present learning based on a new experiment}}, \tag{2.4}$$

where $\alpha \in [0, 1]$ captures the forgetting rate, $s_t$ is the state that the iteration $t$ concentrates on, $s_{t+1}$ is randomly drawn from the Markovian transition probabilities conditional on $s_t$, $\widehat{Q}_{i,t}(s, x)$ is the estimated Q-matrix of agent $i$ in the $t$-th iteration, and $\pi_{i,t}$ is the payoff in the $t$-th iteration, corresponding to agent $i$'s choice of action $x_{i,t}$.

Equation (2.4) indicates that for agent $i$ in the $t$-th iteration, only the value of the estimated

Q-matrix $\widehat{Q}_{i,t}(s,x)$ corresponding to the state-action pair $(s_t, x_{i,t})$ is updated to $\widehat{Q}_{i,t+1}(s_t, x_{i,t})$. All other state-action pairs remain unchanged. In other words, $\widehat{Q}_{i,t+1}(s,x) = \widehat{Q}_{i,t}(s,x)$ for cases where $s \neq s_t$ or $x \neq x_{i,t}$. The updated value $\widehat{Q}_{i,t+1}(s_t, x_{i,t})$ is computed as a weighted average of accumulated knowledge based on the previous experiments, $\widehat{Q}_{i,t}(s_t, x_{i,t})$, and learning based on a new experiment, $\pi_{i,t} + \rho \max_{x' \in \mathcal{X}} \widehat{Q}_{i,t}(s_{t+1}, x')$. A key distinction between the Q-learning recursive algorithm (2.4) and the Bellman recursive equation (2.1) lies in how they handle expectations. Q-learning algorithm (2.4) does not form expectations about the continuation value because the Markovian transition probabilities from $s_t$ to $s_{t+1}$ are unknown. Instead, it directly discounts the continuation value associated with the randomly realized state $s_{t+1}$ in the $(t+1)$-th iteration.

It is crucial to note that the forgetting rate $\alpha$ plays a significant role in the Q-learning algorithm, balancing past knowledge against present learning based on a new experiment. A higher $\alpha$ not only indicates a greater impact of present learning on the Q-value update but also implies that the algorithm forgets past knowledge more quickly, potentially leading to biased learning. To elaborate, let $\tau$ be the number of times that the Q-value of the state-action pair $(s,x)$ has been updated in the past. We derive in Appendix G.1 that as $\tau \to \infty$, the Q-value of $(s,x)$ is as follows:

$$\widehat{Q}_{i,t(\tau)}(s,x) \approx \sum_{h=0}^{\tau-1} \alpha(1-\alpha)^h \left[ \pi_{i,t(\tau-h)} + \rho \max_{x' \in \mathcal{X}} \widehat{Q}_{i,t(\tau-h)}(s_{t(\tau-h)+1}, x') \right], \tag{2.5}$$

where $t(h)$ represents the period in which the Q-value of $(s,x)$ receives the $h$-th update. Clearly, when $\alpha$ is not close to 0, the weights given by $\alpha(1-\alpha)^h$ decay so rapidly with $\tau$ that it jeopardizes the applicability of the law of large number. When the underlying environment has randomness, a sufficiently small value of $\alpha$ is crucial for ensuring small learning biases. Otherwise, the law of large numbers may fail, leading to biased estimation for the underlying distribution $\mathbb{E}[\cdot|s, x_i]$. However, a smaller value of $\alpha$ requires more iterations for the algorithm to converge, and thus greater computational costs. Moreover, if $\alpha$ is excessively small relative to the decaying speed of the exploration rate $\varepsilon_t$ in equation (2.6), biased learning may arise due to insufficient exploration.

### 2.3 Experimentation

Conditional on the state variable $s_t$, agent $i$ chooses its action $x_{i,t}$ in two experimentation modes, exploitation and exploration, as follows:

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

To determine the mode, we employ the simple $\varepsilon$-greedy method. As outlined in equation (2.6), during the $t$-th iteration, agent $i$ engages in the exploration and exploitation modes with exogenous probabilities $\varepsilon_t$ and $1-\varepsilon_t$, respectively. In the exploitation mode, agent $i$ chooses its action to maximize the current state's Q-value, given by $x_{i,t} = \operatorname{argmax}_{x \in \mathcal{X}} \widehat{Q}_{i,t}(s_t, x)$. Conversely, in the exploration mode, agent $i$ randomly chooses its action $\widetilde{x}$ from the set of all possible values in $\mathcal{X}$,

each with equal probability.[4] Essentially, the exploration mode guides the Q-learning algorithm to experiment with suboptimal actions based on the current Q-matrix approximation, $\widehat{Q}_{i,t}$. As $t$ approaches infinity, the pre-specified exploration probability $\varepsilon_t$ monotonically decreases to zero.

Given that agent $i$ lacks prior knowledge about its Q-matrix, it is evident that sufficient exploration is crucial to increase the accuracy of approximating the true Q-matrix. At a minimum, all actions must be attempted multiple times in all states, and even more so in complex environments. However, in addition to the computational costs associated with exploration, there exists a tradeoff. An overly comprehensive exploration scheme may have adverse effects when multiple agents interact with one another, because the random selected actions by one agent introduce noises to other agents, impeding their learning processes.

# 3 Model

This model extends the influential framework of Kyle (1985) along three novel dimensions. First, it considers multiple informed speculators in a repeated-game context. Second, it introduces a representative preferred-habitat investor, whose net demand flows need to be absorbed by other agents in the market (e.g., Vayanos and Vila, 2021). Third, it introduces a market maker who takes into account both inventory and pricing error, going beyond the limited focus on price error alone as in the model of Kyle (1985).

By blending theoretical rigor with practical relevance, this model offers a laboratory for exploring the implications of AI-powered trading on both algorithmic collusion and price efficiency. Importantly, the theoretical results produced by the model act as a foundational benchmark for the characterization and categorization of AI-powered trading in simulation experiments in Sections 4 to 6.

## 3.1 Economic Environment

Time is discrete, indexed by $t = 1, 2, ...$, and runs forever. There are $I \geq 2$ risk-neutral informed speculators, a representative noise trader, a representative preferred-habitat investor, and a market maker. The economic environment is stationary, and all exogenous shocks are independent and identically distributed across periods.

In each period $t$, an asset is available for trading, with its fundamental value, denoted as $v_t$, being realized at the end of the period. Each period consists of two distinct steps: the beginning and the end. We examine the problem in period $t$ in reverse order. At the end of the period, $v_t$ is observed by all agents. It is drawn from a normal distribution $N(\overline{v}, \sigma_v^2)$, where $\sigma_v^2$ represents the variance and $\overline{v}$ the mean, with $\overline{v} \equiv 1$ for convenience. After the realization of $v_t$, trading profits for all agents in period $t$ are determined.

[4]For simplicity, we adopt a uniform distribution. However, a more intelligent distribution choice could make exploration more efficient and less costly.

At the beginning of the period, the informed speculators, noise trader, and preferred-habitat investor submit their order flows. Simultaneously, the market maker sets the asset's price, denoted as $p_t$. Specifically, the noise trader submits its order flow $u_t$ to either buy $u_t$ units of the asset if $u_t > 0$ or take a short position of $u_t$ if $u_t < 0$, with $u_t$ following a normal distribution $N(0, \sigma_u^2)$, where zero is the mean and $\sigma_u^2$ is the variance. The informed speculators are indexed by $i \in \{1, \cdots, I\}$. Each informed speculator $i$ perfectly knows the value $v_t$, but is unaware of $u_t$ when submitting his order flows; he understands that the choice of order flow $x_{i,t}$ will influence $p_t$ by shifting the market-clearing condition and revealing information. The informed speculator $i$ chooses its order flows $\{x_{i,t}\}_{t\geq 0}$ to maximize the expected present value of the profit stream:

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \rho^t (v_t - p_t) x_{i,t}\right], \tag{3.1}$$

where $\rho \in (0,1)$ is the subjective discount rate.

***Preferred-Habitat Investor's Demand Curve.*** Contrary to the uninformed speculator in Kyle (1989), the preferred-habitat investor does not derive information about $v_t$ from $p_t$. Instead, this investor has a linear downward-sloping demand curve for the net trading flow $z_t$:

$$z_t = -\xi(p_t - \overline{v}), \quad \text{with } \xi > 0. \tag{3.2}$$

The rationale behind this specification is straightforward: the preferred-habitat investor focuses solely on the ex-ante expected fundamental value, $\overline{v}$, and tends to buy more of the asset when $p_t - \overline{v}$ is more negative, interpreting this as a stronger indication that the asset is currently undervalued. The demand curve is proportional to the spread between the ex-ante expected fundamental value and the market price. Graham (1973) names this spread a safety margin.

The average asset holding of the preferred-habitat investor, denoted as $\overline{z}$, is often substantial. This implies a small price elasticity of demand, given by $\varepsilon = \mathbb{E}[(\mathrm{d}z_t/\mathrm{d}p_t)(p_t/z_t)] = -\xi\mathbb{E}[p_t/z_t] \approx -\xi/\overline{z}$. Studies indicate that preferred-habitat investors with low price elasticity of demand play an important role in shaping asset prices (e.g., Greenwood and Vayanos, 2014; Vayanos and Vila, 2021; Greenwood et al., 2023).

The preferred-habitat investor's demand curve (3.2) mirrors that of the "long-term investor" in the model by Kyle and Xiong (2001). This becomes clear, especially when we recognize that $\overline{v}$ is the fair value of the asset to risk neutral investors as $\overline{v} = \mathbb{E}[v_t]$. According to this demand curve, the preferred-habitat investor always provides liquidity to the market. When the price falls further below the ex-ante expected fundamental value, $\overline{v}$, in the market, the preferred-habitat investor will buy more of the asset. Analogous to Kyle and Xiong (2001), the demand curve (3.2) can be justified by a rational choice made by the preferred-habitat investor under certain assumptions. These assumptions are summarized in Lemma 1. The proof is in Appendix A.

**Lemma 1** (Demand Curve)**.** *If the preferred-habitat investor possesses exponential utility with an absolute*

*risk aversion coefficient of $\eta$, then the demand curve has the functional form of (3.2), where the slope $\xi$ is given by $1/(\eta\sigma_v^2)$.*

Moreover, the concept of specifying exogenous net demand curves within the framework of a noisy rational expectation equilibrium also shares similarities with studies conducted by Hellwig, Mukherji and Tsyvinski (2006) and Goldstein, Ozdenoren and Yuan (2013), among others. The fundamental idea is to capture relevant institutional frictions and preferences in a parsimonious and tractable manner. Notably, our net demand curves can be reinterpreted as "noisy supply curves" in these prior works by introducing a new variable $\widetilde{z}_t \equiv -(u_t + z_t)$. Specifically, $\widetilde{z}_t$ represents the total trading supply provided by the noisy trader and the preferred-habitat investor to absorb the trading demand of informed speculators. The total supply $\widetilde{z}_t$ follows an exogenous noisy supply curve defined as:

$$\widetilde{z}_t = -u_t + \xi(p_t - \overline{v}), \tag{3.3}$$

where $-u_t$ can be reinterpreted as the unobservable demand or supply shock in the context of the above prior works.

***Market Maker's Pricing Rules.*** Trading occurs through the market maker, whose role is to absorb the order flow while minimizing pricing errors. The market maker observes the combined order flow of informed speculators and the noise trader, represented by $y_t = \sum_{i=1}^{I} x_{i,t} + u_t$, as well as the order flow $z_t$ of the preferred-habitat investor. However, the market maker cannot distinguish between order flows from informed speculators and the noise trader. Thus, the market maker can only make statistical inferences about the fundamental value $v_t$ based on the combined order flow $y_t$ rather than individual order flows. The market maker sets the price $p_t$ to jointly minimize inventory and pricing errors according to the following objective function:

$$\min_{p_t} \mathbb{E}\left[(y_t + z_t)^2 + \theta(p_t - v_t)^2 \middle| y_t\right], \tag{3.4}$$

where $\theta > 0$ represents the weight that the market maker places on minimizing pricing errors. Here, $\mathbb{E}[\cdot|y_t]$ denotes the market maker's expectation over $v_t$, conditioned on the observed combined order flow $y_t$ and its belief about how informed speculators would behave in the equilibrium.

The market maker's objective function (3.4) captures both the inventory cost and asymmetric information faced by the market maker. Because the market maker takes the position $-(y_t + z_t)$ to clear the market, the term $(y_t + z_t)^2$ represents its inventory-holding costs. The quadratic form is adopted for tractability, consistent with the literature (e.g., Mildenstein and Schleef, 1983). The term $\theta(p_t - v_t)^2$ captures the market maker's efforts to reduce pricing errors arising from asymmetric information. The weight $\theta$ serves as a reduced-form way to capture the various benefits of reducing pricing errors, such as increased trading flows from a growing client base or enhanced competitive advantages over other trading platforms.[5] As $\theta$ approaches zero, the price

[5]Similarly, in the context of e-commerce platforms, it is often assumed that the platform aims to maximize a weighted

$p_t$ is primarily determined by the market clearing condition, $y_t + z_t = 0$, as in the model of Kyle and Xiong (2001). Conversely, as $\theta$ increases towards infinity, the price $p_t$ is primarily determined by the pricing-error minimization condition, $p_t = \mathbb{E}[v_t|y_t]$, as in the model of Kyle (1985).

Because multiple informed speculators engage in a repeated-game of trading in our model, multiple equilibria may emerge. We identify three types of equilibria: the non-collusive equilibrium, the perfect cartel equilibrium, and the collusive equilibrium sustained by price-trigger strategies. Throughout our analysis, we assume that the market maker is aware of the specific equilibrium in which informed speculators are participating. Specifically, we consider the linear and symmetric equilibrium in which the trading strategy of the informed speculators is characterized by

$$x_{i,t} = \chi(v_t - \overline{v}), \quad \text{for all } i = 1, \cdots, I. \tag{3.5}$$

The first-order condition of the minimization problem (3.4) leads to

$$p_t = \frac{\xi}{\xi^2 + \theta} y_t + \frac{\xi^2}{\xi^2 + \theta} \overline{v} + \frac{\theta}{\xi^2 + \theta} \mathbb{E}[v_t|y_t],$$

where $\mathbb{E}[v_t|y_t]$, according to Bayesian updating, is

$$\mathbb{E}[v_t|y_t] = \overline{v} + \gamma y_t, \quad \text{with } \gamma = \frac{I\chi}{(I\chi)^2 + \sigma_u^2/\sigma_v^2}.$$

Therefore, the market maker's pricing rule is

$$p_t = \overline{v} + \lambda y_t, \quad \text{with } \lambda = \frac{\theta\gamma + \xi}{\theta + \xi^2}.$$

### 3.2 Noncollusive Nash Equilibrium

We use the superscript $N$ to denote the variables in the noncollusive Nash equilibrium. At the beginning of each period $t$, each informed speculator $i$ solves the following problem:

$$x^N(v_t) = \underset{x_i}{\operatorname{argmax}} \mathbb{E}\left[(v_t - p_t)\, x_i \middle| v_t\right], \tag{3.6}$$

where $\mathbb{E}[\cdot|v_t]$ is informed investor $i$'s expectation conditional on the privately observed $v_t$ and its belief about how the market maker would set the price in the equilibrium $p_t = p^N(y_t)$. The pricing function $p^N(\cdot)$ is determined in equilibrium, as follows:

$$p^N(y_t) = \overline{v} + \lambda^N y_t, \quad \text{with } \lambda^N = \frac{\theta\gamma^N + \xi}{\theta + \xi^2} \text{ and } \gamma^N = \frac{I\chi^N}{(I\chi^N)^2 + (\sigma_u/\sigma_v)^2}, \tag{3.7}$$

average of per-unit fee revenues and consumer surplus (see, e.g., Johnson, Rhodes and Wildenbeest, 2023). The weight on consumer surplus in this context is a reduced-form way to capture various aspects of increasing consumer surplus. For example, increasing consumer surplus allows the platform to dynamically expand its consumer base over time and better compete with rival platforms.

where $y_t$ is the combined order flow of informed speculators and the noise trader, given by

$$y_t = x_i + (I-1)x^N(v_t) + u_t. \tag{3.8}$$

The non-collusive Nash equilibrium can be summarized in the following proposition.

**Proposition 3.1.** *The order flow of informed speculators and price in the non-collusive Nash equilibrium are*

$$x^N(v_t) = \chi^N(v_t - \overline{v}) \;\; \textit{and} \;\; p^N(v_t) = \overline{v} + \lambda^N y_t, \;\; \textit{respectively,}$$

*where $\chi^N$ and $\lambda^N$ satisfy*

$$\chi^N = \frac{1}{(I+1)\lambda^N} \quad \textit{and} \quad \lambda^N = \frac{\theta\gamma^N + \xi}{\theta + \xi^2} \;\; \textit{with} \;\; \gamma^N = \frac{I\chi^N}{(I\chi^N)^2 + (\sigma_u/\sigma_v)^2}.$$

*The expected profit of informed speculators is*

$$\pi^N = \left(1 - \lambda^N I\chi^N\right)\chi^N\sigma_v^2.$$

*The price informativeness, denoted by $\mathcal{I}^N$, is defined as the logged signal-noise ratio of prices,*

$$\mathcal{I}^N = \log\left[\frac{\mathrm{var}(x_{i,t}^N)}{\mathrm{var}(u_t)}\right] = \log\left[\left(I\chi^N\right)^2(\sigma_v/\sigma_u)^2\right].$$

*The market liquidity, denoted by $\mathcal{L}^N$, is defined as the inverse sensitivity of the market maker's inventory $|z_t + y_t|$ to the noise order flow $u_t$*

$$\mathcal{L}^N = \frac{1}{\partial|z_t + y_t|/\partial u_t} = \frac{1}{|1 - \xi\lambda^N|}.$$

*The mispricing, denoted by $\mathcal{E}^N$, is defined by the percentage deviation of the asset's price $p_t$ from its conditional expected value*

$$\mathcal{E}^N = \left|\frac{p^N(v_t) - \mathbb{E}^N[v_t|y_t]}{\mathbb{E}^N[v_t|y_t] - \overline{v}}\right| = \left|\frac{\lambda^N - \gamma^N}{\gamma^N}\right|.$$

Intuitively, the price informativeness measure captures the fact that relative to the noise trader, informed speculators' order flows contain information about the asset's value $v_t$. Thus, the order flow of informed speculators can be considered as informative signals about the value of $v_t$ whereas noise order flows contain no information. The market liquidity measure captures the fact that when the market is less liquid, trade flows can have a larger impact on the market maker's inventory, leading to greater adjustments in the asset's price.

### 3.3 Perfect Cartel Equilibrium

Consider a cartel that consists all $I$ informed speculators under perfect collusion. The cartel is a monopolist who chooses each informed speculator's order flow to maximize total profits. Because informed speculators are symmetric, the cartel solves the following problem

$$x^M(v_t) = \underset{x_i}{\operatorname{argmax}}\, \mathbb{E}\left[(v_t - p_t)\, x_i \middle| v_t\right], \tag{3.9}$$

where $\mathbb{E}[\cdot|v_t]$ is informed investor $i$'s expectation conditional on the privately observed $v_t$ and its belief about how the market maker would set the price in the equilibrium $p_t = p^M(y_t)$. The pricing function $p^M(\cdot)$ s determined in equilibrium, as follows:

$$p^M(y_t) = \overline{v} + \lambda^M y_t, \;\; \text{with} \;\; \lambda^M = \frac{\theta\gamma^M + \xi}{\theta + \xi^2} \;\; \text{and} \;\; \gamma^M = \frac{I\chi^M}{(I\chi^M)^2 + (\sigma_u/\sigma_v)^2}, \tag{3.10}$$

where $y_t$ is the combined order flow of informed speculators and the noise trader, given by

$$y_t = Ix_i + u_t. \tag{3.11}$$

The perfect cartel equilibrium can be summarized in the following proposition.

**Proposition 3.2.** *The order flow of informed speculators and price in the perfect cartel equilibrium are*

$$x^M(v_t) = \chi^M(v_t - \overline{v}) \;\; \textit{and} \;\; p^M(v_t) = \overline{v} + \lambda^M y_t, \;\; \textit{respectively,}$$

*where $\chi^M$ and $\lambda^M$ satisfy*

$$\chi^M = \frac{1}{2I\lambda^M} \quad \textit{and} \quad \lambda^M = \frac{\theta\gamma^M + \xi}{\theta + \xi^2} \;\; \textit{with} \;\; \gamma^M = \frac{I\chi^M}{(I\chi^M)^2 + (\sigma_u/\sigma_v)^2}.$$

*The expected profit of informed speculators is*

$$\pi^M = \left(1 - \lambda^M I\chi^M\right)\chi^M\sigma_v^2.$$

*The price informativeness, denoted by $\mathcal{I}^M$, is defined as the logged signal-noise ratio of prices,*

$$\mathcal{I}^M = \log\left[\frac{\operatorname{var}(x_{i,t}^M)}{\operatorname{var}(u_t)}\right] = \log\left[\left(I\chi^M\right)^2(\sigma_v/\sigma_u)^2\right].$$

*The market liquidity, denoted by $\mathcal{L}^M$, is defined as the inverse sensitivity of the market maker's inventory $|z_t + y_t|$ to the noise order flow $u_t$*

$$\mathcal{L}^M = \frac{1}{\partial|z_t + y_t|/\partial u_t} = \frac{1}{|1 - \xi\lambda^M|}.$$

*The mispricing, denoted by $\mathcal{E}^M$, is defined by the percentage deviation of the asset's price $p_t$ from its conditional expected value*

$$\mathcal{E}^M = \left| \frac{p^M(v_t) - \mathbb{E}^M[v_t|y_t]}{\mathbb{E}^M[v_t|y_t] - \overline{v}} \right| = \left| \frac{\lambda^M - \gamma^M}{\gamma^M} \right|.$$

### 3.4 Collusive Nash Equilibrium

Information asymmetry is a significant characteristic of capital markets, rendering standard grim-trigger strategies less viable to sustain tacit collusion, due to the challenges in accurately observing and monitoring each other's actions.[6] However, tacit collusion can still be sustained under information asymmetry through price-trigger strategies with imperfect monitoring. If an informed speculator can reliably infer other informed speculators' total order flows from the market price, collusive incentives can be created.

The concept of tacit collusion sustained by price-trigger strategies was first introduced by Green and Porter (1984). Even with imperfect monitoring, agents can establish collusive incentives by allowing noncollusive competition to occur with positive probabilities. Abreu, Pearce and Stacchetti (1986) further characterize optimal symmetric equilibria in this context, revealing two extreme regimes: a collusive regime and a punishment regime featuring a noncollusive reversion. In the collusive regime, informed speculators implicitly coordinate on submitting order flows in a less aggressive manner than what they would do in the noncollusive Nash equilibrium. If the price breaches a critical level, suspicion of cheating arises, leading to a noncollusion reversion. In the punishment regime, informed speculators trade noncollusively and obtain low profits.

***Price-Trigger Strategies.*** We now describe the collusive Nash equilibrium sustained by price-trigger strategies under information asymmetry, as studied by Green and Porter (1984). Specifically, we focus on the symmetric collusive Nash equilibrium in which all $I$ informed speculators choose the same collusive order flow, denoted by $x^C(v_t)$. Such trading strategies are sustained by a price-trigger strategy: Firms will initially submit their respective order flows $x^C(v_t)$, and will continue to do so until the market price falls below a trigger price $q(v_t)$ if $v_t < \overline{v}$ or goes above a trigger price $q(v_t)$ if $v_t > \overline{v}$, and then they will trade noncollusively for a reversionary episode that lasts for $T-1$ periods. In period $t$, the state of world is "normal," denoted by $s_t = 0$, if (a) $v_{t-1} = \overline{v}$ and $s_{t-1} = 0$, or (b) $p_{t-1} \leq q(v_{t-1})$ and $v_{t-1} > \overline{v}$ and $s_{t-1} = 0$, or (c) $p_{t-1} \geq q(v_{t-1})$ and $v_{t-1} < \overline{v}$ and $s_{t-1} = 0$, or (d) $p_{t-T} > q(v_{t-T})$ and $v_{t-T} > \overline{v}$ and $s_{t-T} = 0$, or (e) $p_{t-T} \leq q(v_{t-T})$ and $v_{t-T} < \overline{v}$ and $s_{t-T} = 0$. Otherwise, in period $t$, the state of world is "reversionary," denoted by $s_t = 1$. In other words, $s_t = 0$ if price trigger is not violated at $t-1$ and $s_{t-1} = 0$, or if price trigger is violated at $t-T$ and $s_{t-T} = 0$; otherwise, $s_t = 1$.

[6] Tacit collusion sustained by grim-trigger strategies has been extensively studied since the pioneering work of Fudenberg and Maskin (1986) and Rotemberg and Saloner (1986), among others. Recent studies delve into the impact of such tacit collusion sustained by grim-trigger strategies on pricing in capital markets (e.g., Opp, Parlour and Walden, 2014; Dou, Ji and Wu, 2021*a*,*b*; Dou, Wang and Wang, 2023).

Similar to Green and Porter (1984), we assume that the state variable $s_t$ is a common knowledge to all agents. We characterize the equilibrium order flows and prices in each period $t$. There are two cases: when $s_t = 1$, the state of world is reversionary, and thus the equilibrium order flows and prices follow the noncollusive equilibrium in Section 3.2; and when $s_t = 0$, the state of world is normal. In this case, we focus on linear policy functions and characterize the equilibrium order flow $x^C(v_t)$ and price $p_t^C$ as follows:

$$x^C(v) \equiv \chi^C(v - \overline{v}), \tag{3.12}$$

$$p^C(y) = \overline{v} + \lambda^C y, \text{ with } \lambda^C = \frac{\theta\gamma^C + \xi}{\theta + \xi^2} \text{ and } \gamma^C = \frac{I\chi^C}{(I\chi^C)^2 + \sigma_u^2/\sigma_v^2}. \tag{3.13}$$

The price-trigger function $q(v)$ is specified based on the expected price when all informed speculators trade coordinately according to $x^C(v)$ conditional on $v$, namely, $\overline{p}^C(v) \equiv \mathbb{E}\left[p^C(y)|v\right]$. Specifically, plugging (3.12) into (3.13) and taking expectation over $u$, we obtain that $\overline{p}^C(v) \equiv \overline{v} + \lambda^C I\chi^C(v - \overline{v})$. The price-trigger function $q(v)$ is specified as follows:

$$q(v) \equiv \begin{cases} \overline{p}^C(v) + \lambda^C\sigma_u\omega, & \text{if } v > \overline{v} \\ \overline{p}^C(v) - \lambda^C\sigma_u\omega, & \text{if } v < \overline{v}, \end{cases} \tag{3.14}$$

where $\omega > 0$ is a parameter that characterizes the tightness of the price trigger.

Equation (3.14) warrants further discussions. First, when $v > \overline{v}$, informed investors have incentives to buy a large amount of the asset, which boosts up its price. As a result, when $v > \overline{v}$, a meaningful price-trigger strategy would punish the potential deviating counterparty by reverting to the noncollusive Nash equilibrium once the market price goes above a certain high-level threshold $q(v)$. In contrast, when $v < \overline{v}$, informed investors have incentives to sell a large amount of the asset, which suppresses down its price. As a result, when $v < \overline{v}$, a meaningful price-trigger strategy would punish the potential deviating counterparty by reverting to the noncollusive Nash equilibrium once the market price falls below a certain low-level threshold $q(v)$. Second, there is no price threshold when $v = \overline{v}$ because no informed investor would have incentives to trade in this case. Third, although there are infinitely many alternative ways to specify the functional form of the threshold $q(v)$, we focus on a specification that is not only statistically meaningful but also ensures a linear model solution as in Kyle (1985). If no one deviates from the coordinated trading, each informed speculator can infer that the noise order is $\widehat{u}_t = [p_t - q(v_t)]/\lambda^C$ based on the observed price $p_t = p^C(y_t)$. If $\widehat{u}_t$ is excessively positive when $v_t > \overline{v}$, say $\widehat{u}_t > \omega\sigma_u$ for some constant $\omega > 0$, the informed speculator would suspect that some other informed speculators might have deviated from the implicit agreement. Analogously, if $\widehat{u}_t$ is excessively negative when $v_t < \overline{v}$, say $\widehat{u}_t < -\omega\sigma_u$ for some constant $\omega > 0$, the informed speculator would suspect that some other informed speculators might have deviated from the implicit agreement. Fourth, the multiplier $\sigma_u$ ensures that the probability of price-trigger violation is independent of the magnitude of noisy trading, $\sigma_u$, in the collusive Nash equilibrium.

Given that $s_t = 0$, let $J^C(\chi_i)$ denote each informed speculator $i$'s expected present value of

future profits, when investor $i$ chooses $x_{i,t} = \chi_i(v_t - \overline{v})$ and all other $I-1$ informed investors choose $x^C(v_t)$. The value of $J^C(\chi_i)$ is determined recursively as follows:

$$\begin{aligned} J^C(\chi_i) =& \mathbb{E}\left[\left(v_t - p^C(y_t)\right)\chi_i(v_t - \overline{v})\right] \\ &+ \rho J^C(\chi_i)\mathbb{P}\left\{\text{Price trigger is not violated in period } t \middle| \chi_i, \chi^C\right\} \\ &+ \mathbb{E}\left[\sum_{\tau=1}^{T-1}\rho^\tau \pi^N(v_{t+\tau}) + \rho^T J^C(\chi_i)\right]\mathbb{P}\left\{\text{Price trigger is violated in period } t \middle| \chi_i, \chi^C\right\}, \end{aligned} \tag{3.15}$$

where the combined order flow of informed investors and the noise trader is

$$y_t = \chi_i(v_t - \overline{v}) + (I-1)x^C(v_t) + u_t, \tag{3.16}$$

and the probability of price-trigger violation is

$$\begin{aligned} &\mathbb{P}\left\{\text{Price trigger is not violated in period } t \middle| \chi_i, \chi^C\right\} \\ =&\mathbb{E}\left[\mathbb{P}\left(p_t \le q(v_t)|v_t\right)\mathbf{1}\{v_t > \overline{v}\}\right] + \mathbb{E}\left[\mathbb{P}\left(p_t \ge q(v_t)|v_t\right)\mathbf{1}\{v_t < \overline{v}\}\right] \\ =&\mathbb{E}\left[\Phi(\sigma_u^{-1}(\chi^C - \chi_i)(v_t - \overline{v}) + \omega)\mathbf{1}\{v_t > \overline{v}\}\right] + \mathbb{E}\left[\Phi(\sigma_u^{-1}(\chi_i - \chi^C)(v_t - \overline{v}) + \omega)\mathbf{1}\{v_t < \overline{v}\}\right], \end{aligned}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution.

***Impossibility of Collusion When Efficient Prices Prevail.*** The following proposition highlights the impossibility of achieving collusion in an environment closely resembling the standard Kyle benchmark (Kyle, 1985), where efficient prices prevail. In this setting, the market maker focuses on minimizing pricing errors and sets the price approximately at $\mathbb{E}[v_t|y_t]$, which is the expected fundamental value conditional on the observed combined order flow of informed speculators and the noise trader. In other words, the efficient price in this context is an unbiased estimate of the asset's fundamental value. The proof is in Appendix B.

**Proposition 3.3** (Impossibility of Collusion When Efficient Prices Prevail)**.** *If $\theta$ is large or $\xi$ is small, there is no collusive Nash equilibrium that can be sustained by price-trigger strategies for any $\sigma_u/\sigma_v > 0$.*

Sustaining coordination through price-trigger strategies requires two conditions: (i) price informativeness needs to be sufficiently high to ensure that there is sufficient capacity for monitoring, which has been emphasized by Abreu, Milgrom and Pearce (1991) and Sannikov and Skrzypacz (2007), and (ii) the price impact of informed speculators' order flows needs to be sufficiently low to ensure that there is sufficient room for achieving significant informational rents.

However, the environments with large $\theta$ or small $\xi$ closely resemble the standard Kyle benchmark (Kyle, 1985), where efficient prices prevail. In this environment, because $\lambda^C$ is

approximately equal to $\gamma^C$, price informativeness is always low and unresponsive to $\sigma_u/\sigma_v$.[7] As a result, the two necessary conditions (i) and (ii) cannot hold simultaneously. In particular, in order to achieve high price informativeness, the environment needs to have low noise trading risks, as captured by a low $\sigma_u/\sigma_v$. However, knowing that noise orders are not significant, the market maker will choose a high $\gamma^C$, resulting in a high price impact of informed trading because $\lambda^C \approx \gamma^C$. The high price impact of informed trading would further induce informed speculators to trade conservatively by placing orders of small amounts. In the end, the positive effect on price informativeness from low noise trading risks would be largely cancelled out by the negative effect from the conservative orders of informed speculators, making the price informativeness low and unresponsive to $\sigma_u/\sigma_v$.

Proposition 3.3 carries intrinsic value in terms of theoretical insights and novelty, setting it apart from existing theories on the impossibility of collusion under information asymmetry, as posited by Abreu, Milgrom and Pearce (1991) and Sannikov and Skrzypacz (2007). These prior theories emphasize that, when prices are not informative, "false positive" errors, made by triggering punishments, occur on the equilibrium path disproportionately often, erasing all benefits from collusion. In contrast, Proposition 3.3 offers a distinctive intuitive perspective, highlighting that informed speculators cannot exploit pricing errors to achieve collusive outcomes because prices are already efficient, accurately reflecting the asset's fundamental value. The absence of substantial pricing errors essentially renders collusion infeasible, as there exists limited scope for market manipulation based on price discrepancies. In summary, Proposition 3.3 sheds light on the interplay between efficient pricing, information asymmetry, and collusive behavior in financial markets. By demonstrating the impracticality of collusion in environments characterized by efficient prices, our results provide a deeper understanding of market dynamics and the implications of information asymmetry on collusion strategies.

***Existence of Collusion with a Significant Preferred-Habitat Investor.*** The following proposition shows that collusion sustained by price-trigger strategies exists when the preferred-habitat investor plays an important role in price formation, making prices not very efficient. However, when information asymmetry, captured by $\sigma_u/\sigma_v$, is too large, or when the number of informed speculators $I$, no collusion can be sustained by price-trigger strategies even with inefficient prices. The proof is in Appendix C.

**Proposition 3.4** (Existence of Collusion with a Significant Preferred-Habitat Investor)**.** *If $\theta$ is sufficiently small or $\xi$ is sufficiently large, there exists a collusive Nash equilibrium that can be sustained by price-trigger strategies provided that $\sigma_u/\sigma_v$ and $I$ are not too large.*

If $\theta$ is small or $\xi$ is large, the market maker determines prices determined primarily to minimize inventory costs rather than pricing errors. Thus, a low price impact of informed trading can arise even in environments with low noise trading risks. The low price impact of informed trading would further induce informed speculators to trade aggressively by placing orders of large

[7] In the extreme case with $\theta = \infty$ or $\xi = 0$, price informativeness is independent from $\sigma_u/\sigma_v$ as in Kyle (1985).

amounts, thereby leading to high price informativeness. Consequently, the necessary conditions (i) and (ii) can hold simultaneously when the preferred-habitat investor plays an important role in price formation.

However, when $\sigma_u/\sigma_v$ is too large, price informativeness is low, and thus price-trigger strategies are difficult to sustain. This is because when prices are not informative, agents make "false positive" errors by triggering punishments on the equilibrium path disproportionately often, erasing all benefits from collusion. This key idea exactly follows the insight of Abreu, Milgrom and Pearce (1991) and Sannikov and Skrzypacz (2007).

***Properties of Collusion Sustained by Price-Trigger Strategies.*** To discern whether informed speculators trade in a tacitly collusive manner based on observable outcomes, we derive testable properties of collusion.

**Proposition 3.5** (Supra-competitive nature of collusion)**.** *In the price-trigger collusive equilibrium, it holds that*

$$\pi^M \geq \pi^C > \pi^N, \tag{3.17}$$

*where* $\pi^C = \left(1 - \lambda^C I \chi^C\right) \chi^C \sigma_v^2$ *is the expected profit of informed speculators in the collusive equilibrium. If we define* $\Delta^C \equiv \frac{\pi^C - \pi^N}{\pi^M - \pi^N}$*, inequalities in (3.17) can be summarized by* $\Delta^C \in (0, 1]$*.*

Clearly, a greater $\Delta^C$ signifies a higher collusion capacity. We use $\Delta^C$ as a measure for collusion capacity, as in Calvano et al. (2020). Similar measures are also adopted in empirical studies to identify collusion capacity (e.g., Dou, Wang and Wang, 2023). Consistent with the definitions in the noncollusive equilibrium and the perfect cartel equilibrium, the price informativeness, denoted by $\mathcal{I}^C$, is defined as the logged signal-noise ratio of prices,

$$\mathcal{I}^C = \log\left[\frac{\mathrm{var}(x_{i,t}^C)}{\mathrm{var}(u_t)}\right] = \log\left[\left(I\chi^C\right)^2 (\sigma_v/\sigma_u)^2\right].$$

The market liquidity, denoted by $\mathcal{L}^C$, is defined as the inverse sensitivity of the market maker's inventory $|z_t + y_t|$ to the noise order flow $u_t$

$$\mathcal{L}^C = \frac{1}{\partial |z_t + y_t| / \partial u_t} = \frac{1}{|1 - \xi\lambda^C|}.$$

The mispricing, denoted by $\mathcal{E}^C$, is defined by the percentage deviation of the asset's price $p_t$ from its conditional expected value

$$\mathcal{E}^C = \left|\frac{p^C(v_t) - \mathbb{E}^C[v_t|y_t]}{\mathbb{E}^C[v_t|y_t] - \overline{v}}\right| = \left|\frac{\lambda^C - \gamma^C}{\gamma^C}\right|.$$

In the next proposition, we derive how $\Delta^C$, $\mathcal{I}^C$, $\mathcal{L}^C$, and $\mathcal{E}^C$ vary across various market structures and information environments. The proof is in Appendix D.

**Proposition 3.6** (Effects of Market Structures and Information Environments)**.** *If $\theta$ is sufficiently small or $\xi$ is sufficiently large, the price-trigger collusive Nash equilibrium satisfies the following properties:*

$$\textit{(i)}\ I\uparrow \implies \Delta^C\downarrow \ \ \& \ \ \mathcal{I}^C/\mathcal{I}^M\uparrow \ \ \& \ \ \mathcal{L}^C/\mathcal{L}^M\uparrow \ \ \& \ \ \mathcal{E}^C\downarrow$$

$$\textit{(ii)}\ \sigma_u/\sigma_v\uparrow \implies \Delta^C\downarrow \ \ \& \ \ \mathcal{I}^C/\mathcal{I}^M\uparrow \ \ \& \ \ \mathcal{L}^C/\mathcal{L}^M\uparrow \ \ \& \ \ \mathcal{E}^C\downarrow$$

$$\textit{(iii)}\ \rho\uparrow \implies \Delta^C\uparrow \ \ \& \ \ \mathcal{I}^C/\mathcal{I}^M\downarrow \ \ \& \ \ \mathcal{L}^C/\mathcal{L}^M\downarrow \ \ \& \ \ \mathcal{E}^C\uparrow$$

$$\textit{(iv)}\ \xi\uparrow \implies \Delta^C\uparrow \ \ \& \ \ \mathcal{I}^C/\mathcal{I}^M\downarrow \ \ \& \ \ \mathcal{L}^C/\mathcal{L}^M\downarrow \ \ \& \ \ \mathcal{E}^C\uparrow$$

# 4 Simulation Experiments with AI-Powered Trading

The theoretical results presented in Section 3 are predicated on the assumption that both the informed speculators and the market maker possess rational expectations. Specifically, they are capable of discerning (i) the order flows of other informed speculators, albeit with noise; (ii) the distribution of noise trading flows; and (iii) the distribution of the fundamental value of the asset. Furthermore, both the informed speculators and the market maker are sufficiently astute, with the speculators being able to communicate amongst themselves. This allows the informed speculators to collectively reach and sustain a price-trigger strategy characterized by $\chi^C(v)$ and $q(v)$, as detailed in (3.12) to (3.14). Meanwhile, this also allows the market maker to perfectly understand the collusion scheme of these speculators.

It remains uncertain whether autonomous, model-free AI algorithms can learn to sustain tacit collusion during trading – and thereby generate supercompetitive profits – in line with the theoretical predictions, which are derived based on stringent, and at times, unrealistic assumptions. As a proof-of-concept illustration, in this section, we design simulation experiments to investigate the capability of Q-learning algorithms to attain tacit collusion under asymmetric information, without the overt acts of communication or agreements typically seen in competition law infringements (Harrington, 2018).

## 4.1 Informed AI Speculators with Q-Learning

We consider informed speculators operating Q-learning algorithms (i.e, informed AI speculators) to learn how to trade. Importantly, informed AI speculators have no direct knowledge of order flows from their counterparts and are oblivious to the distribution of noisy trading flows and the fundamental value of the asset. Our experimental design and methodology are similar to the studies of Calvano et al. (2020) and Asker, Fershtman and Pakes (2022), who explore product market competition under which asymmetric information and endogenous pricing rules are absent.

Specifically, each informed AI speculator $i \in \{1, \cdots, I\}$ adopts the Q-learning algorithm described in Section 2. Observing $s_t$, informed AI speculator $i$ chooses its order flow $x_{i,t}$, following one of the two experimentation modes described in Section 2.3. After receiving the total quantity

of market orders, the market maker determines the price $p_t$ according to its own pricing rules (see Subsection 4.2 below). The profit of informed AI speculator $i$ in period $t$ is given by $\pi_{i,t} = (v_t - p_t)x_{i,t}$.

***State Variables.*** State variables, $s_t$, are essential for characterizing the recursive relation presented in equation (2.4). The choice of state variables is not unique. In principle, $s_t$ can encompass any information that informed AI speculator $i$ has observed up to the beginning of period $t$. This includes both public information and speculator $i$'s own private information. We utilize the smallest possible set of state variables in $s_t$ that can theoretically generate tacit collusion sustained by price-trigger strategies. First, drawing from the insights in Section 3.4, we include the asset's price $p_{t-1}$ in the preceding period $t-1$ as part of $s_t$. Second, we incorporate $v_t$, instead of $v_{t-1}$, as part of $s_t$ because informed AI speculators engage in trading activities in period $t$ after observing $v_t$ at the beginning of period $t$. Thus, the state variable $s_t$ is defined as $s_t \equiv \{p_{t-1}, v_t\}$. Put simply, we equip the informed AI speculator with a one-period memory to trace the history for decision making, similar to the approach adopted by Calvano et al. (2020).

One could also expand informed AI speculator $i$'s state variables in $s_t$ with its own lagged order flow $x_{i,t-1}$, a piece of private information only known by informed AI speculator $i$, and a longer memory for lagged asset prices and order flows. In our simulation experiments, we observe that enlarging the state variable $s_t$ augments the degree of tacit collusion among informed AI speculators, leading to higher trading profits. Thus, our deliberate choice to solely incorporate $p_{t-1}$ and $v_t$ as state variables sets a stringent bar for the Q-learning algorithms to reach tacit collusion within our economic environment. Furthermore, the Q-learning algorithm with state variables $s_t \equiv \{p_{t-1}, v_t\}$ has a convergence speed significantly faster than those incorporating a more extensive list of state variables.[8]

The evolution of state variable $s_t$ is given by $s_{t+1} \equiv \{p_t, v_{t+1}\}$, where $v_{t+1}$ is randomly drawn from the distribution $N(\overline{v}, \sigma_v^2)$. The price $p_t$ is determined by the market maker, and it depends on the noise trading flow and the order from the preferred-habitat investor in period $t$, which remain unknown to informed AI speculators when they make decisions in period $t$.

***Role of Exploration and Exploitation in Generating Collusive Outcomes.*** Exploration is not only critical for approximating the true Q-matrix but also for informed AI speculators to learn and sustain the collusion through price-trigger strategies discussed in Section 5.1. In each iteration, the randomly selected order flow typically differs significantly from the exploited order flow that generates collusive profits. Thus, such deviation, triggered by exploration, provides the only opportunity for the algorithms to learn the price-trigger strategies to sustain the collusion through punishment threat.

Exploitation, as a defining characteristic of RL algorithms, plays a vital role in generating collusion through homogenized learning biases discussed in Section 5.2. Specifically, exploitation biases the estimation of the Q-matrix away from its true values. This bias leads to excessive

[8]When dealing with an extensive list of state variables, deep Q-learning algorithms become indispensable.

overestimation of Q-values for certain choices that can sustain collusive profits, while simultaneously underestimating Q-values for other choices in $\mathcal{X}$. The collusion through homogenized learning biases shares a foundation with the fundamental concept of the "bias-variance tradeoff" in supervised machine learning algorithms – sacrificing unbiasedness to gain stronger identification. Although Q-learning algorithms are inherently self-oriented, they can achieve and maintain collusive profits through interactions by overestimating the Q-values of choices that facilitate high collusive profits. Consequently, under the influence of the biased estimated Q-matrix, informed AI speculators lack incentives to deviate from collusive behavior. Such behaviors constitute a unique character of AI algorithms, which is intrinsically different from how human traders would behave.

### 4.2 Pricing Rule of the Adaptive Market Maker

The market maker does not know the distributions of randomness. It stores and analyzes historical data on asset values, asset prices, the order flows from the preferred-habitat investor, and the combined order flows from informed AI speculators and the noise trader, i.e., $\mathcal{D}_t \equiv \{(v_{t-\tau}, p_{t-\tau}, z_{t-\tau}, y_{t-\tau})\}_{\tau=1}^{T_m}$, where $T_m$ is a large integer. The market maker estimates the demand curve of the preferred-habitat investor and the conditional expectation of the asset value, $\mathbb{E}[v_t|y_t]$, using the following linear regression models:

$$z_{t-\tau} = \xi_0 - \xi_1 p_{t-\tau}, \tag{4.1}$$

$$v_{t-\tau} = \gamma_0 + \gamma_1 y_{t-\tau} + \epsilon_{t-\tau}, \tag{4.2}$$

where $\tau = 1, \cdots, T_m$. The estimated coefficients are $\widehat{\xi}_{0,t}$, $\widehat{\xi}_{1,t}$, $\widehat{\gamma}_{0,t}$, and $\widehat{\gamma}_{1,t}$, respectively, based on the dataset $\mathcal{D}_t$ in period $t$. The pricing rule adaptively adheres to the theoretical optimal policy using a plug-in procedure:

$$p_t(y) = \widehat{\gamma}_{0,t} + \widehat{\lambda}_t y \ \text{ with } \ \widehat{\lambda}_t = \frac{\theta \widehat{\gamma}_{1,t} + \widehat{\xi}_{1,t}}{\theta + \widehat{\xi}_{1,t}^2}, \tag{4.3}$$

where $\theta$ is the market maker's own choice. Therefore, the market maker is adaptive using simple statistical models. To show robustness of our results, we also consider the economic environment where the market maker determines the pricing rule with rational expectations or the market maker adopts Q-learning algorithms to learn the pricing rule (see Appendix F). All the results are similar to those obtained in the baseline economic environment.

### 4.3 Repeated Games of Machines

At $t = 0$, each informed AI speculator $i \in \{1, \cdots, I\}$ is assigned with an arbitrary initial Q-matrix $\widehat{Q}_{i,0}$ and state $s_0$. Then, the economy evolves from period $t$ to period $t+1$ as follows:

(1) Informed AI speculator $i$ draws a random value that determines whether it will be in the exploration mode with probability $\varepsilon_t$ or the exploitation mode with probability $1 - \varepsilon_t$ in

period $t$. The random values drawn by different informed AI speculators are independent. Subsequently, each informed AI speculator $i$ submits its own order flow $x_{i,t}$ according to its mode.

(2) The noise trader submits its order flow $u_t$, which is randomly drawn from a normal distribution $N(0, \sigma_u^2)$.

(3) The preferred-habitat investor submits its order flow $z_t$ according to (3.2).

(4) The market maker observes the historical data $\mathcal{D}_t \equiv \{v_{t-\tau}, p_{t-\tau}, z_{t-\tau}, y_{t-\tau}\}_{\tau=1}^{T_m}$ and estimates the optimal pricing rule according to (4.1) – (4.3).

(5) Each informed AI speculator $i$ realizes its profits $(v_t - p_t)x_{i,t}$ and updates its Q-matrix according to equation (2.4).

(6) At the beginning of period $t+1$, the state variable for each informed AI speculator evolves to $s_{t+1} = \{p_t, v_{t+1}\}$, where $v_{t+1}$ is drawn from $N(\overline{v}, \sigma_v^2)$ and is independent of any other variables.

The interactions of informed AI speculators and an adaptive market maker, together with the randomness caused by the noise trader and stochastic asset values in the background, make the stationary equilibrium difficult to achieve. The economic environment in our study is substantially more complex than that of Calvano et al. (2020) whose setting does not have randomness, information asymmetry, or endogenous pricing rules. As noted by Calvano et al. (2020), the player's optimization problem is inherently nonstationary when its rivals vary their actions over time due to experimentation or learning. Theoretical analysis of the Q-learning algorithms playing repeated games is generally not tractable. Rather than applying stochastic approximation techniques to AI agents, we follow Calvano et al. (2020) by simulating the exact stochastic dynamic system a large number of times to smooth out uncertainty. There is no theoretical guarantee that Q-learning agents will settle on a stable outcome, nor that they will correctly learn an optimal policy. However, we can always verify this in our simulations ex post to ensure that our analyses are conducted based on the stationary equilibrium.

### 4.4 Discretization of State and Action Space

We choose the following grids for the state variable $s_t \equiv \{p_{t-1}, v_t\}$ and action variable $x_{i,t}$. For computational efficiency, we approximate the normal distribution $N(\overline{v}, \sigma_v)$ using a sufficiently larger number of $n_v$ grid points, $\mathbb{V} = \{v_1, \cdots, v_{n_v}\}$. Our discretization ensures that these $n_v$ grid points have equal probabilities but are unequally spaced. Specifically, the probability of each grid point is $\mathcal{P}_k = 1/n_v$. The locations of grid points are chosen based on $v_k = \overline{v} + \sigma_v \Phi^{-1}((2k-1)/(2n_v))$ for $k = 1, \cdots, n_v$, where $\Phi^{-1}$ is the inverse cumulative density function of a standard normal distribution. The mathematical property of $\Phi^{-1}$ implies that grid points around the

mean $\overline{v}$ are closer to each other than those far away from the mean. The speed of convergence is significantly increased because all $n_v$ grid points of $v_t$ have equal probabilities.[9]

We construct the discrete grid points for informed AI speculators' order $x_{i,t}$ based on their optimal actions in the noncollusive Nash equilibrium and perfect cartel equilibrium. According to our model in Section 3, the order values in the two equilibria are given by $x^N = (v - \overline{v})/((I+1)\lambda)$ and $x^M = (v - \overline{v})/(2I\lambda)$. We specify informed AI speculators' action space by discretizing the interval $[x^M - \iota(x^N - x^M), x^N + \iota(x^N - x^M)]$ for $v > \overline{v}$ and $[x^N - \iota(x^M - x^N), x^M + \iota(x^M - x^N)]$ for $v < \overline{v}$ into $n_x$ equally spaced grid points, i.e., $\mathbb{X} = \{x_1, \cdots, x_{n_x}\}$. The parameter $\iota > 0$ ensures that informed AI speculators can choose order flows beyond the theoretical levels corresponding to the noncollusive Nash equilibrium and perfect cartel equilibrium. As the action space is discrete, the exact order flows corresponding to the perfect cartel equilibrium may not be feasible. Despite this, our simulations show that informed AI speculators can collude with each other to a large degree.

The grid points of price $p_t$ are similarly chosen as those of $x_{i,t}$, except for considering the noise trader's impact on prices. Specifically, in our numerical experiments, the noise trader's order is drawn randomly from the normal distribution $N(0, \sigma_u)$, without imposing any discretization or truncation. In our theoretical framework in Section 3, the market maker sets the price according to the total order flow $y_t$, which is the sum of informed AI speculators' order $\sum_{i=1}^{I} x_{i,t}$ and the noise trader's order $u_t$. Because $u_t$ follows an unbounded normal distribution, the theoretical range of the price $p_t$ is unbounded. To maintain tractability, in our numerical experiments, we set the upper bound at $p_H = \overline{v} + \lambda(I \max(x^M, x^N) + 1.96\sigma_u)$ and the lower bound at $p_L = \overline{v} + \lambda(I \min(x^M, x^N) - 1.96\sigma_u)$, corresponding to the 95% confidence interval of the noise trader's order distribution, $N(0, \sigma_u)$. The grid points of $p_t$ are chosen by discretizing the interval $[p_L - \iota(p_H - p_L), p_H + \iota(p_H - p_L)]$ into $n_p$ grids, i.e., $\mathbb{P} = \{p_1, \cdots, p_{n_p}\}$.

### 4.5 Initial Q-Matrix and States

We initialize the Q-matrix at $t = 0$ using the discounted payoff that would accrue to informed AI speculator $i$ if the other informed AI speculators randomize their actions uniformly over the grid points defined by $\mathbb{X}$.[10] Moreover, we consider a zero order flow from the noise trader,

[9] All the results are robust to the use of alternative methods to discretize the state variable $v_t$. For example, one commonly used method is to use $n_v$ equally spaced points over a sufficiently large interval, e.g., $[\overline{v} - 6\sigma_v, \overline{v} + 6\sigma_v]$. The probability of each grid point is computed based on the probability mass function of the normal mass function, i.e., $\mathcal{P}_k = \exp\left(-(k - \overline{v})^2/(2\sigma_v^2)\right)$ for $k = 1, \cdots, n_v$. Compared to the discretization method we use, this alternative method yields similar quantitative results but has a much slower convergence. The reason is that it assigns very small probabilities to the left-most and right-most grid points. As a result, the Q-matrix's cells far away from the mean $\overline{v}$ are updated at much lower frequencies than those closer to the mean. An infrequent update for the cells far away from the mean in turn requires many more updates for other cells of the Q-matrix to stabilize. Thus, the global convergence speed is reduced significantly due to the buckets effect. In fact, as $n_v \to \infty$, the two alternative methods can both perfectly capture the theoretical distribution of $v_t$ but yield vastly different convergence speed for the Q-learning algorithms.

[10] Adopting different initial values for the Q-matrix do not significantly alter the results. In RL algorithms, another common strategy to initialize the Q-matrix is to use optimistic initial values. That is, initializing the Q-matrix with sufficiently high values so that subsequent iterations tend to reduce the values of the Q-matrix. This approach enables Q-learning algorithms to visit all actions multiple times at the beginning, resulting in early improvement in estimated action values. Thus, setting optimistic initial values is in some sense equivalent to adopting a thorough exploration

corresponding to the expected value of the distribution $N(0, \sigma_u^2)$. Specifically, for each informed AI speculator $i = 1, \cdots, I$, we set its initial Q-matrix $\widehat{Q}_{i,0}$ at $t = 0$ as follows:

$$\widehat{Q}_{i,0}(p_m, v_k, x_n) = \frac{\sum_{x_{-i} \in \mathbb{X}} \left[ v_k - (\overline{v} + \lambda(x_n + (I-1)x_{-i})) \right] x_n}{(1-\rho) n_x}, \tag{4.4}$$

for $(p_m, v_k, x_n) \in \mathbb{P} \times \mathbb{V} \times \mathbb{X}$. The initial states of our simulation, $s_0 = \{p_{-1}, v_0\}$, are randomized uniformly over $\mathbb{V} \times \mathbb{P}$.

### 4.6 Specification of Learning Modes

We adopt an exponentially time-declining state-dependent exploration rate for informed AI speculators,

$$\varepsilon_{t(v_k)} = e^{-\beta t(v_k)}, \tag{4.5}$$

where the parameter $\beta > 0$ governs the speed that informed AI speculators' exploration rate diminishes over time and the variable $t(v_k)$ captures the number of times that the exogenous state $v_k \in \mathbb{V}$ has occurred in the past.[11] The specification of $t(v_k)$ implies that the exploration rate is state dependent, which ensures that informed AI speculators can sufficiently explore their actions for all grid points of the exogenous state variable $v_t$.

The specification (4.5) implies that initially, Q-learning algorithms are almost always in the exploration mode, choosing actions randomly. However, as time passes, Q-learning algorithms gradually switch to the exploitation mode.

### 4.7 Parameter Choice

The parameters used in our numerical experiments can be categorized into three groups according to their roles. The environment parameters are the parameters that characterize the underlying economic environment in our experiments. Importantly, the values of most of these parameters are neither known to informed AI speculators nor to the market maker.[12] They instead adopt Q-learning algorithms to learn how to make decisions in an unknown environment. The simulation parameters are the parameters that determine our numerical experiments, such as the number of discrete grid points, simulation sessions, etc. The hyperparameters are the parameters that control the machine learning process. Below, we describe the choice of parameters for each category.

***Environment Parameters.*** Across all simulation experiments, we set $\overline{v} = 1$, $\sigma_v = 1$, and $\theta = 0.1$. The parameter $\overline{v}$ determines the expected value of $v_t$, and thus we normalize its value to unity without loss of generality. The parameter $\sigma_v$ plays a similar role as $\sigma_u$ because what matters in our

over the entire action space early in the learning phase and then exploitation later on.

[11] In principle, we can allow informed AI speculators to choose their exploration rate conditional on the realized value of $v_t$ because they perfectly observe $v_t$, which is one of their state variables $s_t = \{p_{t-1}, v_t\}$.

[12] An exception is $\rho$ and $\theta$. The parameter $\rho$ is known to informed AI speculators as this parameter captures their own discount rates. The parameter $\theta$ is known to the market maker as this is its own choice.

model in Section 3 is the ratio $\sigma_u/\sigma_v$. We thus normalize the value of $\sigma_v$ to unity. The parameter $\theta$ determines the extent to which the market maker focuses on price discovery. We find that the implications of different values of $\theta$ can be analyzed similarly by varying the value of $\xi$. Thus, for simplicity, we fix the value of $\theta$ at 0.1 throughout our simulation experiments.

In the baseline economic environment, we set $I = 2$, $\sigma_u = 0.1$, $\rho = 0.95$, and $\xi = 500$. We extensively study the implications of different values for these parameters. Specifically, we consider different number of informed AI speculators ranging from $I = 2$ to $I = 6$, different levels of background noise ranging from $\sigma_u = e^{-5}$ to $\sigma_u = e^5$, different discount rates ranging from $\rho = 0.5$ to $\rho = 0.95$, and different values of $\xi$ ranging from $\xi = 0$ to $\xi = 500$.

***Simulation Parameters.*** We set $\iota = 0.1$ so that informed AI speculators can go beyond the theoretical bounds of order flows by 10%. We choose $n_x = 15$ and $n_p = 31$. These grid points are sufficiently dense to capture the economic mechanism we are interested in. Importantly, our choice of $n_p \approx 2n_x$ ensures that, all else equal, a one-grid point change in one informed AI speculator's order will result in a change in price $p_t$ over the grid defined by $\mathbb{P}$. If the grid defined by $\mathbb{P}$ is coarser, informed AI speculators will not be able to detect small deviations of peers even in the absence of noise, which in turn lowers the possibility of algorithmic collusion through price-trigger strategies.

We use $n_v = 10$ grid points to approximate the normal distribution of $v_t$. Under our discretization, the standard deviation of $v_t$ is $\widehat{\sigma}_v = \sqrt{\sum_{k=1}^{N} \mathcal{P}(v_k)(v_k - \overline{v})^2} = 0.938$, which is close to the theoretical value $\sigma_v = 1$. In the remainder of this paper, the theoretical benchmarks of noncollusive Nash equilibrium and perfect cartel equilibrium are computed using $\widehat{\sigma}_v$, to be consistent with the discretization of $v_t$ adopted in our simulation experiments.

All the results are robust if we choose a larger $n_v$, $n_x$, $n_p$, or $\iota$, as long as the hyperparameters, $\alpha$ and $\beta$, are adjusted accordingly to ensure sufficiently good learning outcomes. However, the cost of using denser grids is that significantly longer time is needed for Q-learning algorithms to fully converge to limit strategies.

We set $T_m = 10,000$ so that the market maker stores sufficiently long time-series data to estimate the linear regressions (4.1) and (4.2). In our simulation experiments, we verify that the estimates of $\widehat{\xi}_{0,t}$, $\widehat{\xi}_{1,t}$, $\widehat{\gamma}_{0,t}$, and $\widehat{\gamma}_{1,t}$ can accurately recover the preferred-habitat investor's demand curve and the conditional expectation of the asset value, $\mathbb{E}[v_t|y_t]$. Increasing the value of $T_m$ will not change any quantitative results, but it adds more computation burden.

For each experiment with a particular choice of environment parameters, we simulate the Q-learning algorithms by $N = 1,000$ times. All the random initial states and shocks (i.e., $v_t$, $u_t$, and exploration status of each informed AI speculator for all $t \geq 0$) are independently drawn from identical distributions across the $N$ simulation sessions of the experiment. In principle, the results of different experiments can differ both because of the difference in environment parameters and the difference in the realized values of random variables. To ensure that comparisons across different experiments are not contaminated by the latter, we generate a large set of random variables for all $N$ simulation sessions offline and store in the high-powered-computing server.

The same set of random values is used when we compare results across the experiments with different environment parameters in Sections 5 and 6.

***Hyperparameters.*** The hyperparameters that control the learning process of Q-learning algorithms are set at $\alpha = 0.01$ and $\beta = 10^{-5}$. All results are robust to choosing different values of $\alpha$ and $\beta$ so long as they are in the reasonable range that ensures sufficiently good learning outcomes. Our baseline choice of $\beta$ implies that any action $x_k \in \mathbb{X}$ is visited purely by random exploration by $n_v/[(1-\exp(-10^{-5}))n_x] = 66,660$ times on average before exploration completes.[13] In Section 6.3, we study the experiments with different values of $\alpha$ and $\beta$ as well as the experiments that allow informed AI speculators to adopt different values of $\alpha$. In Section 7.2, we develop a two-tier Q-learning algorithms that allow informed AI speculators to learn the choice of $\alpha$.

### 4.8 Convergence

Strategic games played by Q-learning algorithms do not have general convergence results. To verify convergence, a practical criterion is to check whether each player's optimal strategy does not change for a long period of time. Note that convergence is determined by the stationarity of players' optimal strategies rather than the stationarity of players' learned Q-matrices. In fact, in a stochastic environment, the Q-matrix can never remain unchanged because randomly realized shocks will always result in an update for some cells of the Q-matrix. However, the slight update in the Q matrix does not necessarily result in a change in the optimal strategies. This is why convergence in optimal strategies can be achieved in principle, even in a stochastic environment with Q-learning algorithms playing repeated games.

In general, setting a smaller value of $\alpha$ or $\beta$ requires longer time for the algorithm to reach convergence. For example, with $\beta = 10^{-5}$, informed AI speculators' Q-learning algorithms are still doing exploration with $e^{-\beta T'/n_v} = 36.8\%$ probability after $T' = 1,000,000$ periods. It is almost by definition that the optimal strategies are nonstationary with an exploration rate that is far away from zero. Thus, a necessary condition for all Q-learning algorithms to reach stationary optimal strategies is that exploration rate is virtually zero, say, after 10,000,000 periods. Moreover, with a small $\alpha$, the Q-matrix is updated slowly when new information arrives. As a result, informed AI speculators can only slowly learn their optimal actions, which are based on their learned Q-matrices. A sufficiently long time is needed to ensure the convergence of optimal strategies.

Per discussions above, we adopt a stringent criterion of convergence by requiring all informed AI speculators' optimal strategies to stay unchanged for 1,000,000 consecutive periods. All $N = 1,000$ simulation sessions are simulated until convergence. The number of periods needed to reach convergence varies considerably across experiments depending on the particular choice of environment parameters. Moreover, even for the same experiment, the number of periods needed to reach convergence can vary significantly across the $N$ simulation sessions, depending

[13] We do not have an explicit formula for the expected number of times a cell in the Q-matrix being visited by random exploration because the state variable $p_{t-1}$ in $s_t = \{p_{t-1}, v_t\}$ is also affected by the noise trader's random order and the pricing rule adopted by the market maker.

on the realized values of random variables. Among all the experiments we study, the number of periods to reach convergence ranges from about 20 million to about 10 billion. To speed up computations, our programs are written in C++, using $-O2$ to optimize the compiling process. The C++ program is run with parallel computing in a high-powered-computing server cluster with 376 CPU cores in total. It takes about 1 min to 6 hours to finish all $N$ simulation sessions in one experiment, depending on the total number of iterations needed to reach convergence.

### 4.9 Metrics Reflecting Collusive Behavior

Motivated by our theoretical results in Section 3, we calculate three simple metrics that can be indicative of potential collusive behavior among informed AI speculators. The values of all three metrics are computed in each simulation session over $T = 100,000$ periods, after informed AI speculators' optimal strategies fully converge to the limit strategies according to the convergence criterion in Section 4.8. By taking the average over a large number of periods, we smooth out the stochastic underlying economic environment, caused by the randomness in the noise trader's order $u_t$ and the stochastic variation of the asset value $v_t$ over time.

***Collusion Capacity.*** The degree of collusion can be reflected by the Delta metric defined as follows:

$$\Delta^C = \frac{1}{I}\sum_{i=1}^{I}\Delta_i^C, \text{ with } \Delta_i^C = \frac{\overline{\pi}_i - \overline{\pi}_i^N}{\overline{\pi}_i^M - \overline{\pi}_i^N}, \tag{4.6}$$

where $\overline{\pi}_i \equiv \sum_{t=T_c}^{T_c+T}\pi_{i,t}(v_t,u_t)$ is the average profits of informed AI speculator $i$ over $T$ periods after Q-learning algorithms reach convergence at $T_c$. The values of $\overline{\pi}_i^N = \sum_{t=T_c}^{T_c+T}\pi_i^N(v_t,u_t)$ and $\overline{\pi}_i^M = \sum_{t=T_c}^{T_c+T}\pi_i^M(v_t,u_t)$ are the average profit that informed speculator $i$ would obtain, theoretically, in the noncollusive Nash equilibrium or perfect cartel equilibrium, respectively. Because informed speculators are symmetric, we have $\pi_i^N(v_t,u_t) \equiv \pi^N(v_t,u_t)$ and $\pi_i^M(v_t,u_t) \equiv \pi^M(v_t,u_t)$ for all $i = 1,..,I$. Specifically, according to the formulas in Section 3.2, conditional on the realized values of $v_t$ and $u_t$ in period $t$, informed speculator $i$'s profit in the noncollusive Nash equilibrium is

$$\pi^N(v_t,u_t) = \left[v_t - p^N(Ix^N(v_t)+u_t)\right]x^N(v_t), \text{ for } i = 1,\cdots,I, \tag{4.7}$$

where $x^N(v_t) = \chi^N(v_t-\overline{v})$ and $p^N(Ix^N(v_t)+u_t) = \overline{v} + \lambda^N(Ix^N(v_t)+u_t)$. Similarly, according to the formulas in Section 3.3, conditional on the realized values of $v_t$ and $u_t$ in period $t$, informed speculator $i$'s profit in the perfect cartel equilibrium is

$$\pi^M(v_t,u_t) = \left[v_t - p^M(Ix^M(v_t)+u_t)\right]x^M(v_t), \text{ for } i = 1,\cdots,I, \tag{4.8}$$

where $x^M(v_t) = \chi^M(v_t-\overline{v})$ and $p^M(Ix^M(v_t)+u_t) = \overline{v} + \lambda^M(Ix^M(v_t)+u_t)$.

In principle, the value of $\Delta^C$ should range from 0 to 1. A larger $\Delta^C$ implies that informed AI speculators attain higher profits. The value of $\Delta^C$ can never be larger than 1 because $\overline{\pi}_i^M$ is the highest theoretically possible average profit. In fact, because informed AI speculators can

only choose actions over discrete grids, by design, it is not possible to obtain $\Delta^C = 1$ in our simulation experiments. However, it is possible to achieve a $\Delta^C$ below 0 under the limit strategies of informed AI speculators. This outcome implies that informed AI speculators failed to learn a good approximation of the actual Q-matrix, and as a result, they achieve average profits lower than those in the noncollusive Nash equilibrium.

***Profit Gain Relative to Noncollusion.*** The Delta metric is informative about collusive behavior. However, it does not tell us the relative magnitude of supra-competitive profits. We thus also calculate the extra profit gain relative to the profits that informed AI speculators would obtain in the noncollusive Nash equilibrium theoretically. Specifically, the relative profit gain is $\sum_{i=1}^{I} \overline{\pi}_i / \sum_{i=1}^{I} \overline{\pi}_i^N$, where $\overline{\pi}_i$ and $\overline{\pi}_i^N$ are calculated similarly as those in equation (4.6).

***Order Sensitivity to Asset Value.*** In our model, each informed speculator's order flows $x_{i,t}$ are linear in the asset value $v_t$, as captured by $x_{i,t} = \chi^C(v_t - \overline{v})$. Our model implies that informed speculators are more conservative in placing their orders if there is implicit collusion. That is, the sensitivity of order flows $x_{i,t}$ to the asset value $v_t - \overline{v}$ is lower when informed speculators collude more, i.e., $\chi^M \leq \chi^C < \chi^N$.

In our simulation experiments, informed AI speculators directly learn $x_{i,t}$ without imposing the linearity restriction between $x_{i,t}$ and $v_t$. Despite this, we find that informed AI speculators learn roughly linear strategies (see Figure 8). We estimate $\widehat{\chi}^C$ based on the recorded asset values and order flows $\{v_t, x_{i,t}\}_{t=T_c}^{T_c+T}$ for each AI speculator $i = 1, \cdots, I$, by running the following linear regression:

$$x_{i,t} = \chi_{i,0}^C + \chi_{i,1}^C v_t + \epsilon_t. \tag{4.9}$$

Consistent with our model, the estimates based on the simulated data satisfy $\widehat{\chi}_{i,0}^C \approx -\overline{v}\widehat{\chi}_{i,1}^C$ in the unrestricted regression (4.9). The estimate $\widehat{\chi}_{i,1}^C$ captures the sensitivity of $x_{i,t}$ to $v_t$ corresponding to the optimal trading strategies after Q-learning algorithms converge. We further compute the average sensitivity of informed AI speculators as $\widehat{\chi}^C = \frac{1}{I}\sum_{i=1}^{I} \widehat{\chi}_{i,1}^C$.

### 4.10 Measures of Price Informativeness, Market Liquidity, and Mispricing

***Price Informativeness.*** Consistent the our model, the degree of price informativeness in our simulation experiments is measured by the log signal-noise ratio as follows:

$$\mathcal{I}^C = \log\left[\frac{\text{var}(x_{i,t}^C)}{\text{var}(u_t)}\right] = \log\left[(I\widehat{\chi}^C)^2(\widehat{\sigma}_v/\sigma_u)^2\right], \tag{4.10}$$

where $\widehat{\sigma}_v$ is the standard deviation of $v_t$ under our discrete grid points in $\mathbb{V}$.

***Market Liquidity.*** Consistent the our model, the market liquidity in period $t$ is measured by the inverse sensitivity of market maker's inventory $|z_t + y_t|$ to noise order flows $u_t$

$$\mathcal{L}_t^C = \frac{1}{\partial |z_t + y_t| / \partial u_t} = \frac{1}{|1 - \xi \widehat{\lambda}_t|}, \tag{4.11}$$

where $z_t = -\xi(p_t - \overline{v}) = -\xi \widehat{\lambda}_t y_t$ and $\widehat{\lambda}_t$ is given by equation (4.3). The average market liquidity is computed as $\mathcal{L}^C = \sum_{t=T_c}^{T_c+T} \mathcal{L}_t^C$.

***Mispricing.*** Consistent the our model, the magnitude of mispricing in period $t$ is measured by the percentage deviation of the asset's price $p_t$ from its conditional expected value

$$\mathcal{E}_t^C = \left| \frac{p_t - \mathbb{E}[v_t|y_t]}{\mathbb{E}[v_t|y_t] - \overline{v}} \right| = \left| \frac{\widehat{\lambda}_t - \widehat{\gamma}_{1,t}}{\widehat{\gamma}_{1,t}} \right|, \tag{4.12}$$

where $p_t = \widehat{\gamma}_{0,t} + \widehat{\lambda}_t y_t$ and $\mathbb{E}[v_t|y_t] = \widehat{\gamma}_{0,t} + \widehat{\gamma}_{1,t} y_t$; $\widehat{\gamma}_{0,t}$ and $\widehat{\gamma}_{1,t}$ are estimated from (4.2). The average mispricing is computed as $\mathcal{E}^C = \sum_{t=T_c}^{T_c+T} \mathcal{E}_t^C$.

# 5 AI Collusion under Information Asymmetry

Our model suggests that informed speculators can achieve supra-competitive profits through implicit collusion when both price efficiency and noise trading risks are low (see Proposition 3.4). In this section, we conduct simulation experiments with informed AI speculators whose trading is powered by Q-learning algorithms. We are mainly interested in four questions. First, can informed AI speculators learn to collude, even without communicating with each other or possessing any information about the underlying economic environment? Second, if collusion exists, what are the mechanisms that generate such collusive behavior among informed AI speculators? Third, how price efficiency and noise trading risk affect the trading strategies of informed AI speculators. Fourth, what are the implications of AI-powered trading for price informativeness, market liquidity, and mispricing in financial markets?

In Subsection 5.1, we show that in environments with low price efficiency and low noise trading risks, informed AI speculators are able to learn price-trigger strategies to achieve implicit collusion, which is quite similar to the mechanism characterized in our model in Section 3.4. In Subsection 5.2, we show that in environments with low price efficiency and high noise trading risks, informed AI speculators are not able to learn price-trigger strategies to achieve collusion, as predicted by our model. However, they can still achieve supra-competitive profits due to biased learning. The equilibrium of informed AI speculators resembles a self-confirming equilibrium (Fudenberg and Kreps, 1988; Fudenberg and Levine, 1993) with collusion rather than a Nash equilibrium. In Subsection 5.3, we study the role of price efficiency and noise trading risks in determining informed AI speculators' profits and collusive behavior. In Subsection 5.4, we illustrate informed AI speculators' trading strategies. Finally, in Subsection 5.5, we study the

implications of AI-powered trading for price informativeness, market liquidity, and mispricing in financial markets.

### 5.1 Artificial Intelligence: Collusion through Price-Trigger Strategies

In this subsection, we study informed AI speculators' behavior when the environment has low price efficiency (i.e., $\xi = 500$) and low noise trading risks (i.e., $\sigma_u/\sigma_v = 10^{-1}$). The other parameters are set according to the baseline economic environment described in Section 4.7. Across all $N = 1,000$ simulation sessions, the average value of $\Delta^C$ is about 0.73 and the average profit of informed AI speculators is about 9% higher than the profit in the noncollusive equilibrium. Thus, our simulation results indicate that informed AI speculators can achieve supra-competitive profits. Below, we examine the mechanism that sustains their collusion. We show that informed AI speculators are intelligent enough to learn price-trigger strategies, which allows them to sustain collusion after their Q-learning algorithms converge. These simulation results with informed AI speculators are similar to the theoretical predictions of our model with rational-expectation informed speculators.

#### 5.1.1 Price-Trigger Strategy

Motivated by our model, we examine whether the optimal strategies learned by informed AI speculators are consistent with the price-trigger strategy illustrated in Section 3. To this end, in Figure 1, we study the impulse response function (IRF) after an exogenous shock to the noise order flow, which further affects the asset's price given the market maker's pricing rule.

Specifically, in each of the $N = 1,000$ simulation sessions, we focus on the economic environment after informed speculators' Q-learning algorithms converge. Throughout the IRF experiment, for all $t \geq 0$, both informed AI speculators play their learned optimal strategies and the asset's price $p_t$ is determined by the market maker according to its learned pricing rule. In period $t = 3$, we introduce an unexpected exogenous shock $\Delta u_t$ to the noise order flow $u_t$. The direction of the shock is made to mimic the price impact of a hypothetical profitable deviation from informed AI speculators. That is, we choose $\Delta u_t > 0$ if $v_t > \overline{v}$ and $\Delta u_t < 0$ if $v_t < \overline{v}$. Thus, all else equal, this exogenous shock will unexpectedly increase the asset's price $p_t$ if $v_t > \overline{v}$ and decrease $p_t$ if $v_t < \overline{v}$.

We are interested in the IRF of three outcome variables. The first outcome variable is the price's percentage deviation from its long-run mean, defined by $(\widetilde{p}_t - \mathbb{E}[\widetilde{p}_t])/\mathbb{E}[\widetilde{p}_t]$, where $\widetilde{p}_t = (p_t - \overline{v})sgn(v_t - \overline{v})$ and $sgn(\cdot)$ is the sign function. The variable $\widetilde{p}_t$ captures the difference between the asset's price $p_t$ and its expected value. The sign function ensures that $\widetilde{p}_t > 0$ because according to our model and simulation results, when $v_t > \overline{v}$, we have $p_t > \overline{v}$ and $sgn(v_t - \overline{v}) = 1$; when $v_t < \overline{v}$, we have $p_t < \overline{v}$ and $sgn(v_t - \overline{v}) = -1$. In addition, the definition of $\widetilde{p}_t$ ensures that the exogenous shock always increases its value, enabling us to take the average of IRF across simulation paths for expositional purposes. Specifically, if $v_t > \overline{v}$, the exogenous shock will increase $p_t$, and because $sgn(v_t - \overline{v}) = 1$, $\widetilde{p}_t$ also increases. If $v_t < \overline{v}$, the exogenous shock will decrease $p_t$, and because $sgn(v_t - \overline{v}) = -1$, $\widetilde{p}_t$ also increases. The second outcome variable is

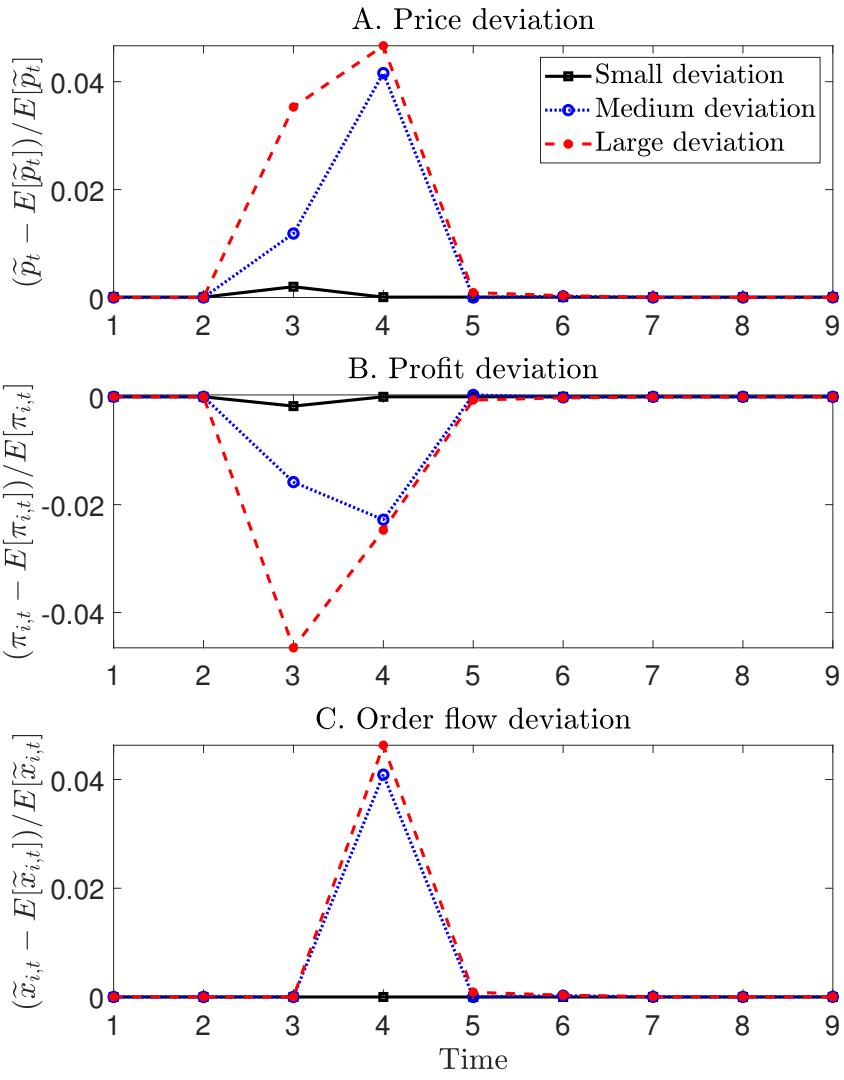
each informed AI speculator's profit's percentage deviation from its long-run mean, defined by $(\pi_{i,t} - \mathbb{E}[\pi_{i,t}])/\mathbb{E}[\pi_{i,t}]$. The third outcome variable is each informed AI speculator's order flow's percentage deviation from its long run mean, defined by $(\widetilde{x}_{i,t} - \mathbb{E}[\widetilde{x}_{i,t}])/\mathbb{E}[\widetilde{x}_{i,t}]$, where $\widetilde{x}_{i,t} = x_{i,t} sgn(v_t - \overline{v})$. The sign function ensures that $\widetilde{x}_{i,t} > 0$ because according to our model and simulation results, we have $x_{i,t} > 0$ when $v_t > \overline{v}$ and $x_{i,t} < 0$ when $v_t < \overline{v}$.

To clearly present the IRF, we calculate the average value of the above three interested outcome variables in two steps. First, for each of the $N = 1,000$ simulation sessions, we use the learned optimal strategies to simulate the IRF 10,000 times, with independently drawn random shocks to $v_t$ and $u_t$. We smooth out the randomness in the economic environment by taking the average IRF across these 10,000 independent paths. This is referred to as the IRF for each simulation session $i = 1, ..., N$. Second, we compute the average IRF across $N = 1,000$ simulation sessions. This allows us to smooth out the randomness (i.e., initial states and exploration choices) during the learning process. However, our results hold not merely to the average IRF of $N = 1,000$ simulation sessions. Figure 2 plots the distribution of the impulse responses across the $N = 1,000$ simulation sessions. Although the magnitudes of the deviations in prices and trading flows differ significantly across simulation sessions, the $[25\%, 75\%]$ and $[5\%, 95\%]$ confidence intervals indicate that price-trigger strategies are consistently adopted by informed AI speculators.

Figure 1 plots the average IRF across the $N = 1,000$ simulation sessions for each outcome variable of interest. We consider exogenous shocks of different magnitudes. In the scenario with "small deviation," $|\Delta u_t|$ is roughly 0.5% of the average magnitude of informed AI speculators' order flow $|x_{i,t}|$. Thus, it generates a small impact on the asset's price $p_t$ at $t = 3$. In the scenario with "medium deviation" and "large deviation," $|\Delta u_t|$ is about 2.5% and 7% of the average magnitude of informed AI speculators' order flow $|x_{i,t}|$, respectively, resulting in much larger changes in $p_t$.

Panel A plots the price's percentage deviation from its long-run mean. Due to the exogenous shock, the asset's price deviates from its long-run mean in period $t = 3$, and the size of price deviation increases with the magnitude of the exogenous shock. Panel B plots the profit's percentage deviation from its long-run mean for one informed AI speculator. The other informed AI speculator has similar profit dynamics. It is shown that in period $t = 3$, the price deviation reduces the informed AI speculator's profit, and the impact increases with the magnitude of the price deviation. Panel C plots the order flow's percentage deviation from its long-run mean for one informed AI speculator. The deviation is zero in period $t = 3$ because informed AI speculators submit their orders before observing the price in the same period.

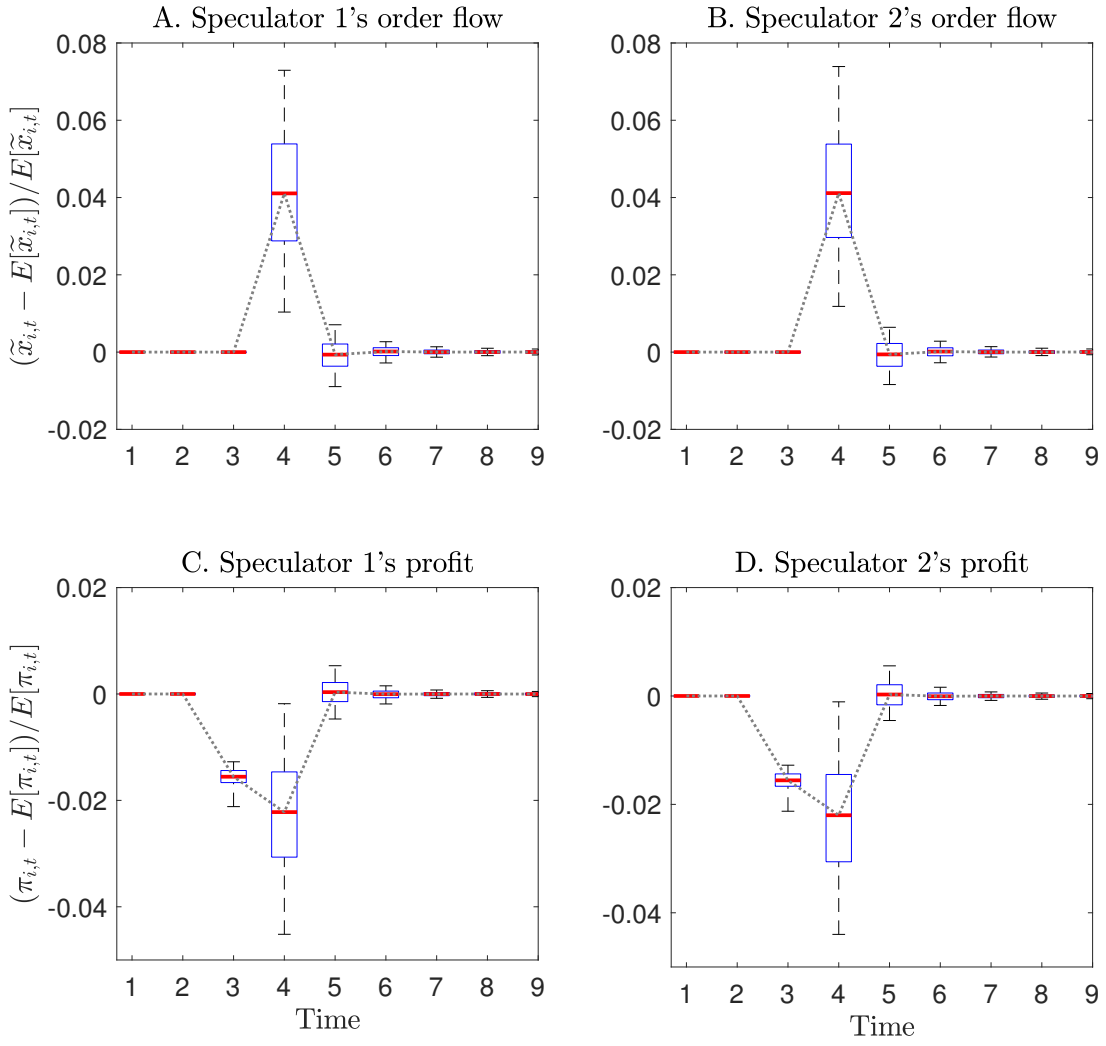
In period $t = 4$, panel C shows that in response to medium and large price deviations occurred in the previous period, the informed AI speculator's order flow significantly deviates from its long-run mean. Moreover, the magnitude of the order flow deviation is similar for the medium and large price deviation. However, the informed AI speculator's order flow does not respond to small price deviation. These patterns resemble the price-triggers strategies described in Section 3. Panel A shows that for the medium and large deviation cases, the percentage deviation of the asset's price continues to increase as a result of increased order flows from informed AI




Note: In each simulation session, we focus on the economic environment after informed speculators' Q-learning algorithms converge. Throughout the IRF experiment, for all $t \geq 0$, both informed AI speculators play their learned optimal strategies and the asset's price $p_t$ is determined by the market maker according to its learned pricing rule. In period $t = 3$, we introduce an unexpected exogenous shock $\Delta u_t$ to the noise order flow $u_t$. The direction of the shock is made to mimic the price impact of a hypothetical profitable deviation from informed AI speculators. That is, we choose $\Delta u_t > 0$ if $v_t > \overline{v}$ and $\Delta u_t < 0$ if $v_t < \overline{v}$. Thus, all else equal, this exogenous shock will unexpectedly increase the asset's price $p_t$ if $v_t > \overline{v}$ and decrease $p_t$ if $v_t < \overline{v}$. The three curves in each panel represent different magnitudes of the shock. Panel A plots the price's percentage deviation from its long-run mean. Panels B and C plot the percentage deviation of profit and order flow from its long-run mean for one informed AI speculator, respectively. All curves are average values across $N = 1,000$ simulation sessions, where each session is independently simulated 10,000 times to smooth out the effect of random shocks to $v_t$ and $u_t$. We set $\sigma_u/\sigma_v = 10^{-1}$. The other parameters are set according to the baseline economic environment described in Section 4.7.

Figure 1: IRF after an exogenous shock to $u_t$ ($\sigma_u/\sigma_v = 10^{-1}$).

speculators. This, in turn, results in continued profit losses for informed AI speculators (see panel B). By contrast, for the small deviation case, both of the asset's price and informed AI speculators'




Note: The experiment is similar to that described for Figure 1. Panels A and B plot the two speculators' order flow's percentage deviation from the long-run mean, and panels C and D plot their profit's percentage deviation from the long-run mean. In each panel, the dotted line represents the median value, the boxes represent the 25th and 75th percentiles, and the dashed intervals represent the 5th and 95th percentiles across $N = 1,000$ simulation sessions. Parameters are set as in Figure 1.

Figure 2: Confidence intervals for the IRF after an exogenous shock to $u_t$ ($\sigma_u/\sigma_v = 10^{-1}$).

profits revert back to the long-run mean.

In period $t = 5$, panel C shows that informed AI speculators' order flows abruptly return to the long-run mean for both the medium and large deviation cases. As a result, both the price and profit deviation abruptly return to zero (see panels A and B).

### 5.1.2 Punishment for Deviation

According to our model in Section 3, price-trigger strategies are implemented based on whether the asset's price in the preceding period deviates from its long-run mean, which could be caused by either the random order flows from the noise trader or the order flows from informed AI speculators. Informed AI speculators cannot distinguish what causes price deviation.

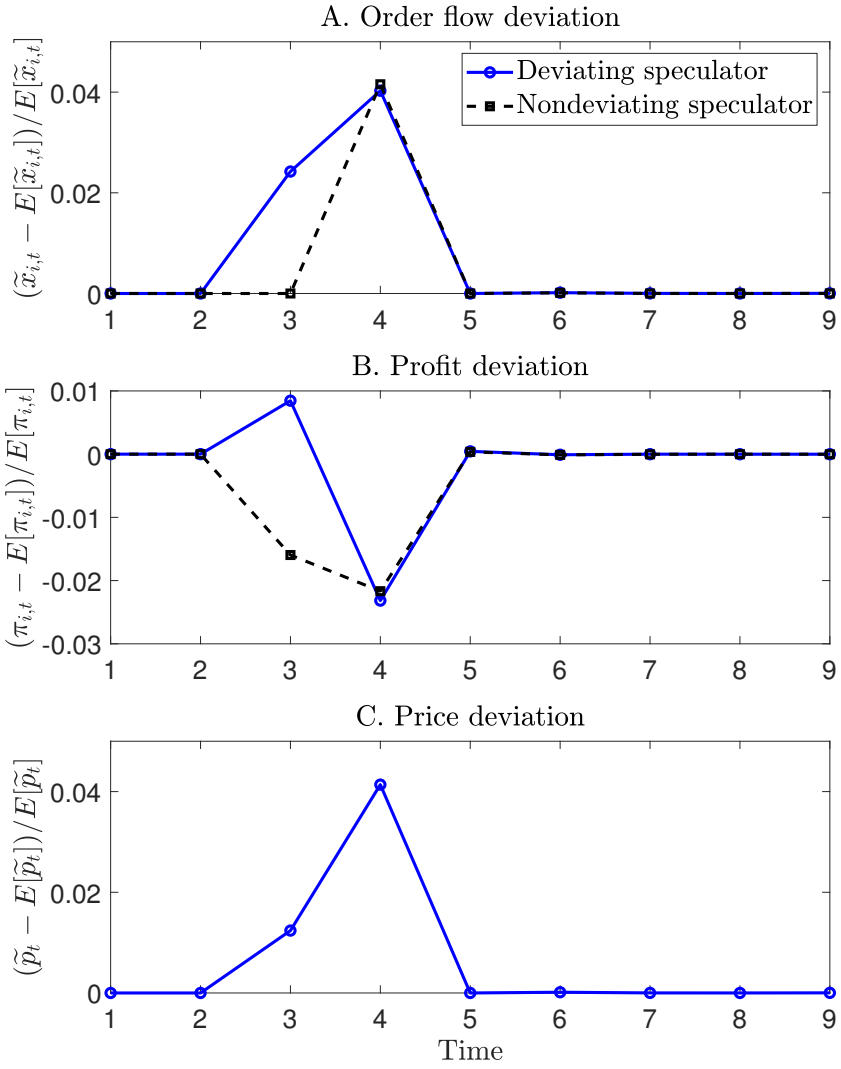
In this section, we complement the experiments in Section 5.1.1 by further studying the IRF after a unilateral deviation by one of the informed AI speculators. Specifically, in each of the $N = 1,000$ simulation sessions, we focus on the economic environment after informed speculators' Q-learning algorithms converge. Throughout the IRF experiment, for all $t \geq 0$, both informed AI speculators play their learned optimal strategies and the asset's price $p_t$ is determined by the market maker according to its learned pricing rule. In period $t = 3$, we exogenously force one informed AI speculator $i$ to have a one-period deviation from its learned optimal strategy. The one-period deviation in period $t = 3$ is made to the direction that increases the contemporaneous profit of the deviating speculator (i.e., we exogenously increase the deviating speculator's order by $\Delta x_{i,t}$ if $v_t > \overline{v}$ and reduce its order by $\Delta x_{i,t}$ if $v_t < \overline{v}$). We choose the deviation size $\Delta x_{i,t}$ to be one grid point in the order space $\mathbb{X}$, which ensures that the resulting price deviation is similar to the medium deviation case in panel A of Figure 1 for comparison purposes.

Panel A of Figure 3 plots the order flow's percentage deviation for both the deviating speculator and the nondeviating speculator. In period $t = 3$, on average, the deviating speculator's order flow deviates from the long-run mean by 2.5% while the nondeviating speculator's order flow remains unchanged. In period $t = 4$, the deviation gets punished as the nondeviating speculator behaves more aggressively, deviating its order flow from the long-run mean by 4.2%.

Rather than reducing its order flow, the deviating speculator further increases its order flow to 4.1% of the long-run mean in period $t = 4$, slightly below that of the nondeviating speculator. This form of overshooting exists for small deviations. As shown in panel A of Figure 5, if we consider a larger deviation, the deviating speculator would reduce its order flow in period $t = 4$. Regardless of whether its a small or a large deviation, both informed AI speculators abruptly return to the predeviation level of order flows in period $t = 5$.

Panel B of Figure 5 plots the profit's percentage deviation from its long-run mean for each informed AI speculator. In period $t = 3$, the deviating speculator's profit increases by 0.8% of the long-run mean while the nondeviating speculator's profit decreases by 1.6%. In period $t = 4$, due to the punishment strategy implemented by the nondeviating speculator, the profit of the deviating speculator drops substantially from 0.8% to $-2.4\%$ of the long-run mean. The expected discounted profit of deviation is about $-1.6\%$ of the long-run mean for the deviating speculator, indicating that deviation from the learned optimal strategies is not profitable.

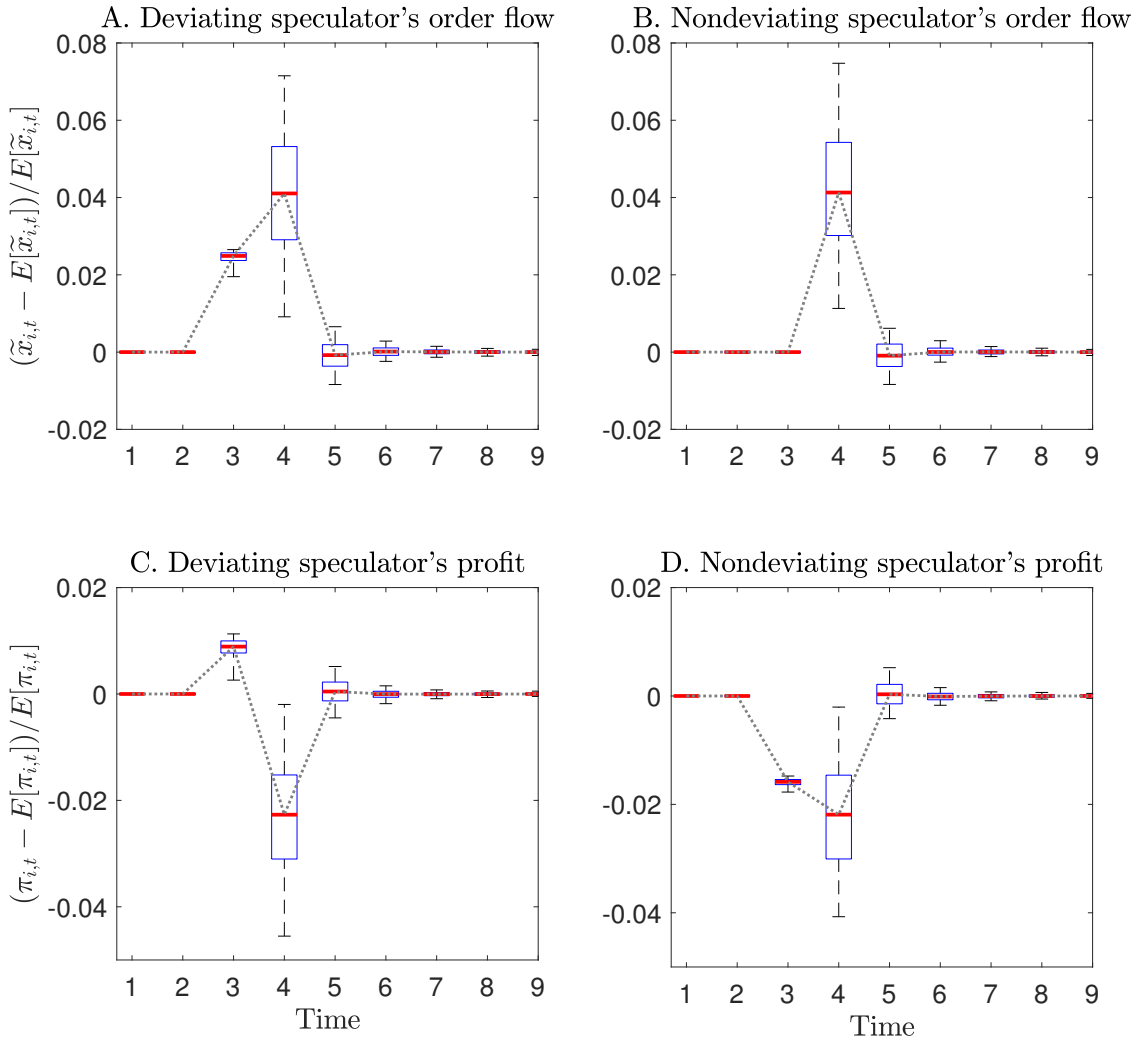
Panel C of Figure 3 plots the price's percentage deviation from its long-run mean. In period $t = 3$, due to the order deviation by one of the informed AI speculators, the asset's price deviates from its long-run mean by 1.2%. In fact, this is the force that triggered both informed AI speculators to change their order flows in period $t = 4$ because $p_{t-1}$ is the only state variable that




Note: In each simulation session, we focus on the economic environment after informed speculators' Q-learning algorithms converge. Throughout the IRF experiment, for all $t \geq 0$, both informed AI speculators play their learned optimal strategies and the asset's price $p_t$ is determined by the market maker according to its learned pricing rule. In period $t = 3$, we exogenously force one informed AI speculator $i$ to have a one-period deviation from its learned optimal strategy. The one-period deviation in period $t = 3$ is made to the direction that increases the contemporaneous profit of the deviating speculator (i.e., we exogenously increase the deviating speculator's order flow by $\Delta x_{i,t}$ if $v_t > \overline{v}$ and reduce its order flow by $\Delta x_{i,t}$ if $v_t < \overline{v}$). The deviation size $\Delta x_{i,t}$ is one grid point in the order space $\mathbb{X}$. Panels A and B plot the percentage deviation of profit and order flow from its long-run mean for both informed AI speculator, respectively. Panel C plots the price's percentage deviation from its long-run mean. All curves are average values across $N = 1,000$ simulation sessions, where each session is independently simulated 10,000 times to smooth out the effect of random shocks to $v_t$ and $u_t$. We set $\sigma_u/\sigma_v = 10^{-1}$. The other parameters are set according to the baseline economic environment described in Section 4.7.

Figure 3: IRF after a unilateral deviation ($\sigma_u/\sigma_v = 10^{-1}$).

records the deviation status in the preceding period $t = 3$. The asset's price continues to increase to 4.2% in period $t = 4$ because of the overshooting in the deviating speculator's order flow, and




Note: The experiment is similar to that described for Figure 3. Panels A and B plot the two speculators' order flow's percentage deviation from the long-run mean, and panels C and D plot their profit's percentage deviation from the long-run mean. In each panel, the dotted line represents the median value, the boxes represent the 25th and 75th percentiles, and the dashed intervals represent the 5th and 95th percentiles across $N = 1,000$ simulation sessions. Parameters are set as in Figure 3.

Figure 4: Confidence intervals for the IRF after a unilateral deviation ($\sigma_u/\sigma_v = 10^{-1}$).

then abruptly returns to the long-run mean in period $t = 5$ as the two informed AI speculators revert to their predeviation behavior.

Figure 4 plots the distribution of the IRF across the $N = 1,000$ simulation sessions and shows that the deviating speculator gets punished through price-trigger strategies in most simulation sessions. To further show robustness, in panels A to C of Figure 5, we present the IRF of a larger deviation by setting $\Delta x_{i,t}$ equal to three grid points in the order space $\mathbb{X}$. The nondeviating speculator still implements a punishment strategy by substantially increasing its order flow in

period $t = 4$ to punish the deviating speculator's defect in period $t = 3$. The expected discounted profit of deviation is negative for the deviating speculator. In panels D to F of Figure 5, we present the IRF in an economic environment with higher noise trading risks by setting $\sigma_u/\sigma_v = 1$. In this environment, the two informed AI speculators achieve a small amount of supra-competitive profits with an average value of $\Delta^C = 0.2$. Even with such a low level of supra-competitive profits, we still see that the nondeviating speculator implements price-trigger strategies to deter deviations. However, the magnitude of both deviations and punishments in panels D to F of Figure 5 are smaller than those in Figure 3. This is consistent with a lower average $\Delta^C$ and the theoretical insight that collusive behavior becomes more difficult to achieve when informed AI speculators are less able to monitor peers' deviations due to the larger information asymmetry caused by higher noise trading risks.

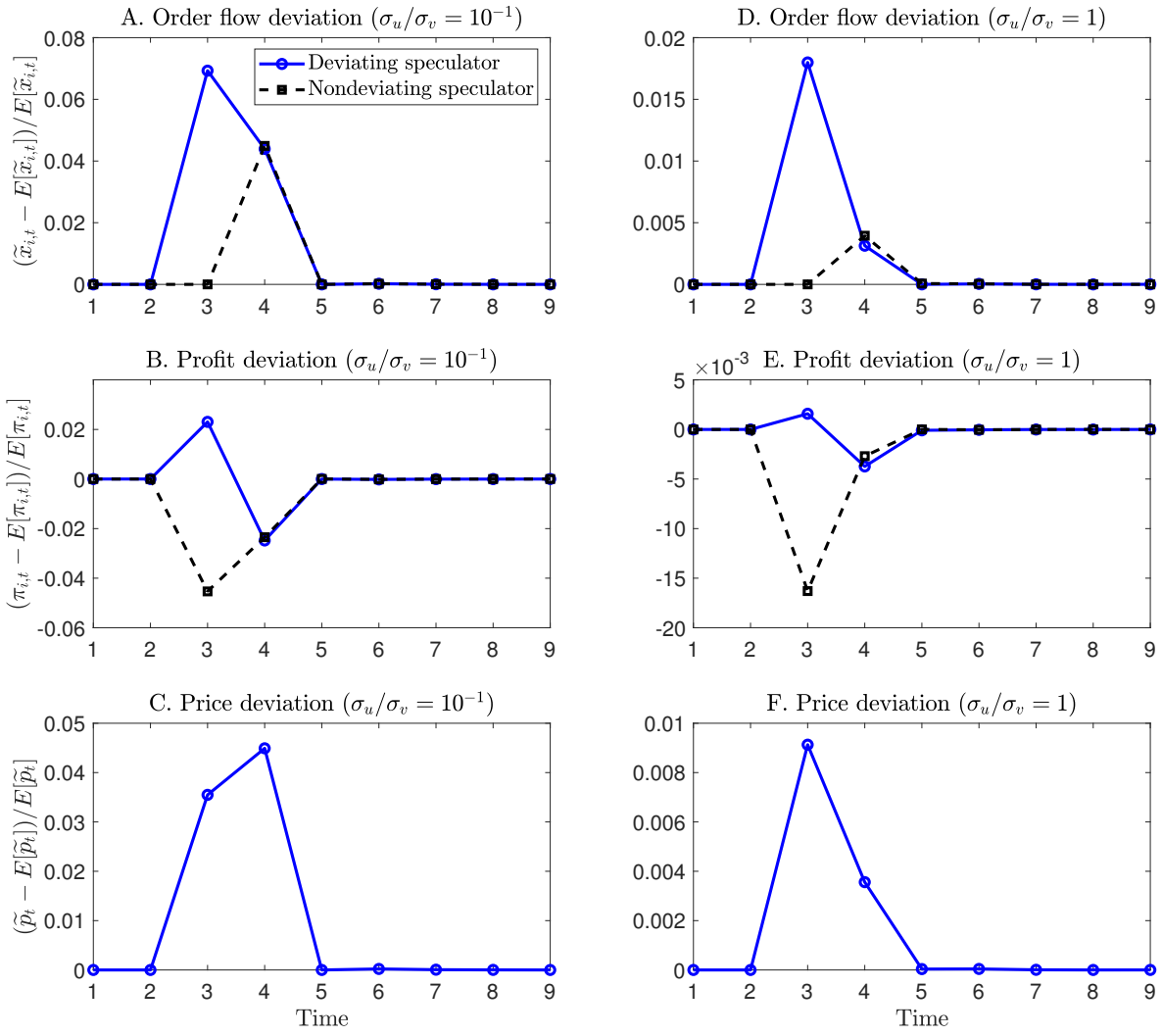
### 5.1.3 Discussions

Except for the duration of punishment, the impulse responses presented in Figures 1, 3 and 5 are quite consistent with the price-trigger strategies described in our model in Section 3. The patterns observed in our experiments coincide with our theoretical predictions that when the environment has low price efficiency and low noise trading risks, informed AI speculators are able to collude with each other by adopting price-trigger strategies to deter deviations. Moreover, collusion is more difficult to attain as noise trading risks become large.

Q-learning algorithms can learn price-trigger strategies because of experimentations. When one informed AI speculator switches to the exploration mode in the process of learning, it would choose actions randomly. Such behavior is effectively similar to defect from an implicit collusive agreement, if any. When this occurs, the two informed AI speculators would be trapped in the punishment phase until further explorations by one or both informed AI speculators occur. Informed AI speculators are able to learn coordination strategies because exploration modes will eventually stop, a necessary condition for Q-learning algorithms to converge.

Our finding that informed AI speculators are able to learn price-trigger strategies is similar to the finding of Calvano et al. (2020) that informed AI speculators learn grim-trigger strategies to sustain collusion in a perfect-information environment with Bertrand competition. However, different from Calvano et al. (2020), after punishment in period $t = 4$, rather than gradually returning to predeviation behavior, the informed AI speculators in our experiments abruptly return to their predeviation behavior. This difference is mainly due to the information asymmetry introduced by noise trading risks (i.e., $\sigma_u > 0$) and the stochastic asset value (i.e., $\sigma_v > 0$). Both model ingredients make informed AI speculators more difficult to sustain collusion by punishment threat, not just in the simulation experiments with informed AI speculators, but also in the model with rational-expectation informed speculators in Section 3.

In particular, our economic environment differs from that of Calvano et al. (2020) in two main aspects. First, we consider a stochastic environment where the asset's value $v_t$ in each period is drawn from an i.i.d. distribution. In this stochastic setting, it becomes more difficult




Note: The experiment is similar to that described for Figure 3. The left three panels consider a larger deviation by setting $\Delta x_{i,t}$ equal to three grid points in the order space $\mathbb{X}$. The right three panels consider an economic environment with higher noise trading risks by setting $\sigma_u/\sigma_v = 1$. The other parameters are set according to the baseline economic environment described in Section 4.7.

Figure 5: Robustness of IRF: larger deviation or higher noise trading risks ($\sigma_u/\sigma_v = 1$).

for the two informed AI speculators to learn punishment strategies to sustain collusion than in the deterministic setting with a constant $v_t$.[14] Second, the noise trader's random order flows generate information asymmetry to informed AI speculators, which makes grim-trigger strategies infeasible. As a result, informed AI speculators have to adopt price-trigger strategies to collude. In both the model with rational-expectation informed speculators and the simulation experiments

[14] In one of the robustness checks, Calvano et al. (2020) consider stochastic demand and show that the average $\Delta^C$ is lower when aggregate demand can take two values randomly. We also find that with stochastic $v_t$, the average $\Delta^C$ declines because it is more difficult for Q-learning algorithms to learn strong punishment strategies. The decline in $\Delta^C$ would be smaller if the evolution of $v_t$ exhibits a smaller degree of randomness, either through a higher level of persistence or a less dispersed distribution.

with informed AI speculators, the ratio $\sigma_u/\sigma_v$ plays a crucial role in determining the level of collusion.

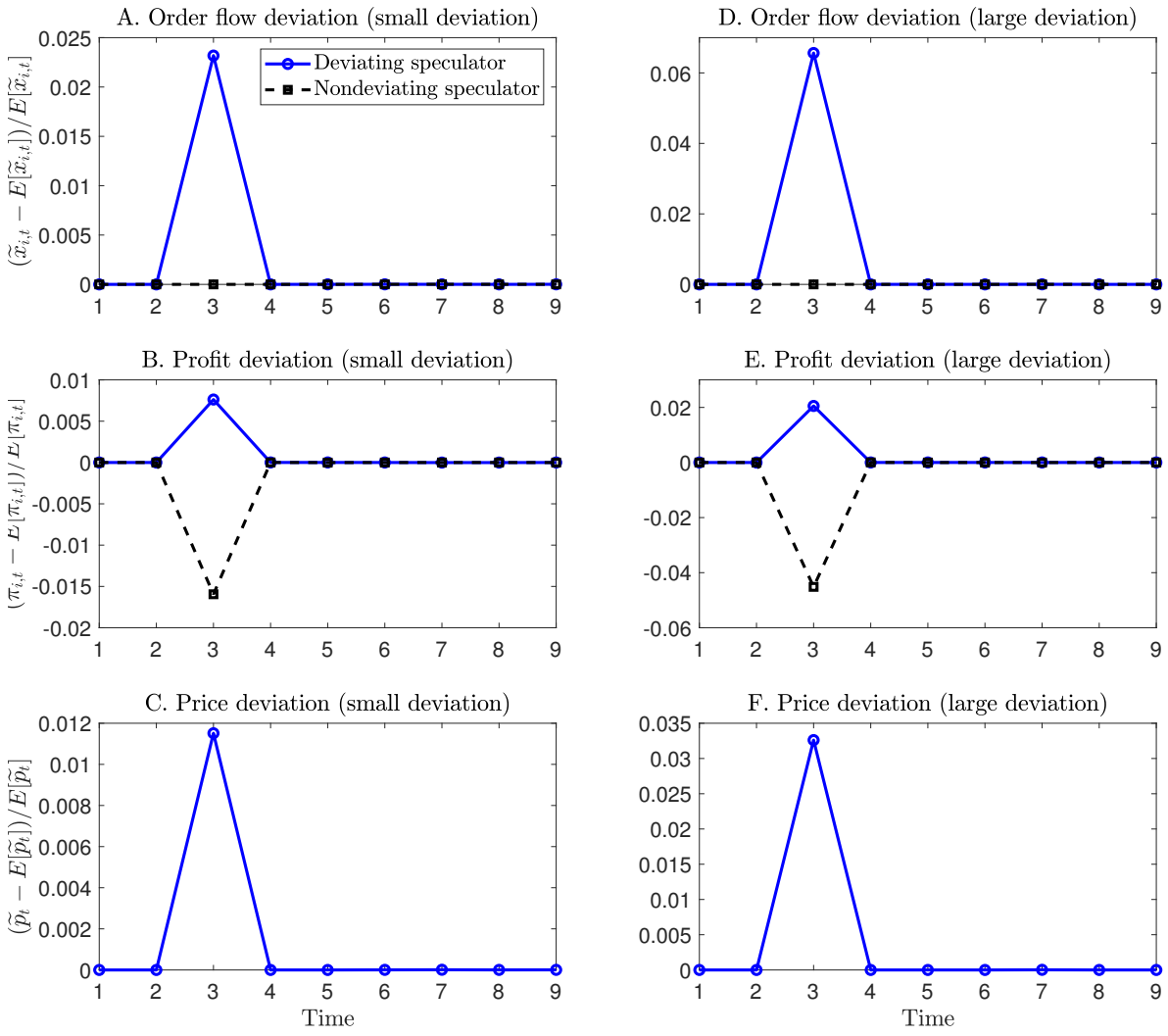
The information asymmetry in our economic environment implies that peer informed AI speculators' lagged actions are unobservable and thus cannot be included as state variables. Thus, as described in Section 4.1, we use the lagged asset's price $p_{t-1}$ as the state variable in period $t$, rather than the lagged actions of the two informed AI speculators. Compared to our baseline setting with state variables $s_t = \{p_{t-1}, v_t\}$, we also examine the settings with alternative specifications of state variables. First, we consider a counterfactual setting with state variables $s_t = \{x_{i,t-1}, x_{-i,t-1}, v_t\}$. This setting essentially assumes that informed AI speculators' can perfectly observe peers' order flows, which is close to the perfect-information setting of Calvano et al. (2020) except for including $v_t$ as an additional state variable. Second, we consider the setting where state variables are $s_t = \{p_{t-1}, x_{i,t-1}, v_t\}$. We find that under the perfect information benchmark (i.e., $\sigma_u/\sigma_v = 0$) with two informed AI speculators, these two alternative settings have almost the same average $\Delta^C$. This is not surprising because under the perfect information benchmark, recording $x_{i,t-1}$ and $p_{t-1}$ allows each informed AI speculator to exactly back out its peer's order flow $x_{-i,t-1}$. However, with information asymmetry (i.e., $\sigma_u/\sigma_v > 0$), the first setting with $s_t = \{x_{i,t-1}, x_{-i,t-1}, v_t\}$ yields a considerably higher average $\Delta^C$ than the other setting with $s_t = \{p_{t-1}, x_{i,t-1}, v_t\}$. In addition, we find that the average $\Delta^C$ in these two alternative settings is higher than that in our baseline setting. Thus, incorporating informed AI speculators' lagged actions as additional state variables indeed helps informed AI speculators to learn collusive strategies, likely through an improved learning of punishment strategies. However, lagged actions are not a necessary ingredient because in both our model with rational-expectation informed speculators and simulation experiments with informed AI speculators, including lagged price $p_{t-1}$ alone can already result in a significant degree of collusion.

### 5.2 Artificial Stupidity: Collusion through Homogenized Learning Biases

In this subsection, we study informed AI speculators' learned optimal strategies when the environment has low price efficiency but large noise trading risks (i.e., $\sigma_u/\sigma_v = 10^2$). Similar to Section 5.1, the other parameters are set according to the baseline economic environment.

According to our model in Section 3, it is impossible for informed speculators to collude with each other in environments with large noise trading risks. However, in our simulation experiments, informed AI speculators can still achieve supra-competitive profits. Across $N = 1,000$ simulation sessions, the average value of $\Delta^C$ is about 0.6 and the average profit of informed AI speculators is about 7.5% higher than the profit in the noncollusive equilibrium. The profit becomes even higher as noise trading risks further increase. Below, we examine the mechanism that leads to such supra-competitive profits. We show that in line with our model's prediction, informed AI speculators do not learn price-trigger strategies to sustain collusion. Instead, they are able to collude to achieve supra-competitive profits due to homogenized learning biases.

To begin with, we study the impulse responses to a unilateral deviation in Figure 6. Clearly,




Note: The experiment is similar to those described for Figure 3, except for setting $\sigma_u/\sigma_v = 10^2$. The left three panels consider a unilateral small deviation with deviation size $\Delta x_{i,t}$ equal to one grid point in the order space $\mathbb{X}$. The right three panels consider a unilateral large deviation with $\Delta x_{i,t}$ equal to three grid points in $\mathbb{X}$.

Figure 6: IRF after a unilateral deviation ($\sigma_u/\sigma_v = 10^2$).

regardless of whether it is a small deviation (panels A to C) or a large deviation (panels D to F), we do not see any punishment from the nondeviating speculator. Instead, panels A and D of Figure 6 show that the nondeviating speculator's order flow is virtually unchanged and the deviating speculator returns to its learned optimal trading strategy immediately in period $t = 4$, which is just one period after the deviation. Panels B and E of Figure 6 show that the deviating speculator obtains an extra amount of one-period profit in period $t = 3$, which causes a one-period profit loss for the nondeviating speculator. Because there is no punishment for $t \geq 4$, the average percentage gains from the deviation in terms of discounted profits is strictly positive for the deviating speculator.

### 5.2.1 Self-Confirming Equilibrium

The collusive outcomes achieved by the two informed AI speculators are clearly not generated by price-trigger strategies when $\sigma_u/\sigma_v$ is large, which is consistent with the prediction of our model (Proposition 3.4). In fact, the collusive outcomes are achieved through homogenized learning biases of informed AI speculators when noise trading risks are large. Although deviation seems to be profitable in terms of increasing the discounted profits, both informed AI speculators choose not to do this according to their learned optimal trading strategies after their Q-learning algorithms converge. The reason is that informed AI speculators' actions are governed by their learned Q-matrix, which indicates that the (no-deviation) strategies they are playing are optimal and any deviations cannot be profitable.

The steady-state behavior of informed AI speculators represents a self-confirming equilibrium, a notion first introduced by Fudenberg and Levine (1993). Compared with the Nash equilibrium, the self-confirming equilibrium is weaker because it allows players to have incorrect (or biased) off-equilibrium beliefs. This equilibrium concept is motivated by the idea that noncooperative equilibria should be interpreted as outcomes of a learning process, in which players form beliefs based on their past experience. While beliefs can be generally correct along the equilibrium path of play because it is frequently observed, beliefs are not necessarily correct off the equilibrium path unless players engage in a sufficient amount of experimentation with non-optimal actions (e.g., Fudenberg and Kreps, 1988, 1995; Cho and Sargent, 2008). Importantly, the incorrect off-equilibrium beliefs are not inconsistent with the evidence (i.e., outcomes along the equilibrium path). As noted by Fudenberg and Levine (1993), any self-confirming equilibrium can be a steady state, especially, including those equilibria with outcomes that cannot arise in Nash equilibrium. The self-confirming equilibrium allows completely arbitrary beliefs and supposes that players do not think strategically like what they do in a rational expectations framework. Instead, players choose actions based on what they have learned from their past experience.

In our simulations, informed AI speculators' beliefs are summarized by their Q-matrices. Specifically, the value of each state-action pair $(s, x)$ in the Q-matrix represents the "perceived" reward that the informed AI speculator can obtain by playing the action $x \in \mathcal{X}$ in the state $s \in S$.[15] In Appendix G.1, we show that the hyperparameter $\alpha$, which determines the informed AI speculator's forgetting rate or memory capacity, plays a crucial role in determining the magnitude of learning biases. Unbiased learning about the Q-matrix requires two conditions to hold simultaneously: 1) the informed AI speculators have sufficiently experimented all possible off-equilibrium plays before Q-learning algorithms converge, and 2) informed AI speculators' memory capacity is infinitely large, i.e., $\alpha \to 0$. As long as $\alpha > 0$, the Q-matrix is learned with biases due to the failure of the law of large numbers. Moreover, learning biases are larger when noise trading risks are higher (i.e., higher $\sigma_u/\sigma_v$) or the forgetting rate $\alpha$ is higher. Intuitively, informed AI speculators average past data to approximate the moments of the conditional probability

[15] As we show in Appendix G.1, when $\rho = 0$, the value of each state-action pair $(s, x)$ in the Q-matrix is equal to the sum of the discounted value of the profits $(v - p)x$ received by the informed AI speculator when it played $x$ in state $s$ in the past, with the discount rate being $1 - \alpha$.

distribution of interest. When the environment's has higher noise trading risks or the forgetting rate $\alpha$ is higher, informed AI speculators lack sufficient memory capacity to store and analyze past data, and thus it becomes more difficult to approximate the moments of interest (i.e., the Q-matrix). The magnitude of learning biases in turn will determine which self-confirming equilibrium would emerge after Q-learning algorithms converge.

#### 5.2.2 Biased Learning Leads to Self-Confirming Equilibrium with Supra-Competitive Profits

Having discussed that the steady state reached by informed AI speculators represents a self-confirming equilibrium, we now further explain why informed AI speculators' biased learning leads to collusive rather than competitive outcomes.

The underlying logic involves the following four key steps. First, collusive outcomes are achieved when informed AI speculators adopt more conservative, rather than more aggressive trading strategies. Specifically, according to our model in Section 3, the sensitivities of informed speculators' order flow to the asset's value $v_t$ in different equilibria satisfy $\chi^M \leq \chi^C < \chi^N$. Because informed speculator $i$'s order $x_{i,t}$ is $x_{i,t} = \chi(v_t - \overline{v})$, its absolute value of order flow satisfies $|x^M_{i,t}| \leq |x^C_{i,t}| < |x^N_{i,t}|$ for any $v_t$, indicating that collusion means that informed speculators adopt more conservative (i.e., trading with smaller absolute value of order flow $|x_{i,t}|$), rather than more aggressive trading strategies.

Second, compared with more conservative trading strategies, when informed AI speculators adopt more aggressive trading strategies, the unconditional variance of per-period profits is larger, namely, the distribution of per-period profits is more dispersed. Specifically, in Appendix G.2, we show that, for any state $s$, there exists complementarity between an informed AI speculator's order flow $x$ and the noise order flow $u_t$ in determining per-period profits. This complementarity implies that more aggressive trading strategies would amplify the impact of the noise order flow $u_t$, generating a more dispersed distribution of per-period profits compared to that generated by more conservative trading strategies.

Third, if playing action $x$ in state $s$ generates a more dispersed distribution of per-period profits, the resulting estimated Q value, $\widehat{Q}_t(s,x)$, for the state-action pair $(s,x)$ also has a more dispersed distribution over time. This is because at any point in time $t$, the estimated $\widehat{Q}_t(s,x)$ is the sum of the discounted value of per-period profits that the informed AI speculator receives when it visits the state-action pair $(s,x)$ in the past.

Fourth, a necessary condition for all Q-learning algorithms to reach stationary optimal strategies is that exploration rate is virtually zero, and informed AI speculators are purely in the exploitation mode. However, because of exploitation, for any state $s$, the action $x$ that generates a more dispersed distribution of $\widehat{Q}_t(s,x)$ over time is less likely to be adopted by informed AI speculators after their Q-learning algorithms converge. Specifically, relative to playing conservative actions, playing an aggressive action (denoted by $x^*$), generates a dispersed distribution of $\widehat{Q}_t(s,x^*)$ over time. This means that an aggressive action $x^*$ is likely to generate both a high $\widehat{Q}_t(s,x^*)$ and a low $\widehat{Q}_t(s,x^*)$. In one case, suppose a sequence of unfavorable noise order flows

were realized when the informed AI speculator was playing $x^*$ in state $s$, so that a low $\widehat{Q}_t(s, x^*)$ is estimated for $x^*$. Then, $x^*$ will not be played when the informed AI speculator conducts exploitation in state $s$ in the future, because this action obviously does not maximize its Q value. In the other case, suppose a sequence of favorable noise order flows were realized when the informed AI speculator was playing $x^*$ in state $s$, so that a high $\widehat{Q}_t(s, x^*)$ is estimated for $x^*$. Then, $x^*$ will be further "exploited" in future periods. Because $x^*$ generates a more dispersed $\widehat{Q}_t(s, x^*)$, it is highly likely that, eventually, the estimated $\widehat{Q}_t(s, x^*)$ will be small. From this point on, like the first case, the informed AI speculator will not play $x^*$ when conducting future exploitation in state $s$. Thus, in the process of reaching convergence, the informed AI speculator's exploitation has the tendency to not adopt the trading strategies that can possibly generate large negative Q values, which are aggressive trading strategies that generate a more dispersed distribution of per-period profits. In some sense, informed AI speculators exhibit a certain degree of aversion to risks in the exploitation mode.

Taking the above four steps together, informed AI speculators' biased learning leads them to adopt more conservative trading strategies after their Q-learning algorithms converge, resulting in collusive outcomes.

#### 5.2.3 Homogenized Bias and Implicit Coordination

We have explained how informed AI speculators' learning biases and exploitation lead to a self-confirming equilibrium that features collusive outcomes. However, it remains unclear why informed AI speculators adopt highly similar trading strategies after their Q-learning algorithms converge. What is the fundamental force that generates this sort of implicit coordination? We find that the key reason is that informed AI speculators rely on the same foundational model in their learning process. This generates homogenized learning biases, eventually leading to implicit coordination.

To elaborate, first consider the economic environment represented by the trough point of the blue solid line in panel A of Figure 7, i.e., $\log(\sigma_u/\sigma_v) = 2$. This represents an environment with high price inefficiency but relatively low noise trading risks in the sense that learning biases are small for informed AI speculators. However, noise trading risks are large enough to rule out the existence of a collusive equilibrium sustained by price-trigger strategies. Because learning biases are small in this environment, informed AI speculators are able to learn to play a noncollusive Nash equilibrium after their Q-learning algorithms converge, resulting in an average $\Delta^C \approx 0$. Implicit coordination in this environment is achieved because both informed AI speculators adopt similar noncollusive trading strategies in the Nash equilibrium.

Next, suppose that noise trading risks in the economic environment become higher, all else equal, both informed AI speculators become more biased in their learning processes. This leads both of them to optimally choose more conservative trading strategies after their Q-learning algorithms converge. Because both informed AI speculators adopt the same Q-learning algorithm with the same forgetting rate $\alpha$, the magnitudes of their learning biases are similar. Thus, they

also become more conservative at a similar pace, resulting in similar optimal trading strategies after their Q-learning algorithms converge, as if they are implicitly coordinating with each other. The homogenized bias in informed AI speculators' Q-learning algorithms allows them to attain similar levels of supra-competitive profits. The extent to which informed AI speculators are biased homogeneously determines the implicitly coordinated level of profits. Importantly, as noted above, the two informed AI speculators reach a self-confirming equilibrium in which no one will deviate, because their biased beliefs, as recorded in their learned Q-matrices, suggest that any deviation cannot be profitable.

By contrast, if the two informed AI speculators' learning processes are not governed by the same foundational model, the learning biases will not be homogeneized. As a result, the two informed AI speculators may not be able to simultaneously attain supra-competitive profits. As an illustrative example, in panel B of Figure 15, we consider an experiment in which one informed AI speculator adopts a more advanced algorithm than the other, as captured by a lower forgetting rate $\alpha$. We find that the more advanced informed AI speculator is able to attain much higher profits than in the experiment with two informed AI speculators adopting the same $\alpha$. However, the average profit of the less advanced informed AI speculator is much lower and similar to the profit in the noncollusive Nash equilibrium. In about half of the $1,000$ simulation sessions, the profits of the less advanced informed AI speculator are even lower than the profit in the noncollusive Nash equilibrium. This experiment highlights the importance of homogenized bias in generating implicit coordination and supra-competitive profits for all informed AI speculators. Further, in Section 7.2, we extend the Q-learning algorithm to a two-tier Q-learning algorithm in which informed AI speculators learn both the optimal choice of the forgetting rate $\alpha$ and the optimal trading strategies corresponding to the choice of $\alpha$. Interestingly, we find that informed AI speculators will learn to coordinately adopt high values of $\alpha$ in the stationary equilibrium, and such coordination allows both of them to obtain supra-competitive profits through homogenized learning baises.

#### 5.2.4 Determinants of the Magnitude of Learning Biases

The extent to which learning is biased determines which self-confirming equilibrium would emerge after Q-learning algorithms converge, which consequently determines the average profits of informed AI speculators. Specifically, the above mechanism is stronger when informed AI speculators' Q-matrices are estimated with larger biases. Thus, the extent to which informed AI speculators collude to attain supra-competitive outcomes increases with the magnitude of learning biases. We now discuss the determinants of the magnitude of learning biases.

As noted in Section 5.2.1, learning biases are larger in environments with higher noise trading risks (i.e., higher $\sigma_u/\sigma_v$) or when informed AI speculators have a higher forgetting rate $\alpha$. In equation (G.6) in Appendix G.1, we formally show that the magnitude of learning biases increases when $\sigma_u/\sigma_v$ is higher, $\lambda$ is higher, $\rho$ is lower, or $\alpha$ is higher. These properties behind Q-learning algorithms predict that informed AI speculators can attain higher supra-competitive profits due

to biased learning when $\sigma_u/\sigma_v$ is higher, $\lambda$ is higher, $\rho$ is lower, or $\alpha$ is higher. Consistent these predictions, first, we show that the average $\Delta^C$ across $N = 1,000$ simulation sessions increases with $\sigma_u/\sigma_v$ in the region with high noise trading risks (i.e., $\log(\sigma_u/\sigma_v) \geq 2$) in panel A of Figure 7. Second, we show that in the environment with high noise trading risks (e.g., $\log(\sigma_u/\sigma_v) = 2$), reducing $\xi$ from 500 to 1 (which results in a larger $\lambda$ and higher price efficiency) leads to a higher average $\Delta^C$ in panel B of Figure 7. Third, we show that in the environment with high noise trading risks, reducing the value of $\rho$ leads to a higher average $\Delta^C$ in Figure 13. Finally, we show that in the environment with high noise trading risks, a higher $\alpha$ would result in a higher average $\Delta^C$ in panel B of Figure 14.

## 5.3 Role of Noise Trading Risks and Price Efficiency

In this subsection, we study the role of noise trading risk and price efficiency in generating collusive outcomes for informed AI speculators.

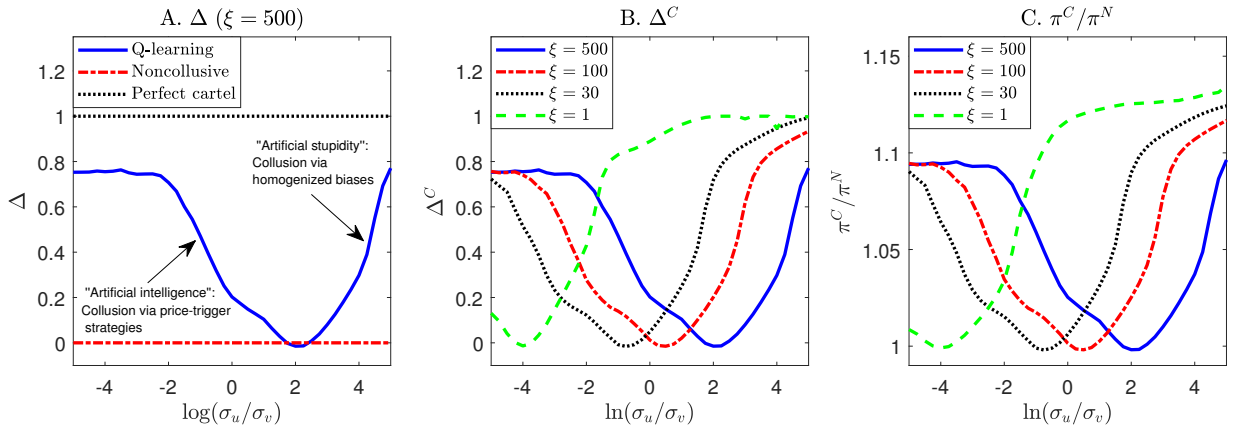
### 5.3.1 Role of Noise Trading Risks

Consider the baseline economic environment described in Section 4.7. In panel A of Figure 7, we plot the average $\Delta^C$ as $\log(\sigma_u/\sigma_v)$ varies from $-5$ to 5 along the x-axis. The black dotted and red dash-dotted lines represent the theoretical benchmarks ($\Delta^M = 1$ and $\Delta^N = 0$) in the perfect cartel and noncollusive Nash equilibrium, respectively. The blue solid line plots the average $\Delta^C$ across $N = 1,000$ simulation sessions, holding all other parameters unchanged. It shows that as $\log(\sigma_u/\sigma_v)$ increases along the x-axis, the average $\Delta^C$ first decreases and then increases. This U-shape pattern is an outcome of the interaction of the two mechanisms discussed in Sections 5.1 and 5.2. Specifically, in the region of low noise trading risks, i.e., $\log(\sigma_u/\sigma_v) < 2$, the average $\Delta^C$ is decreasing in $\log(\sigma_u/\sigma_v)$. In this region, informed AI speculators learn price-trigger strategies to sustain collusion and attain supra-competitive profits, as discussed in Section 5.1. The negative relationship between the average $\Delta^C$ and $\log(\sigma_u/\sigma_v)$ observed in our simulation experiments is consistent with the prediction of our model (see Proposition 3.6.(ii)).

In the region of large noise trading risks, i.e., $\log(\sigma_u/\sigma_v) \geq 2$, the average $\Delta^C$ is increasing in $\log(\sigma_u/\sigma_v)$. In this region, informed AI speculators attain supra-competitive profits because of homogenized learning biases, as discussed in Section 5.2. The positive relationship between the average $\Delta^C$ and $\log(\sigma_u/\sigma_v)$ observed in our simulation experiments is consistent with the theoretical property that biased learning becomes more significant when $\log(\sigma_u/\sigma_v)$ increases (see Section 5.2.4).

### 5.3.2 Role of Price Efficiency

According to our model in Section 3, the market maker focuses more on minimizing pricing errors when $\xi$ is small or $\theta$ is large. In this case, price efficiency is high and there does not exists collusive Nash equilibrium sustained by price-trigger strategies for any $\sigma_u/\sigma_v > 0$ (Proposition 3.3). By




Note: This figure plots the average $\Delta^C$ and the profit gain relative to noncollusion ($\pi^C/\pi^N$) across $N = 1,000$ simulation sessions as $\log(\sigma_u/\sigma_v)$ varies along the x-axis, for different values of $\xi = 500, 100, 30, 1$. The other parameters are set according to the baseline economic environment described in Section 4.7.

Figure 7: $\Delta^C$ and $\pi^C/\pi^N$ for $\log(\sigma_u/\sigma_v) \in [-5, 5]$ and $\xi = 500, 100, 30, 1$.

contrast, when $\xi$ is large or $\theta$ is small, the market maker focuses more on minimizing inventory costs. In this case, price efficiency is low and there exists a collusive Nash equilibrium that can be sustained by price-trigger strategies for small $\sigma_u/\sigma_v$ and $I$ (Proposition 3.4).

By varying the value of $\xi$ in our simulation experiments, we study how price efficiency affects informed AI speculators' trading profits.[16] Specifically, the four curves in panel B of Figure 7 represent the experiments with $\xi = 500, 100, 30$ and 1. The overall U-shaped relationship between the average $\Delta^C$ and $\log(\sigma_u/\sigma_v)$ is not peculiar to the choice of $\xi$. All four curves display U-shape patterns. Panel C of Figure 7 plots the profit gain relative to noncollusion ($\pi^C/\pi^N$), the pattern is similar to that in panel A.

As we compare the four curves in panel B of Figure 7, one salient feature is that the trough of the U-shape shifts to the left as $\xi$ decreases. This suggests that with a smaller $\xi$, a lower level of noise trading risks is necessary for informed AI speculators to learn price-trigger strategies to collude. A similar point can be made if we focus on the region with low noise trading risks, in which price-trigger strategies are learned by informed AI speculators. For example, holding $\ln(\sigma_u/\sigma_v) = -4$ unchanged, it is clear that the average $\Delta^C$ declines monotonically as $\xi$ decreases from 500 to 1. Thus, collusion becomes more difficult to achieve through price-trigger strategies as $\xi$ decreases, as predicted by our model (see Proposition 3.6.(iv)). By contrast, the relationship between $\xi$ and average $\Delta^C$ is opposite if we focus on the region with large noise trading risks, in which informed AI speculators' trading strategies are dominantly affected by learning biases. For example, holding $\ln(\sigma_u/\sigma_v) = 2$ unchanged, it is clear that the average $\Delta^C$ increases monotonically as $\xi$ decreases from 500 to 1. This is consistent with the theoretical property of biased learning discussed in Section 5.2.4, that is, the magnitude of learning biases increases with $\lambda$ (i.e., decreases with $\xi$). Thus, a lower $\xi$ leads to larger learning biases, allowing informed AI speculators to

[16] We do not conduct experiments with different $\theta$ because a smaller $\theta$ has similar impacts as a larger $\xi$ on price efficiency.

achieve higher supra-competitive profits.

### 5.4 Trading Strategy of Informed AI Speculators

In this subsection, we illustrate informed AI speculators' trading strategies in the baseline economic environment described in Section 4.7.

In panel A of Figure 8, we plot the average sensitivity of informed AI speculators' order to the asset's value, $\widehat{\chi}^C$, across $N = 1,000$ simulation sessions as a function of the noise trading risk $\log(\sigma_u/\sigma_v)$. Consistent with panel A of Figure 7, $\widehat{\chi}^C$ displays an inverted U-shape as $\log(\sigma_u/\sigma_v)$ increases along the x-axis. By contrast, the theoretical benchmarks $\chi^N$ and $\chi^M$ stay roughly unchanged as $\log(\sigma_u/\sigma_v)$ increases.

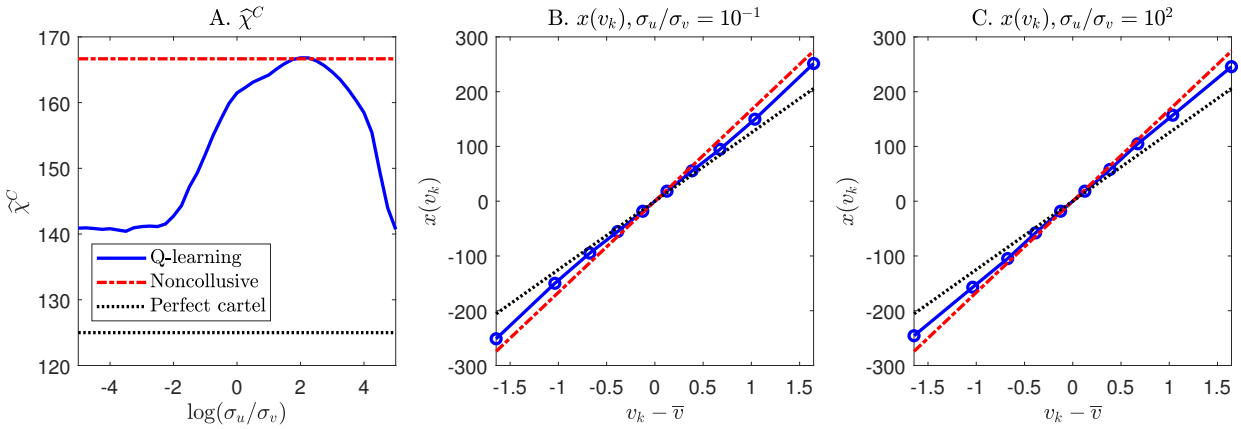
In fact, the estimated $\widehat{\chi}^C$ almost sufficiently describes informed AI speculators' trading strategy because their orders are almost linear in the assets's value, a property that holds both in the model and the simulation experiments. As an illustration, in panels B and C of Figure 8, we present the average trading strategy of informed AI speculators across $N = 1,000$ simulation sessions. Panel B is for the environment with low noise trading risks ($\sigma_u/\sigma_v = 10^{-1}$) and panel C is for the environment with high noise trading risks ($\sigma_u/\sigma_v = 10^2$). The trading strategy in each simulation session is calculated as $x(v_k) = \frac{1}{In_p}\sum_{i=1}^{I}\sum_{m=1}^{n_p} x_i(p_m, v_k)$, which is the average order flow of $I$ informed AI speculators across all grid points of $\mathbb{P}$, after Q-learning algorithms converge. The dots on the blue solid lines represent the average order flow corresponding to the discrete grid points of $\mathbb{V}$. The black dotted and red dash-dotted lines represent the theoretical benchmarks, $\chi^M(v_k - \overline{v})$ and $\chi^N(v_k - \overline{v})$, in the perfect cartel equilibrium and noncollusive Nash equilibrium, respectively.

It is clear that informed AI speculators learn an optimal trading strategy that is roughly linear in the asset's value after their Q-learning algorithms converge, even though the linearity restriction is not imposed during the learning process. Moreover, the slope of a linear fit for the trading strategy of informed AI speculators, i.e., $\widehat{\chi}^C$, lies between $\chi^M$ and $\chi^N$ in both panels B and C of Figure 8. Thus, the trading strategy learned by informed AI speculators is more conservatively than that in the noncollusive Nash equilibrium, which explains why informed AI speculators are able to attain supra-competitive profits.

### 5.5 Price Informativeness, Market Liquidity, and Mispricing

In this subsection, we study the impacts of AI collusion for price informativeness, market liquidity, and mispricing in financial markets. We show that AI collusion leads to lower price informativeness, lower market liquidity, and higher mispricing. The magnitude of such effects depends on the extent to which informed AI speculators collude with each other, which is largely determined by the noise trading risk $\sigma_u/\sigma_v$.

Panel A of Figure 9 plots the market's price informativeness relative to the theoretical benchmark of the perfect cartel equilibrium. By definition, the black dotted line shows that the relative price informativeness in the perfect cartel equilibrium is $\mathcal{I}^M/\mathcal{I}^M \equiv 1$. The red dash-dotted line




Note: Panel A plots the average $\widehat{\chi}^C$ across $N = 1,000$ simulation sessions as $\log(\sigma_u/\sigma_v)$ varies along the x-axis. Panels B and C plot the average trading strategy of informed AI speculators across $N = 1,000$ simulation sessions. The trading strategy in each simulation session is calculated as $x(v_k) = \frac{1}{In_p}\sum_{i=1}^{I}\sum_{m=1}^{n_p} x_i(p_m, v_k)$, which is the average order flow of $I$ informed AI speculators across all grid points of $\mathbb{P}$, after Q-learning algorithms converge. The dots on the blue solid lines represent the average order flow corresponding to the discrete grid points of $\mathbb{V}$. Panels A and B focus on the environments with low ($\sigma_u/\sigma_v = 10^{-1}$) and high ($\sigma_u/\sigma_v = 10^{2}$) noise trading risks, respectively. The other parameters are set according to the baseline economic environment described in Section 4.7.

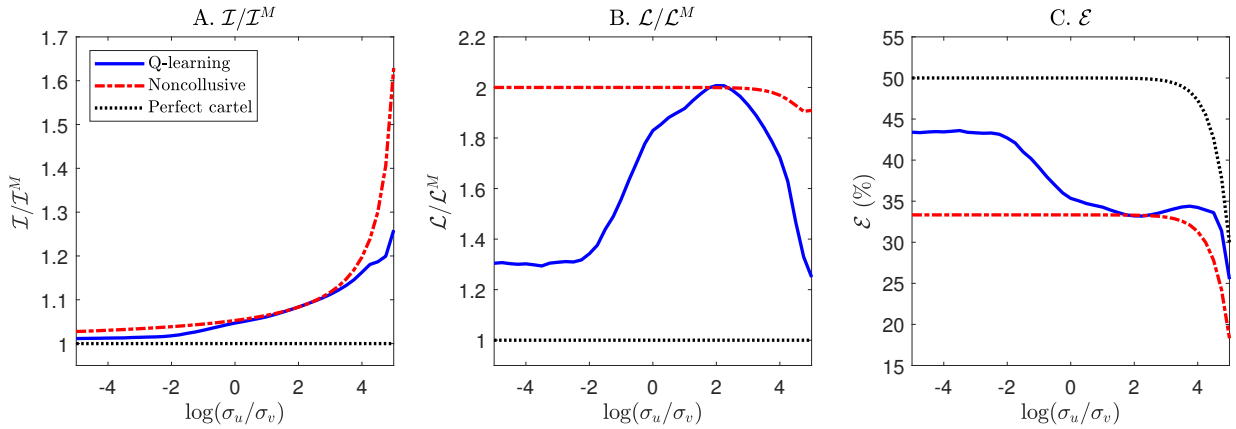
Figure 8: The trading strategy of informed AI speculators.

shows that the ratio of price informativeness in the theoretical benchmark of the noncollusive Nash equilibrium and perfect cartel equilibrium, $\mathcal{I}^N/\mathcal{I}^M$, is greater than 1 and increasing in $\log(\sigma_u/\sigma_v)$.[17] The blue solid line plots the average relative price informativeness, $\mathcal{I}^C/\mathcal{I}^M$, across $N = 1,000$ simulation sessions with informed AI speculators. Its value is close to the relative price informativeness in the theoretical benchmark of the non-collusive equilibrium when $\log(\sigma_u/\sigma_v)$ is around 2 due to the lack of collusion. When $\log(\sigma_u/\sigma_v)$ is very small or very large, the relative price informativeness in our simulation experiments with informed AI speculators is significantly lower than that in the theoretical benchmark of the noncollusive Nash equilibrium. The reason is that informed AI speculators place orders in a more conservative manner, with $\widehat{\chi}^C < \chi^N$, as shown in panel A of Figure 8.

Our findings suggest that perfect price informativeness is not achievable in the presence of informed AI speculators. In our simulation environments, when the noise trading risk $\sigma_u/\sigma_v$ decreases, informed AI speculators would withhold their private information about the asset's value and collude more through price-trigger strategies, placing orders more conservatively than what they would do in the noncollusive Nash equilibrium. This AI collusion reduces price informativeness. Crucially, informed AI speculators never need to communicate with each other, whether explicitly or implicitly, the adoption of Q-learning algorithms automatically leads to such collusive behavior.

Panel B of Figure 9 plots the market liquidity relative to the theoretical benchmark of the perfect cartel equilibrium. The red dash-dotted line shows that the ratio of market liquidity in

[17] This is because $\widehat{\chi}^N > \widehat{\chi}^M$ for all $\log(\sigma_u/\sigma_v)$. Moreover, when $\xi = 500$, $\widehat{\chi}^N$ and $\widehat{\chi}^M$ are roughly unchanged (only slightly increase) as $\log(\sigma_u/\sigma_v)$ increases. Then, according to the equation (4.10), both $\mathcal{I}^N$ and $\mathcal{I}^M$ are decreasing in $\log(\sigma_u/\sigma_v)$, but the ratio $\mathcal{I}^N/\mathcal{I}^M$ is increasing in $\log(\sigma_u/\sigma_v)$.




Note: This figure plots the average values of different metrics across $N = 1{,}000$ simulation sessions as $\log(\sigma_u/\sigma_v)$ varies. Panels A and B plot the price informativeness and market liquidity relative to the theoretical benchmark of the perfect cartel equilibrium, i.e., $\mathcal{I}/\mathcal{I}^M$ and $\mathcal{L}/\mathcal{L}^M$, respectively. Panel C plots the magnitude of mispricing $\mathcal{E}$. The blue solid line represents the simulation experiments with informed AI speculators; the red dash-dotted and black dotted lines represent the theoretical benchmarks of the noncollusive Nash equilibrium and perfect cartel equilibrium, respectively. The other parameters are set according to the baseline economic environment described in Section 4.7.

Figure 9: Price informativeness, market liquidity, and mispricing for $\log(\sigma_u/\sigma_v) \in [-5, 5]$.

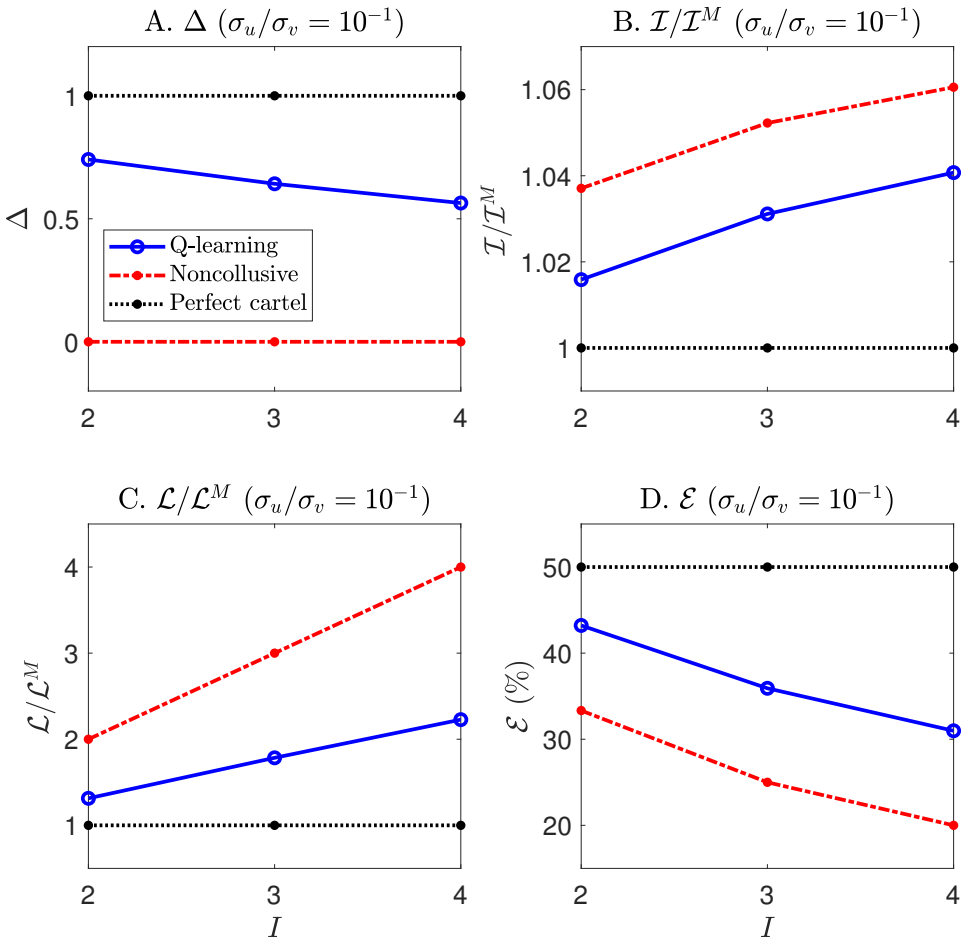
the theoretical benchmark of the noncollusive Nash equilibrium and perfect cartel equilibrium, $\mathcal{L}^N/\mathcal{L}^M$ is greater than 1 and decreasing in $\log(\sigma_u/\sigma_v)$.[18] The blue solid line shows that the market liquidity in our simulation experiments with informed AI speculators is higher than that in the theoretical benchmark of the perfect cartel equilibrium and lower than that of the noncollusive equilibrium. The blue solid line displays an U shape similar to panel A of Figure 8, indicating that the market liquidity is closer to the theoretical benchmark of the perfect cartel equilibrium if there is more AI collusion.

Panel C of Figure 9 plots the magnitude of mispricing in financial markets. Mispricing is higher in the theoretical benchmark of the perfect cartel equilibrium (the black dotted line) than in the noncollusive equilibrium (the red dash-dotted line). The blue solid line shows that AI collusion increases mispricing, and the magnitude is larger when there is a higher degree of collusion among informed AI speculators.

## 6 Further Inspection of Model Ingredients

In this section, we further inspect several key parameters in our simulation experiments. In Subsection 6.1, we study how the number of informed AI speculators affects their trading strategies. In Subsection 6.2, we study the implication of informed AI speculators' subjective discount rates. Finally, in Subsection 6.3, we study the impacts of hyperparameters $\alpha$ and $\beta$ on informed AI speculators' learning outcomes.

[18] This is because $\lambda^N < \lambda^M$ for all $\log(\sigma_u/\sigma_v)$. Intuitively, in the perfect cartel equilibrium, the market maker knows that informed speculators submit orders jointly like a monopoly, and thus the market maker adopts a pricing rule that is more responsive to the combined order flow of informed speculators and the noise trader, i.e., $\gamma^N < \gamma^M$. As $\log(\sigma_u/\sigma_v)$ increases, both $\lambda^N$ and $\lambda^M$ decline, so that market liquidity defined by equation (4.11) increases.

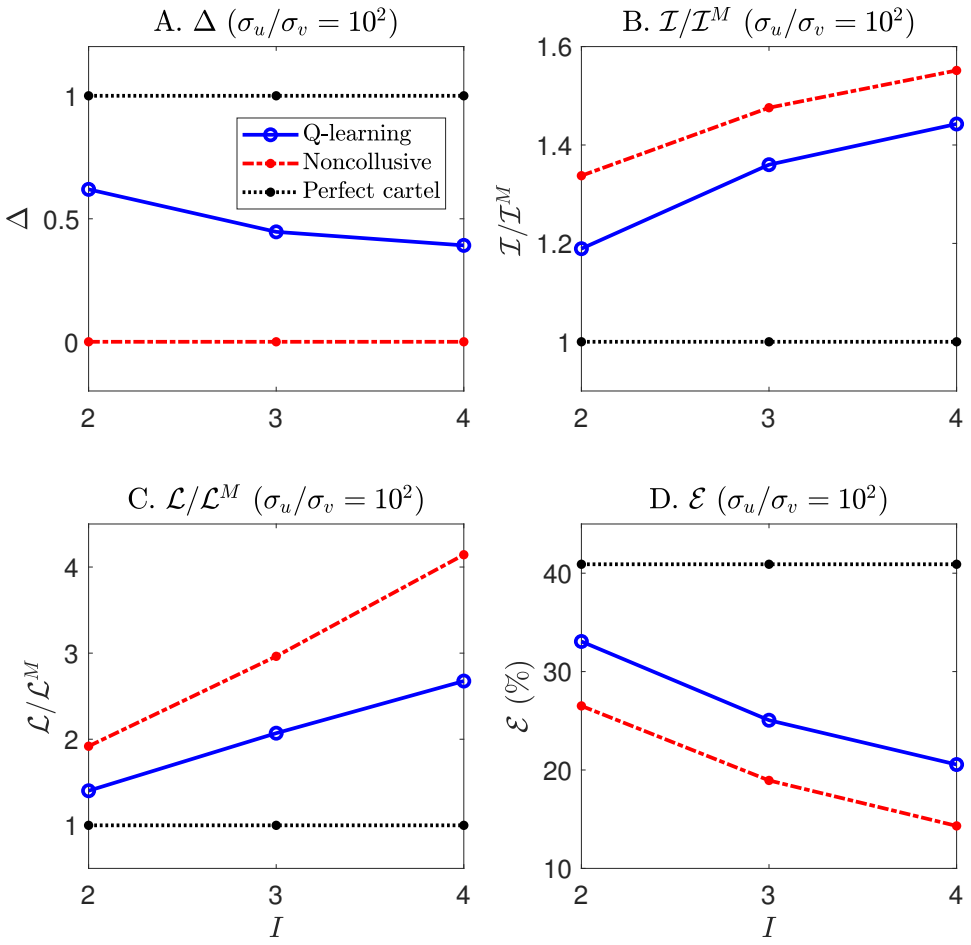



Note: The blue solid line plots the average values of $\Delta^C$, $\mathcal{I}^C/\mathcal{I}^M$, $\mathcal{L}^C/\mathcal{L}^M$, and $\mathcal{E}^C$ across $N = 1{,}000$ simulation sessions as the number of informed AI speculators $I$ varies, in the environment with low noise trading risks, i.e., $\sigma_u/\sigma_v = 10^{-1}$. The red dash-dotted and black dotted lines represent the theoretical benchmarks of the noncollusive Nash equilibrium and perfect cartel equilibrium, respectively. The other parameters are set according to the baseline economic environment described in Section 4.7.

Figure 10: Implications of the number of informed AI speculators ($\sigma_u/\sigma_v = 10^{-1}$).

### 6.1 Number of Informed AI Speculators $I$

Our model in Section 3 predicts that in the environment with low price efficiency (i.e., $\xi$ is large or $\theta$ is small) and low noise trading risks (i.e., small $\sigma_u/\sigma_v$), informed speculators are less able to collude through price-trigger strategies when the number of informed speculators increases (see Proposition 3.6.(i)). In the simulation experiments with informed AI speculators, we find similar patterns. Specifically, consider the baseline economic environment described in Section 4.7. In Figure 10, we conduct simulation experiments in the environment with low noise trading risks ($\sigma_u/\sigma_v = 10^{-1}$). Panel A shows that as the number of informed AI speculators $I$ increases from 2 to 4, the average $\Delta^C$ decreases from 0.74 to 0.56, indicating a decline in the extent of collusion among informed AI speculators. Moreover, panels B to D show that as $I$ increases, the relative




Note: The blue solid line plots the average values of $\Delta^C$, $\mathcal{I}^C/\mathcal{I}^M$, $\mathcal{L}^C/\mathcal{L}^M$, and $\mathcal{E}^C$ across $N = 1{,}000$ simulation sessions as the number of informed AI speculators $I$ varies, in the environment with high noise trading risks, i.e., $\sigma_u/\sigma_v = 10^2$. The red dash-dotted and black dotted lines represent the theoretical benchmarks of the noncollusive Nash equilibrium and perfect cartel equilibrium, respectively. The other parameters are set according to the baseline economic environment described in Section 4.7.

Figure 11: Implications of the number of informed AI speculators ($\sigma_u/\sigma_v = 10^2$).

price informativeness $\mathcal{I}^C/\mathcal{I}^M$ and market liquidity $\mathcal{L}^C/\mathcal{L}^M$ increase whereas the magnitude of mispricing $\mathcal{E}^C$ decreases.

For comparisons, in Figure 11, we conduct simulation experiments in the environment with high noise trading risks ($\sigma_u/\sigma_v = 10^2$). In these experiments, informed AI speculators collude through homogenized learning biases, as discussed in Subsection 5.2. The implications of $I$ for informed AI speculators' strategies are similar to the experiments with low noise trading risks. Specifically, panel A shows that as $I$ increases from 2 to 4, the average $\Delta^C$ decreases from 0.62 to 0.39. These results suggest that the coordination through homogenized learning biases becomes more difficult to achieve when there are more informed AI speculators in the market. Intuitively, the equilibrium degree of collusion is determined by the interaction of two counterveiling forces. One is the magnitude of learning biases, which is the mechanism that generates collusion. The

other is the deviation gain from the self-confirming collusive equilibrium. A larger deviation gain makes it more difficult for informed AI speculators to reach the collusive equilibrium because in the process of exploration (which, in essence, generates deviation behavior), these speculators will more likely learn to play noncollusive actions despite the existence of learning biases. As the number of informed AI speculators $I$ increases, the deviation gain from the equilibrium trading strategies becomes larger, but the magnitude of learning biases remain unchanged.[19] Therefore, as $I$ increases, collusion becomes more difficult and $\Delta^C$ declines.

Panels B to D show that as $I$ increases, the relative price informativeness $\mathcal{I}^C/\mathcal{I}^M$ and market liquidity $\mathcal{L}^C/\mathcal{L}^M$ increase whereas the magnitude of mispricing $\mathcal{E}^C$ decreases.

### 6.2 Subjective Discount Rate $\rho$

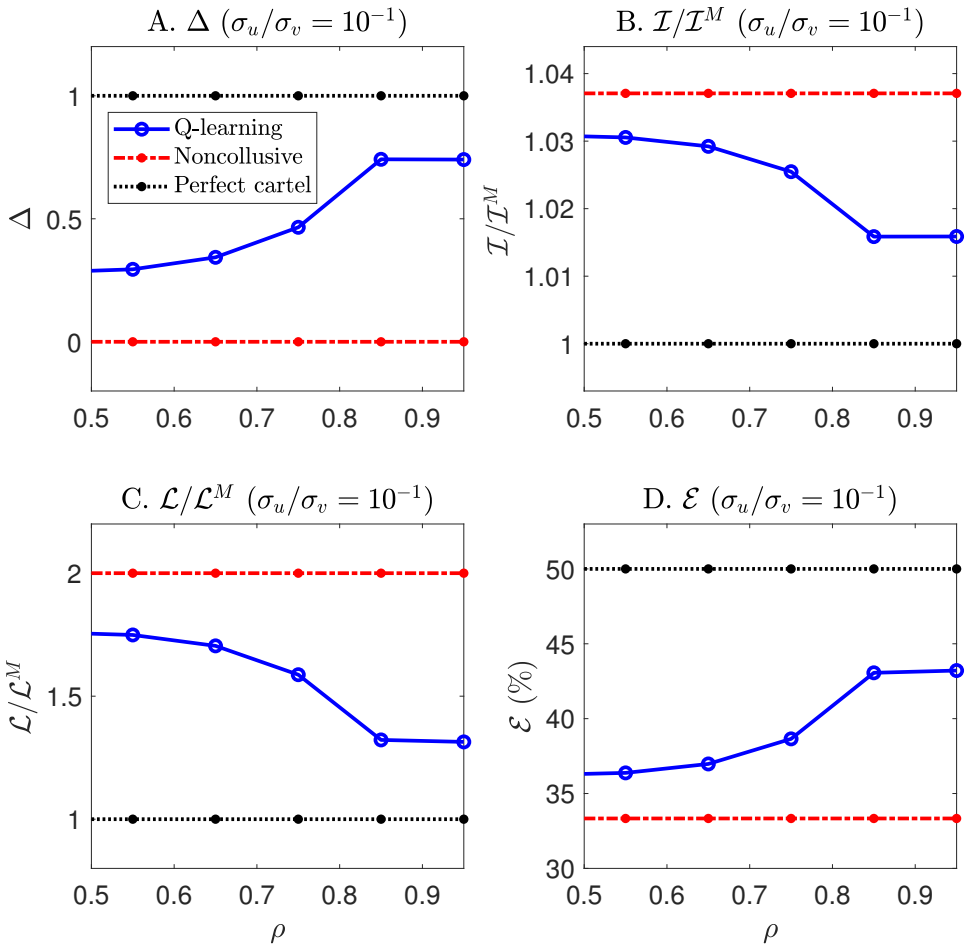
Our model in Section 3 predicts that in the environment with low price efficiency (i.e., $\xi$ is large or $\theta$ is small) and low noise trading risks (i.e., small $\sigma_u/\sigma_v$), informed speculators are able to collude on higher profits through price-trigger strategies as the subjective discount rate $\rho$ increases (see Proposition 3.6.(iii)). In the simulation experiments with informed AI speculators, we find similar patterns. Specifically, consider the baseline economic environment described in Section 4.7. In Figure 12, we conduct simulation experiments in the environment with low noise trading risks ($\sigma_u/\sigma_v = 10^{-1}$). Panel A shows that as $\rho$ increases from 0.5 to 0.95, the average $\Delta^C$ increases from 0.29 to 0.74, indicating an increase in the extent of collusion among informed AI speculators. Moreover, panels B to D show that as $\rho$ increases, the relative price informativeness $\mathcal{I}^C/\mathcal{I}^M$ and market liquidity $\mathcal{L}^C/\mathcal{L}^M$ decline whereas the magnitude of mispricing $\mathcal{E}^C$ increases.

Turning to the environment with high noise trading risks, the theoretical properties discussed in Section 5.2.4 indicate that as the subjective discount rate $\rho$ increases, the magnitude of learning biases declines, and as a result, informed AI speculators would find it more difficult to collude. The patterns observed in our simulation experiments are consistent with this prediction. In particular, in Figure 13, we conduct simulation experiments in the environment with high noise trading risks ($\sigma_u/\sigma_v = 10^2$). Panel A shows that as $\rho$ increases from 0.5 to 0.95, the average $\Delta^C$ decreases from 0.76 to 0.62. Moreover, panels B to D show that as $\rho$ increases, the relative price informativeness $\mathcal{I}^C/\mathcal{I}^M$ and market liquidity $\mathcal{L}^C/\mathcal{L}^M$ increase whereas the magnitude of mispricing $\mathcal{E}^C$ declines.

### 6.3 Hyperparameters $\alpha$ and $\beta$

In this subsection, we study how the hyperparameters $\alpha$ and $\beta$ affect informed AI speculators' profits in equilibrium. Similar to the baseline economic environment, we consider informed AI speculators adopting the same values of $\alpha$ and $\beta$. In panel A of Figure 14, we plot the average $\Delta^C$ in the environment with low noise trading risks ($\sigma_u/\sigma_v = 10^{-1}$) for different values of $\alpha$ and $\beta$.

[19] When $I$ increases, individual informed AI speculators trading flows $x_i$ decrease. However, in equation (G.6), the trading flow $x_i$ proportionally affects every term. Thus, the decrease in $x_i$ does not affect the importance of the term $\alpha\lambda x_i \sum_{\tau=0}^{T}(1-\alpha)^\tau u_{t(T-\tau)}$, which causes learning biases, relative to other terms in equation (G.6). This is why the magnitude of learning biases does not depend on $I$.

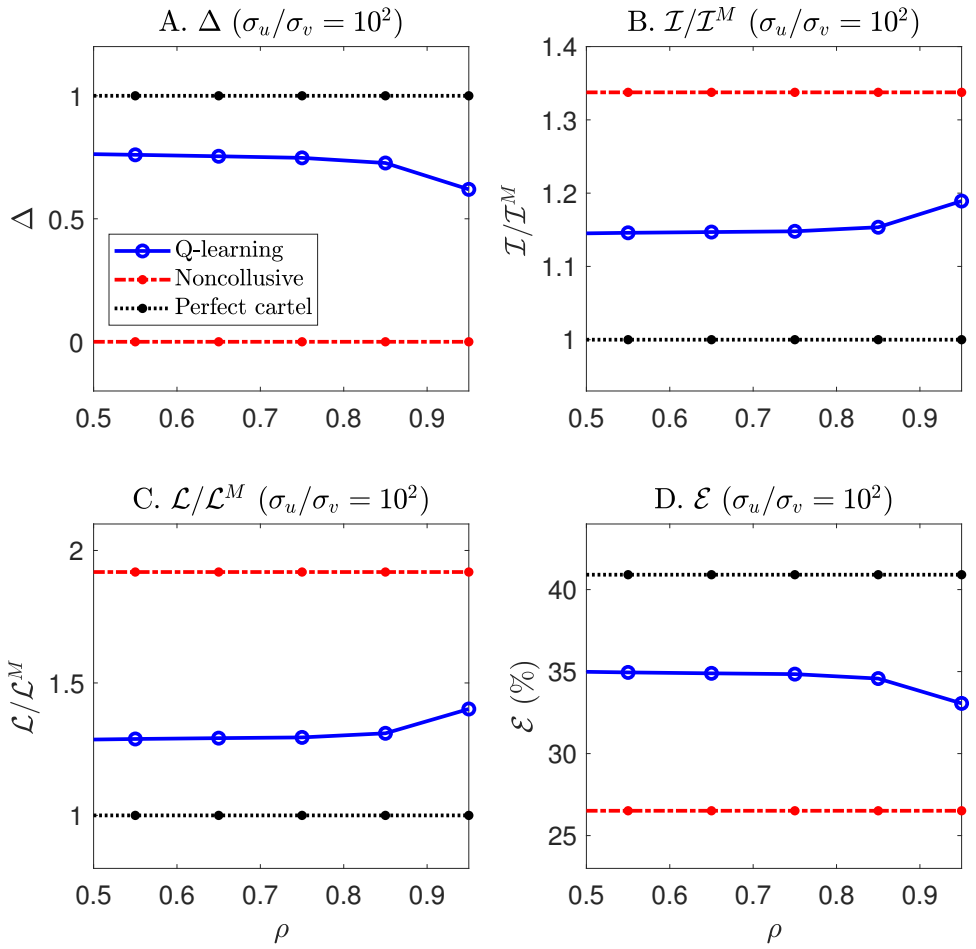



Note: The blue solid line plots the average values of $\Delta^C$, $\mathcal{I}^C/\mathcal{I}^M$, $\mathcal{L}^C/\mathcal{L}^M$, and $\mathcal{E}^C$ across $N = 1,000$ simulation sessions as the subjective discount rate $\rho$ varies, in the environment with low noise trading risks, i.e., $\sigma_u/\sigma_v = 10^{-1}$. The red dash-dotted and black dotted lines represent the theoretical benchmarks of the noncollusive Nash equilibrium and perfect cartel equilibrium, respectively. The other parameters are set according to the baseline economic environment described in Section 4.7.

Figure 12: Implications of the subjective discount rate ($\sigma_u/\sigma_v = 10^{-1}$).

As discussed in Subsection 5.1, informed AI speculators need to conduct sufficient explorations to learn punishment strategies, which is achieved by setting a sufficiently low $\beta$. Indeed, when $\beta = 10^{-6}$, the red bars in panel A of Figure 14 show that informed AI speculators can easily achieve a very high value of $\Delta^C = 0.90$ (corresponding to $\alpha = 0.001$) whereas when $\beta = 10^{-3}$, the yellow bars show that informed AI speculators can only achieve a low value of $\Delta^C = 0.40$ (corresponding to $\alpha = 0.1$).

Panel A of Figure 14 further shows that, to achieve the best collusive outcomes, the values of $\alpha$ and $\beta$ have to be jointly determined. That is, the choice of a smaller $\beta$ needs to be matched with a smaller $\alpha$, and conversely, the choice of a larger $\beta$ needs to be matched with a larger $\alpha$. Intuitively, setting a small $\beta$ ensures that informed AI speculators will spend a long time in the exploration mode in which they randomly choose different actions, resulting in extensive experimentation.

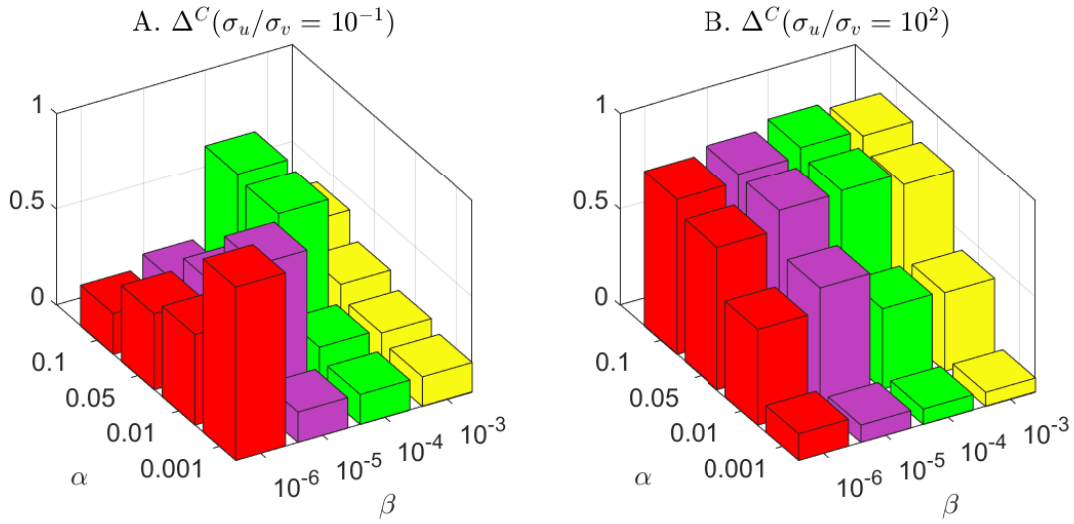



Note: The blue solid line plots the average values of $\Delta^C$, $\mathcal{I}^C/\mathcal{I}^M$, $\mathcal{L}^C/\mathcal{L}^M$, and $\mathcal{E}^C$ across $N = 1{,}000$ simulation sessions as the subjective discount rate $\rho$ varies, in the environment with high noise trading risks, i.e., $\sigma_u/\sigma_v = 10^2$. The red dash-dotted and black dotted lines represent the theoretical benchmarks of the noncollusive Nash equilibrium and perfect cartel equilibrium, respectively. The other parameters are set according to the baseline economic environment described in Section 4.7.

Figure 13: Implications of the subjective discount rate ($\sigma_u/\sigma_v = 10^2$).

Then, setting a small $\alpha$ is necessary to record the value learned in the past whereas setting a large $\alpha$ will disrupt learning as the algorithm would forget what it has learned in the past too rapidly. By contrast, setting a large $\beta$ means that informed AI speculators only spend a short period of time in the exploration mode. Then, if we still set a small $\alpha$, the Q-matrices of informed AI speculators would not be updated significantly until the algorithms complete exploration. Thus, when $\beta$ is large, setting a small $\alpha$ would backfire, making the initial exploration futile. Instead, setting a large $\alpha$ in this case would help informed AI speculators to learn punishment strategies to achieve more collusive outcomes.

In panel B of Figure 14, we plot the average $\Delta^C$ in the environment with high noise trading risks ($\sigma_u/\sigma_v = 10^2$) for different values of $\alpha$ and $\beta$. Holding $\beta$ unchanged at each value of $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$, panel B shows that the value of $\Delta^C$ declines monotonically as $\alpha$ decreases. This




Note: Panel A plots $\Delta^C$ in the environment with low noise trading risks ($\sigma_u/\sigma_v = 10^{-1}$); panel B plots $\Delta^C$ in the environment with high noise trading risks ($\sigma_u/\sigma_v = 10^{2}$). The other parameters are set according to the baseline economic environment described in Section 4.7.

Figure 14: Implications of hyperparameters $\alpha$ and $\beta$ on $\Delta^C$.

is because when noise trading risks are large, the supra-competitive profits are attained because informed AI speculators have homogenized learning biases. As discussed in Section 5.2.4, the learning biases due to the failure of the law of large numbers are mitigated when $\alpha$ becomes small.

Taken together, a key feature that distinguishes collusion through price-trigger strategies (panel A of Figure 14) and collusion through homogenized learning biases (panel B of Figure 14) is whether improved learning through setting a sufficiently small $\alpha$ would significantly reduce the supra-competitive profits of informed AI speculators.

# 7 Coordinated Choice of Q-Learning Algorithms

As shown in panel B of Figure 14, setting a lower forgetting rate $\alpha$ reduces the magnitude of learning biases but it takes longer time and more computation power to train the algorithm. Thus, we can think of $\alpha$ as capturing the "intelligence level" of the algorithm: the algorithm is more advanced if it has a lower $\alpha$.

In this section, we focus on the environment with high noise trading risks and allow informed AI speculators to choose different values of the hyperparameter $\alpha$ for their Q-learning algorithms. We evaluate the implications for trading profits. Specifically, in Subsection 7.1, we show that the more advanced algorithm will make more profit than the less advanced algorithm. Moreover, given the peer's choice of $\alpha$, by setting a lower $\alpha$, the informed AI speculator can increase its own profit. However, importantly, both informed AI speculators can obtain supra-competitive

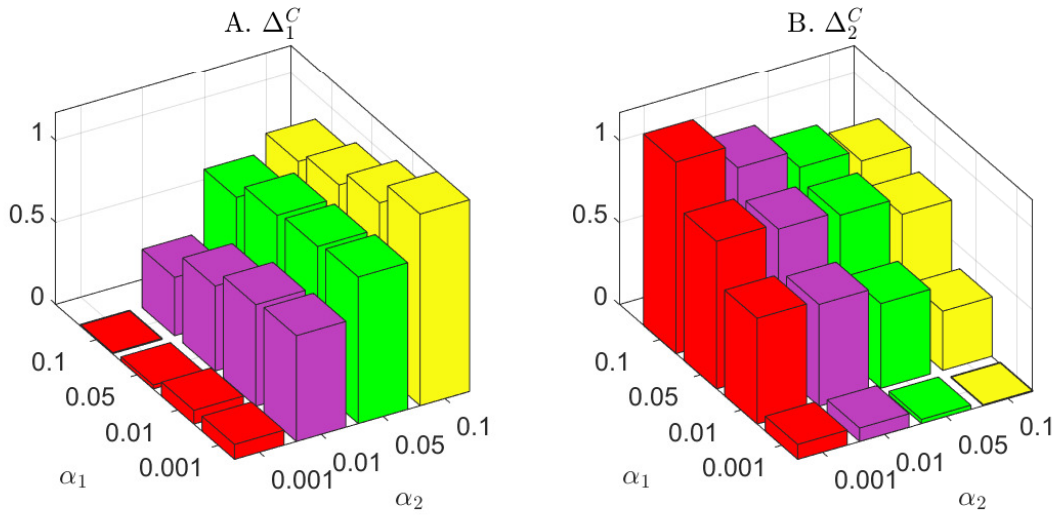
profits if they both adopt less advanced algorithms with similar learning biases. In Subsection 7.2, we extend the Q-learning algorithm to a two-tier Q-learning algorithm in which informed AI speculators learn both the optimal choice of the forgetting rate $\alpha$ and the optimal trading strategies associated with the choice of $\alpha$. We show that informed AI speculators will learn to adopt high values of $\alpha$ in the stationary equilibrium, and such coordination allows both of them to obtain supra-competitive profits.

### 7.1 Homogenized Learning Biases

Focusing on the baseline economic environment with two informed AI speculators, as described in Section 4.7 except for setting $\sigma_u/\sigma_v = 10^2$, representing an environment with high noise trading risks. We allow the two informed AI speculators to adopt different values of $\alpha$, but the same value of $\beta$. Intuitively, the informed AI speculator adopting a more advanced Q-learning algorithm (i.e., a lower $\alpha$) would have smaller learning biases than the one adopting a less advanced algorithm (i.e., a higher $\alpha$). As discussed in Subsection 5.2.4, learning biases induce informed AI speculators to adopt more conservative trading strategies, i.e., smaller order flows. Therefore, the informed AI speculator with a less advanced algorithm would adopt a more conservative trading strategy than the one with a more advanced algorithm. This essentially enables the informed AI speculator with a more advanced algorithm to take advantage of the other informed AI speculator and obtain more profits than what it would obtain when the other speculator adopts an algorithm with the same $\alpha$. Conversely, the informed AI speculator with a less advanced algorithm would obtain less profits than what it would obtain when the other speculator adopts an algorithm with the same $\alpha$.

The results of our simulation experiments are consistent with the above intuition. In Figure 15, we allow each informed AI speculator $i$ to adopt algorithms with different values of $\alpha_i$, with $\alpha_i = 0.001, 0.01, 0.05$ and $0.1$ for $i = 1, 2$. Panels A and B plot the average $\Delta_1^C$ and $\Delta_2^C$ for informed AI speculators 1 and 2, respectively. It is shown that for any combination of $(\alpha_1, \alpha_2)$, the informed AI speculator with a lower $\alpha_i$ attains a higher average $\Delta_i^C$ than the other informed AI speculator. Moreover, holding $\alpha_1$ unchanged at each value of $\{0.001, 0.01, 0.05, 0.1\}$, as informed AI speculator 2's $\alpha_2$ decreases, the average $\Delta_1^C$ for informed AI speculator 1 decreases and the average $\Delta_2^C$ for informed AI speculator 2 increases. Similarly, holding $\alpha_2$ unchanged at each value of $\{0.001, 0.01, 0.05, 0.1\}$, as informed AI speculator 1's $\alpha_1$ decreases, the average $\Delta_2^C$ for informed AI speculator 2 decreases and the average $\Delta_1^C$ for informed AI speculator 1 increases.

Our results indicate that both informed AI speculators can obtain supra-competitive profits if both of them adopt less advanced algorithms with a high value of $\alpha$. Holding one informed AI speculator's algorithm unchanged, the other speculator could increase its profit by adopting a more advanced algorithm with a lower value of $\alpha$, and at the same time, the profit of the speculator with a less advanced algorithm would decrease. However, if both informed AI speculators adopt advanced algorithms with a small value of $\alpha$, the profit for both of them will decrease relative to the equilibrium where both speculators adopt unadvanced algorithms. The results we observe bear similarity to the general equilibrium effects in active management, as characterized by Stambaugh




Note: We allow the two informed AI speculators to adopt Q-learning algorithms with different values of the forgetting rate, denoted by $\alpha_1$ and $\alpha_2$ for informed AI speculators 1 and 2, respectively. Panels A and B plot $\Delta_1^C$ and $\Delta_2^C$ in the environment with high noise trading risks ($\sigma_u/\sigma_v = 10^2$). The other parameters are set according to the baseline economic environment described in Section 4.7.

Figure 15: Profit gain when informed AI speculators adopt algorithms with different values of $\alpha$.

(2020). According to his model, if all managers lack the ability to select positive-alpha stocks, they can collectively achieve high profits. When a small fraction of managers gains more skill, it results in increased profits for the skilled ones, while the less skilled managers see a decline in their profits. However, if a large proportion of managers becomes more skilled, the profits for all managers start to diminish. This decline is due to a shrinking alpha magnitude, caused by more substantial price corrections in general equilibrium. Interestingly, the total profit of the active management industry typically decreases whenever any of the managers become more skilled. In a recent work, Dugast and Foucault (2024) derive a similar result by showing that improvements in the skills of active asset managers, due to lower information processing costs or the proliferation of new datasets, can reduce their average performance as asset prices become more informative.

### 7.2 Adaptive Forgetting Rates

In practice, the forgetting rate $\alpha$ is not necessarily fixed throughout the simulation experiments. Instead, many Q-learning algorithms are implemented with adaptive forgetting rates, which are adjusted dynamically in response to the performance of the model. In this subsection, we show that informed AI speculators can learn to coordinately choose high values of $\alpha$ in environments with high noise trading risks, despite the fact that choosing a low forgetting rate unilaterally may boost self-profit. This result implies that an equilibrium with unadvanced algorithms (i.e., high $\alpha$)

may arise endogenously due to the optimal decisions of informed AI speculators.

***Two-Tier Q-Learning Algorithm.*** Each informed AI speculator $i$ adopts a two-tier Q-learning algorithm. In the lower tier, the informed AI speculator adopts a Q-learning algorithm to learn the lower-tier Q-matrix $\widehat{Q}_{i,t}(s_t, x_{i,t})$ for state $s_t = \{p_{t-1}, v_t\}$ and order flow $x_{i,t}$, given the choice of $\alpha_{i,t}$ in the upper tier. The lower-tier Q-learning algorithm is identical to the algorithm described in Section 4.1, except for the use of a time-varying adaptive forgetting rate $\alpha_{i,t}$. In the upper tier, the informed AI speculator adopts a Q-learning algorithm to learn the upper-tier Q-matrix $\widehat{Q}^u_{i,t}(s^u_{i,t}, \alpha_{i,t})$ for state $s^u_{i,t}$ and action $\alpha_{i,t}$.

For any given choice of $\alpha_{i,t}$ in the upper tier, it is necessary to ensure that the lower tier Q-learning algorithm is run for a sufficiently long period of time, so that the profits corresponding to the choice of $\alpha_{i,t}$ fully stablize. This means that compared with the choice of $x_{i,t}$ in the lower tier, the choice of $\alpha_{i,t}$ in the upper tier has to be experimented at a much lower frequency. Therefore, we specify that each informed AI speculator $i$ adjusts its upper tier's action $\alpha_{i,t}$ only after the lower tier finishes a training epoch that lasts for a total of $T$ periods, with $T$ being a large integer.

Specifically, let $\tau = 1, 2, ...$ denote all training epochs of the lower-tier Q-learning algorithm. The training epoch $\tau$ represents the period from $(\tau-1)T+1$ to $\tau T$. Within each training epoch $\tau$, each informed AI speculator $i$'s upper-tier Q-matrix $\widehat{Q}^u_{i,t}(s^u_{i,t}, \alpha_{i,t})$ or action $\alpha_{i,t}$ stay unchanged from period $(\tau-1)T+1$ to period $\tau T-1$; the values of $\widehat{Q}^u_{i,t}(s^u_{i,t}, \alpha_{i,t})$ and action $\alpha_{i,t}$ are updated only at the end of the training epoch, occurring at $t = \tau T$. Therefore, without loss of generality, we only need to specify the recursive learning equation of the upper-tier Q-learning algorithm at the end of each period, $t = \tau T$, as follows:

$$\widehat{Q}^u_{i,(\tau+1)T}(s^u_{i,\tau T}, \alpha_{i,\tau T}) = (1-\alpha^u)\widehat{Q}^u_{i,\tau T}(s^u_{i,\tau T}, \alpha_{i,\tau T}) + \alpha^u \left[\pi^u_{i,\tau T} + \rho^u \max_{\alpha' \in \mathcal{A}} \widehat{Q}^u_{i,\tau T}(s^u_{i,(\tau+1)T}, \alpha')\right], \quad (7.1)$$

for $\tau = 1, 2, ...$ In equation (7.1), $\pi^u_{i,\tau}$ is the reward in the training epoch $\tau$, given by $\pi^u_{i,\tau T} = \frac{1}{T}\sum_{t=(\tau-1)T+1}^{\tau T}(v_t - p_t)x_{i,t}$, which is the average trading profit over the last $T$ periods, from period $(\tau-1)T+1$ to period $\tau T$. The parameters $\alpha^u$ and $\rho^u$ are the forgetting rate and the subjective discount rate for the upper tier Q-learning algorithm. For tractability, we choose the state variable $s^u_{i,\tau T} = \{\pi^u_{i,(\tau-1)T}\}$, which is the reward in the previous training epoch. The choice of $\alpha_{i,\tau T}$ is chosen as follows:

$$\alpha_{i,\tau T} = \begin{cases} \text{argmax}_{\alpha' \in \mathcal{A}}\, \widehat{Q}^h_{i,\tau T}(s^u_{i,\tau T}, \alpha'), & \text{with prob. } 1-\varepsilon^u_\tau, \quad \text{(exploitation)} \\ \widetilde{\alpha} \sim \text{uniform distribution on } \mathcal{A}, & \text{with prob. } \varepsilon^u_\tau. \quad \text{(exploration)} \end{cases} \quad (7.2)$$

The exploration rate is specified as $\varepsilon_\tau = e^{-\beta^u \tau}$, where $\beta^u$ is a parameter governing the decaying speed of exploration rates across training epochs.

***Simulation Results.*** The two-tier Q-learning algorithm takes a substantially longer time to converge because there are experimentations on both $\alpha_{i,t}$ and $x_{i,t}$. We consider the following

parameter values: $\alpha^u = 0.1$, $\beta^u = 10^{-4}$, and $\rho^u = 0.95$. Each training epoch has a total of $T = 10,000,000$ periods. The convergence criterion requires the decisions of $\alpha_{i,t}$ to stay unchanged for $100,000$ consecutive training epochs. For tractability, we choose three grids for the choice of $\alpha_{i,t}$, with $\mathcal{A} = \{0.001, 0.01, 0.1\}$. The parameters and grids for the lower-tier Q-learning algorithm are similar to those described in Section 4. In particular, there are two informed AI speculators. We separately conduct $N = 1,000$ independent simulations for the environments with high and low noise trading risks.

Our primary focus is on the environment with high noise trading risks (i.e., $\sigma_u/\sigma_v = 10^2$). As shown in panel B of Figure 15, the two informed AI speculators encounter a problem resembling the prisoner's dilemma. Specifically, given informed AI speculator $i$'s choice of $\alpha_i$, informed AI speculator $j$ can gain by adopting the smallest $\alpha_j = 0.001$. However, both informed AI speculators would not make much profit if they reach the Nash equilibrium of $(\alpha_1, \alpha_2) = (0.001, 0.001)$. Instead, both of them would attain supra-competitive profits by coordinately reaching the equilibrium with $(\alpha_1, \alpha_2) = (0.01, 0.01)$ or $(\alpha_1, \alpha_2) = (0.1, 0.1)$, that is, by adopting unadvanced algorithms to trade. In theory, these two equilibria with high values of $\alpha$ can only be sustained in a repeated game. Turning to our simulation experiments with informed AI speculators adopting the two-tier Q-learning algorithms, we find that across the $N = 1,000$ simulations sessions, 272 sessions converge to the equilibrium with $(\alpha_1, \alpha_2) = (0.1, 0.1)$, and 710 sessions converge to the equilibrium with $(\alpha_1, \alpha_2) = (0.01, 0.01)$. There does not exist a single simulation session that converges to the equilibrium with $(\alpha_1, \alpha_2) = (0.001, 0.001)$, even though this is the unique Nash equilibrium in a one-shot game. Our results indicate that in the environment with high noise trading risks, the two informed AI speculators are able to learn to adopt less advanced algorithms, which have high values of $\alpha$, in the stationary equilibrium. This coordination allows both AI speculators to obtain supra-competitive profits.

For comparisons, we also conduct simulation experiments in the environment with low noise trading risks (i.e., $\sigma_u/\sigma_v = 10^{-1}$). As shown in panel A of Figure 14, the optimal outcome is achieved if the two informed AI speculators choose to play the equilibrium with $(\alpha_1, \alpha_2) = (0.01, 0.01)$, given that $\beta = 10^{-5}$. We find that across the $N = 1,000$ simulations sessions, 957 sessions converge to this equilibrium. This suggests that our simple two-tier Q-learning algorithm enables the two informed AI speculators to learn to play the optimal equilibrium. The algorithm's excellent performance is due to the fact that in this environment, informed AI speculators do not face a prisoner's dilemma problem. That is, the equilibrium with $(\alpha_1, \alpha_2) = (0.01, 0.01)$, which yields the highest trading profits for both informed AI speculators, is also the Nash equilibrium of a one-shot game. In other words, choosing the forgetting rate $\alpha_i = 0.01$ maximizes informed AI speculator $i$'s trading profits regardless of the forgetting rate that the other informed AI speculator chooses.

# References


**Abreu, Dilip, David Pearce, and Ennio Stacchetti.** 1986. "Optimal cartel equilibria with imperfect monitoring." *Journal of Economic Theory*, 39(1): 251–269.

**Abreu, Dilip, Paul Milgrom, and David Pearce.** 1991. "Information and Timing in Repeated Partnerships." *Econometrica*, 59(6): 1713–1733.

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

**Assad, Stephanie, Robert Clark, Daniel Ershov, and Lei Xu.** 2023. "Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market." *Journal of Political Economy*, Forthcoming.

**Bagattini, Giulio, Zeno Benetti, and Claudia Guagliano.** 2023. "Artificial intelligence in EU securities markets." *ESMA50-164-6247*. European Securities and Markets Authority.

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

**Bommasani, Rishi, Kathleen Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang.** 2022. "Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization?"

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

**Cho, In-Koo, and Thomas J. Sargent.** 2008. "Self-confirming Equilibria." 407–408. Palgrave Macmillan.

**Colliard, Jean-Edouard, Thierry Foucault, and Stefano Lovo.** 2022. "Algorithmic Pricing and Liquidity in Securities Markets." HEC Paris Working Papers.

**Dou, Winston Wei, Wei Wang, and Wenyu Wang.** 2023. "The Cost of Intermediary Market Power for Distressed Borrowers." The Wharton School at University of Pennsylvania Working Papers.

**Dou, Winston Wei, Yan Ji, and Wei Wu.** 2021*a*. "Competition, Profitability, and Discount Rates." *Journal of Financial Economcis*, 140(2): 582–620.

**Dou, Winston Wei, Yan Ji, and Wei Wu.** 2021*b*. "The Oligopoly Lucas Tree." *The Review of Financial Studies*, 35(8): 3867–3921.

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

**Fudenberg, Drew, and David Levine.** 1993. "Self-Confirming Equilibrium." *Econometrica*, 61(3): 523–45.

**Fudenberg, Drew, and David M. Kreps.** 1988. "A theory of learning, experimentation, and equilibrium in games." Working Papers.

**Fudenberg, Drew, and David M. Kreps.** 1995. "Learning in extensive-form games I. Self-confirming equilibria." *Games and Economic Behavior*, 8(1): 20–55.

**Fudenberg, Drew, and Eric Maskin.** 1986. "The Folk theorem in repeated games with discounting or with incomplete information." *Econometrica*, 54(3): 533–54.

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

**Goldstein, Itay, Emre Ozdenoren, and Kathy Yuan.** 2013. "Trading frenzies and their impact on real investment." *Journal of Financial Economics*, 109(2): 566–582.

**Graham, Benjamin.** 1973. *The Intelligent Investor.* . 4 ed., Publisher: Harper & Row, New York, NY.

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

**Greenwood, Robin, and Dimitri Vayanos.** 2014. "Bond Supply and Excess Bond Returns." *The Review of Financial Studies*, 27(3): 663–713.

**Greenwood, Robin, Samuel Hanson, Jeremy C Stein, and Adi Sunderam.** 2023. "A Quantity-Driven Theory of Term Premia and Exchange Rates*." *The Quarterly Journal of Economics*, qjad024.

**Harrington, Joseph E.** 2018. "Developing Competition Law for Collusion by Autonomous Artificial Agents." *Journal of Competition Law & Economics*, 14(3): 331–363.

**Hellwig, Christian, Arijit Mukherji, and Aleh Tsyvinski.** 2006. "Self-Fulfilling Currency Crises: The Role of Interest Rates." *The American Economic Review*, 96(5): 1769–1787.

**Johnson, Justin, and D. Daniel Sokol.** 2021. "Understanding AI Collusion and Compliance." *The Cambridge Handbook of Compliance*, , ed. Benjamin van Rooij and D. DanielEditors Sokol *Cambridge Law Handbooks*, 881–894. Cambridge University Press.

**Johnson, Justin Pappas, Andrew Rhodes, and Matthijs Wildenbeest.** 2023. "Platform Design when Sellers Use Pricing Algorithms." *Econometrica*, Forthcoming.

**Klein, Timo.** 2021. "Autonomous algorithmic collusion: Q-learning under sequential pricing." *The RAND Journal of Economics*, 52(3): 538–558.

**Kyle, Albert S.** 1985. "Continuous Auctions and Insider Trading." *Econometrica*, 53(6): 1315–1335.

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

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

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

**Long, J. Bradford De, Andrei Shleifer, Lawrence H. Summers, and Robert J. Waldmann.** 1990. "Noise Trader Risk in Financial Markets." *Journal of Political Economy*, 98(4): 703–738.

**Mildenstein, Eckart, and Harold Schleef.** 1983. "The Optimal Pricing Policy of a Monopolistic Marketmaker in the Equity Market." *The Journal of Finance*, 38(1): 218–231.

**Opp, Marcus M., Christine A. Parlour, and Johan Walden.** 2014. "Markup cycles, dynamic misallocation, and amplification." *Journal of Economic Theory*, 154: 126–161.

**Rotemberg, Julio J, and Garth Saloner.** 1986. "A supergame-theoretic model of price wars during booms." *American Economic Review*, 76(3): 390–407.

**Sandholm, Tuomas W., and Robert H. Crites.** 1996. "On multiagent Q-learning in a semi-competitive domain." 191–205. Berlin, Heidelberg:Springer Berlin Heidelberg.

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

**SEC.** 2023. "Conflicts of Interest Associated with the Use of Predictive Data Analytics by BrokerDealers and Investment Advisers." *Release Nos. 34-97990*. U.S. Securities and Exchange Commission.

**Stambaugh, Robert F.** 2020. "Skill and Profit in Active Management." National Bureau of Economic Research, Inc NBER Working Papers 26027.

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

**Tesauro, Gerald, and Jeffrey O. Kephart.** 2002. "Pricing in Agent Economies Using Multi-Agent Q-Learning." *Autonomous Agents and Multi-Agent Systems*, 5(3): 289–304.

**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.

# Appendix

## A Proof of Lemma 1

The preferred-habitat investor solves the following portfolio optimization problem for a given $p_t$:

$$\max_z \mathbb{E}\left[-e^{-\eta(v_t-p_t)z}/\eta\right]. \tag{A.1}$$

Because $v_t - p_t$ is distributed as $N(\overline{v}-p_t, \sigma_v^2)$, the first-order condition with respect to $z$ is

$$0 = \left[(\overline{v}-p_t) - \eta z\sigma_v^2\right] e^{-\eta z(\overline{v}-p_t)+(\eta z)^2\sigma_v^2/2}. \tag{A.2}$$

Thus, the optimal holding, $z$, is characterized as

$$z = -\frac{1}{\eta\sigma_v^2}(p_t - \overline{v}). \tag{A.3}$$

## B Proof of Proposition 3.3

Given that $s_t = 0$, let $J^C(\chi_i)$ denote each informed speculator $i$'s expected present value of future profits, when investor $i$ chooses $x_{i,t} = \chi_i(v_t - \overline{v})$ and all other $I-1$ informed investors choose $x^C(v_t) = \chi^C(v_t - \overline{v})$. That is,

$$\begin{aligned} J^C(\chi_i) =& \mathbb{E}\left[\left(v_t - p^C(y_t)\right)\chi_i(v_t-\overline{v})\right] \\ &+ \rho J^C(\chi_i)\mathbb{P}\left\{\text{Price trigger is not violated in period } t \middle| \chi_i, \chi^C\right\} \\ &+ \mathbb{E}\left[\sum_{\tau=1}^{T-1}\rho^\tau \pi^N(v_{t+\tau}) + \rho^T J^C(\chi_i)\right]\mathbb{P}\left\{\text{Price trigger is violated in period } t \middle| \chi_i, \chi^C\right\}, \end{aligned} \tag{B.1}$$

where $p^C(\cdot)$ is the pricing function of market makers in the collusive Nash equilibrium and

$$p^C(y_t) = \overline{v} + \lambda^C y_t, \quad \text{with } \lambda^C = \frac{\theta\gamma^C + \xi}{\theta + \xi^2} \text{ and } \gamma^C = \frac{I\chi^C}{(I\chi^C)^2 + (\sigma_u/\sigma_v)^2}, \tag{B.2}$$

$$y_t = \chi_i(v_t - \overline{v}) + (I-1)x^C(v_t) + u_t. \tag{B.3}$$

The probability of price trigger violation is

$$\begin{aligned} &\mathbb{P}\left\{\text{Price trigger is not violated in period } t \middle| \chi_i, \chi^C\right\} \\ =&\mathbb{E}\left[\mathbb{P}\left(p_t \le q(v_t)|v_t\right)\mathbf{1}\{v_t > \overline{v}\}\right] + \mathbb{E}\left[\mathbb{P}\left(p_t \ge q(v_t)|v_t\right)\mathbf{1}\{v_t < \overline{v}\}\right] \\ =&\mathbb{E}\left[\Phi(\sigma_u^{-1}(\chi^C - \chi_i)(v_t-\overline{v}) + \omega)\mathbf{1}\{v_t > \overline{v}\}\right] + \mathbb{E}\left[\Phi(\sigma_u^{-1}(\chi_i - \chi^C)(v_t-\overline{v}) + \omega)\mathbf{1}\{v_t < \overline{v}\}\right], \end{aligned}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution.

Evaluating equality (B.1) at $\chi_i = \chi^C$ leads to

$$\begin{aligned} J^C(\chi^C) = &\left(1 - \lambda^C I \chi^C\right) \chi^C \sigma_v^2 \\ &+ \rho J^C(\chi^C) \Phi(\omega) \\ &+ \frac{\rho - \rho^T}{1-\rho} \left[1 - \Phi(\omega)\right] \mathbb{E}\left[\pi^N(v)\right] + \rho^T J^C(\chi^C) \left[1 - \Phi(\omega)\right]. \end{aligned} \tag{B.4}$$

Thus, we can obtain that

$$J^C(\chi^C) = \frac{\left(1 - \lambda^C I \chi^C\right) \chi^C \sigma_v^2 + \frac{\rho - \rho^T}{1-\rho} \left[1 - \Phi(\omega)\right] \mathbb{E}\left[\pi^N(v)\right]}{1 - \rho\Phi(\omega) - \rho^T \left[1 - \Phi(\omega)\right]}. \tag{B.5}$$

The first-order derivative of the both sides of (B.1) with respect to $\chi_i$, evaluated at $\chi_i = \chi^C$, is

$$\begin{aligned} \nabla J^C(\chi^C) = &\left[1 - \lambda^C (I+1) \chi^C\right] \sigma_v^2 \\ &+ \rho \left[\nabla J^C(\chi^C)\right] \Phi(\omega) - \rho J^C(\chi^C) \frac{1}{\sigma_u} \phi(\omega) \mathbb{E}\left[|v - \overline{v}|\right] \\ &+ \frac{\rho - \rho^T}{1-\rho} \frac{1}{\sigma_u} \phi(\omega) \mathbb{E}\left[|v - \overline{v}|\right] \mathbb{E}\left[\pi^N(v)\right] \\ &+ \rho^T \left[\nabla J^C(\chi^C)\right] \left[1 - \Phi(\omega)\right] + \rho^T J^C(\chi^C) \frac{1}{\sigma_u} \phi(\omega) \mathbb{E}\left[|v - \overline{v}|\right], \end{aligned} \tag{B.6}$$

where $\phi(\cdot)$ is the probability density function of the standard normal distribution.

Because $v - \overline{v}$ is distributed as $N(0, \sigma_v^2)$, it follows that $\mathbb{E}\left[|v - \overline{v}|\right] = \sigma_v \sqrt{\frac{2}{\pi}}$. Plugging it into (B.6), we obtain that

$$\begin{aligned} \nabla J^C(\chi^C) = &\left[1 - \lambda^C (I+1) \chi^C\right] \sigma_v^2 \\ &+ \rho \left[\nabla J^C(\chi^C)\right] \Phi(\omega) - \rho J^C(\chi^C) \frac{\sigma_v}{\sigma_u} \phi(\omega) \sqrt{\frac{2}{\pi}} \\ &+ \frac{\rho - \rho^T}{1-\rho} \mathbb{E}\left[\pi^N(v)\right] \frac{\sigma_v}{\sigma_u} \phi(\omega) \sqrt{\frac{2}{\pi}} \\ &+ \rho^T \left[\nabla J^C(\chi^C)\right] \left[1 - \Phi(\omega)\right] + \rho^T J^C(\chi^C) \frac{\sigma_v}{\sigma_u} \phi(\omega) \sqrt{\frac{2}{\pi}}. \end{aligned} \tag{B.7}$$

The policy variable $\chi^C$ constitutes a collusive Nash equilibrium if speculator $i$ has no incentive to deviate by setting $\chi_i \neq \chi^C$. The first-order condition with respect to $\chi_i$, characterized by

$\nabla J^C(\chi^C) = 0$, leads to

$$\begin{aligned} 0 = &\left[1 - \lambda^C(I+1)\chi^C\right]\sigma_v^2 \\ &- \rho J^C(\chi^C)\frac{\sigma_v}{\sigma_u}\phi(\omega)\sqrt{\frac{2}{\pi}} \\ &+ \frac{\rho - \rho^T}{1-\rho}\mathbb{E}\left[\pi^N(v)\right]\frac{\sigma_v}{\sigma_u}\phi(\omega)\sqrt{\frac{2}{\pi}} \\ &+ \rho^T J^C(\chi^C)\frac{\sigma_v}{\sigma_u}\phi(\omega)\sqrt{\frac{2}{\pi}}. \end{aligned} \tag{B.8}$$

According to (B.2), as $\theta \to \infty$ or as $\xi \to 0$, $\lambda^C \to \gamma^C$, that is, the market approaches to the environment of Kyle (1985). In this case, the demand of the preferred-habitat investor is irrelevant. Because the system is continuous, it is sufficient to show that there is no solution $\chi^C \in [\chi^M, \chi^N)$ in the environment of Kyle (1985), where $\chi^N = \frac{1}{\sqrt{I}}\frac{\sigma_u}{\sigma_v}$ and $\chi^M = \frac{1}{I}\frac{\sigma_u}{\sigma_v}$ as a result of $\lambda^C = \gamma^C$. Let $\chi^C = \widehat{\chi}^C\frac{\sigma_u}{\sigma_v}$. Then, we show that there is no solution $\widehat{\chi}^C \in [\widehat{\chi}^M, \widehat{\chi}^N)$, with $\widehat{\chi}^M = \frac{1}{I}$ and $\widehat{\chi}^N) = \frac{1}{\sqrt{I}}$. In the Kyle case, $\mathbb{E}\left[\pi^N(v)\right] = \frac{\sigma_u\sigma_v}{(I+1)\sqrt{I}}$. Therefore, equations (B.5) and (B.8) can be rewritten, respectively, as follows:

$$J^C(\chi^C) = \frac{\left(1 - \gamma^C I\chi^C\right)\chi^C\sigma_v^2 + \frac{\rho - \rho^T}{1-\rho}\left[1 - \Phi(\omega)\right]\frac{\sigma_v\sigma_u}{(I+1)\sqrt{I}}}{1 - \rho\Phi(\omega) - \rho^T\left[1 - \Phi(\omega)\right]}. \tag{B.9}$$

and

$$0 = \left[1 - \gamma^C(I+1)\chi^C\right]\sigma_v^2 - \left[\rho J^C(\chi^C) - \frac{\rho - \rho^T}{1-\rho}\frac{\sigma_v\sigma_u}{(I+1)\sqrt{I}} - \rho^T J^C(\chi^C)\right]\frac{\sigma_v}{\sigma_u}\phi(\omega)\sqrt{\frac{2}{\pi}}. \tag{B.10}$$

Therefore, $\widehat{\chi}^C$ is the root of the following quadratic equation:

$$\begin{aligned} 0 = &\left[1 - I(\widehat{\chi}^C)^2\right]\frac{1}{\rho - \rho^T} \\ &- \left\{1 - \rho + (\rho - \rho^T)[1 - \Phi(\omega)]\right\}^{-1}\left\{\widehat{\chi}^C - \frac{1}{(I+1)\sqrt{I}}\left[1 + (I\widehat{\chi}^C)^2\right]\right\}\phi(\omega)\sqrt{\frac{2}{\pi}}, \end{aligned}$$

which can be simplified as

$$0 = 1 - I(\widehat{\chi}^C)^2 - \vartheta\left\{\widehat{\chi}^C - \frac{1}{(I+1)\sqrt{I}}\left[1 + (I\widehat{\chi}^C)^2\right]\right\}, \tag{B.11}$$

where

$$\vartheta = \frac{\phi(\omega)}{\frac{1-\rho}{\rho-\rho^T} + 1 - \Phi(\omega)}\sqrt{\frac{2}{\pi}}. \tag{B.12}$$

Solving the above problem, we obtain

$$\widehat{\chi}^C = \frac{\vartheta \pm \left| -2\sqrt{I} + \frac{I-1}{I+1}\vartheta \right|}{-2I + 2\vartheta\frac{I\sqrt{I}}{I+1}}.$$

There are three cases.

Case 1: if $-2\sqrt{I} + \frac{I-1}{I+1}\vartheta \leq 0$ and $-2I + 2\vartheta\frac{I\sqrt{I}}{I+1} < 0$, the larger root is

$$\widehat{\chi}^C = \frac{\vartheta + \left(-2\sqrt{I} + \frac{I-1}{I+1}\vartheta\right)}{-2I + 2\vartheta\frac{I\sqrt{I}}{I+1}} = \frac{1}{\sqrt{I}} = \widehat{\chi}^N,$$

and the other root, which is smaller, is given by

$$\widehat{\chi}^C = \frac{\vartheta - \left(-2\sqrt{I} + \frac{I-1}{I+1}\vartheta\right)}{-2I + 2\vartheta\frac{I\sqrt{I}}{I+1}} = \frac{\sqrt{I} + \frac{\vartheta}{I+1}}{-I + \vartheta\frac{I\sqrt{I}}{I+1}},$$

which is negative. Thus, there does not exist a solution $\widehat{\chi}^C$ that lies in $[\frac{1}{I}, \frac{1}{\sqrt{I}})$, meaning that the collusive equilibrium does not exist.

Case 2: if $-2\sqrt{I} + \frac{I-1}{I+1}\vartheta \leq 0$ and $-2I + 2\vartheta\frac{I\sqrt{I}}{I+1} > 0$, the smaller root is

$$\widehat{\chi}^C = \frac{\vartheta - \left(-2\sqrt{I} + \frac{I-1}{I+1}\vartheta\right)}{-2I + 2\vartheta\frac{I\sqrt{I}}{I+1}} = \frac{1}{\sqrt{I}} = \widehat{\chi}^N,$$

and the other root, which is larger, should be greater than $\widehat{\chi}^N$. Thus, there does not exist a solution $\widehat{\chi}^C$ that lies in $[\frac{1}{I}, \frac{1}{\sqrt{I}})$, meaning that the collusive equilibrium does not exist.

Case 3: if $-2\sqrt{I} + \frac{I-1}{I+1}\vartheta > 0$. In this case, we can prove that

$$-2I + 2\vartheta\frac{I\sqrt{I}}{I+1} = \sqrt{I}\left[-2\sqrt{I} + 2\vartheta\frac{I}{I+1}\right] > \sqrt{I}\left[-\frac{I-1}{I+1}\vartheta + 2\vartheta\frac{I}{I+1}\right] > 0.$$

Thus, the larger root is

$$\widehat{\chi}^C = \frac{\vartheta + \left(-2\sqrt{I} + \frac{I-1}{I+1}\vartheta\right)}{-2I + 2\vartheta\frac{I\sqrt{I}}{I+1}} = \frac{1}{\sqrt{I}} = \widehat{\chi}^N.$$

The smaller root is

$$\widehat{\chi}^C = \frac{\vartheta - \left(-2\sqrt{I} + \frac{I-1}{I+1}\vartheta\right)}{-2I + 2\vartheta\frac{I\sqrt{I}}{I+1}} = \frac{\sqrt{I} + \frac{\vartheta}{I+1}}{-I + \vartheta\frac{I\sqrt{I}}{I+1}}.$$

For $\widehat{\chi}^C$ to lie in $[\frac{1}{I}, \frac{1}{\sqrt{I}})$, we need $\widehat{\chi}^C \geq 1/I$, which implies

$$\frac{1}{I+1}\frac{\sqrt{I}-1}{\sqrt{I}+1}\vartheta \leq 1,$$

Thus, if $\vartheta \in \left(2\sqrt{I}\frac{I+1}{I-1}, (I+1)\frac{\sqrt{I}+1}{\sqrt{I}-1}\right]$, there exists a collusive equilibrium. To rule out this, we either need $\vartheta \leq 2\sqrt{I}\frac{I+1}{I-1}$ (to rule out case 3) or $\vartheta > (I+1)\frac{\sqrt{I}+1}{\sqrt{I}-1}$ (to ensure the smaller root $\widehat{\chi}^C < 1/I$ in case 3).

# C Proof of Proposition 3.4

As $\theta \to 0$ or as $\xi \to \infty$, $\lambda^C \to 1/\xi$, that is, the market approaches to the environment where prices are primarily determined by market clearing conditions. In this case, the demand of the preferred-habitat investor plays an important role. In particular, when $\theta = 0$ (or $\xi \to \infty$), the market maker's pricing rule is $\lambda^C = 1/\xi$.

Because the system is continuous, it is sufficient to show that there is a solution $\chi^C \in [\chi^M, \chi^N)$ in the environment with $\lambda^C = 1/\xi$, where $\chi^N = \frac{\xi}{I+1}$, $\chi^M = \frac{\xi}{2I}$, and $\mathbb{E}\left[\pi^N(v)\right] = \frac{\xi\sigma_v^2}{(I+1)^2}$. In this environment, equations (B.5) and (B.8) can be rewritten, respectively, as follows:

$$J^C(\chi^C) = \frac{\left(1-\xi^{-1}I\chi^C\right)\chi^C\sigma_v^2 + \frac{\rho-\rho^T}{1-\rho}\left[1-\Phi(\omega)\right]\frac{\xi\sigma_v^2}{(I+1)^2}}{1-\rho\Phi(\omega)-\rho^T\left[1-\Phi(\omega)\right]} \tag{C.1}$$

and

$$0 = \left[1-\xi^{-1}(I+1)\chi^C\right]\sigma_v^2 - \left[\rho J^C(\chi^C) - \frac{\rho-\rho^T}{1-\rho}\frac{\xi\sigma_v^2}{(I+1)^2} - \rho^T J^C(\chi^C)\right]\frac{\sigma_v}{\sigma_u}\phi(\omega)\sqrt{\frac{2}{\pi}}. \tag{C.2}$$

Therefore, $\chi^C$ is the root of the following quadratic equation:

$$0 = 1-\xi^{-1}(I+1)\chi^C - K\left[\left(1-\xi^{-1}I\chi^C\right)\chi^C - \frac{\xi}{(I+1)^2}\right],$$

where

$$K = \frac{\sigma_v}{\sigma_u}\vartheta = \frac{\sigma_v}{\sigma_u}\frac{\phi(\omega)}{\frac{1-\rho}{\rho-\rho^T}+1-\Phi(\omega)}\sqrt{\frac{2}{\pi}}. \tag{C.3}$$

Solving the above problem, we obtain

$$\widehat{\chi}^C = \frac{K+\frac{I+1}{\xi} \pm \left|\frac{K(I-1)}{I+1} - \frac{I+1}{\xi}\right|}{\frac{2KI}{\xi}}.$$

There are two cases.

Case 1: if $\frac{K(I-1)}{I+1} - \frac{I+1}{\xi} < 0$, then the smaller root is

$$\chi^C = \frac{K+\frac{I+1}{\xi}+\left(\frac{K(I-1)}{I+1}-\frac{I+1}{\xi}\right)}{\frac{2KI}{\xi}} = \chi^N.$$

The larger root must be greater than $\chi^N$. Thus, there does not exist a collusive equilibrium. To rule out this case, we need $K\xi > \frac{(I+1)^2}{I-1}$, which can be achieved by choosing a sufficiently small $\sigma_u/\sigma_v$ according to (C.3).

Case 2: if $\frac{K(I-1)}{I+1} - \frac{I+1}{\xi} > 0$, i.e., $K\xi > \frac{(I+1)^2}{I-1}$, then the larger root is

$$\chi^C = \frac{K + \frac{I+1}{\xi} + \left(\frac{K(I-1)}{I+1} - \frac{I+1}{\xi}\right)}{\frac{2KI}{\xi}} = \chi^N.$$

The smaller root is

$$\chi^C = \frac{K + \frac{I+1}{\xi} - \left(\frac{K(I-1)}{I+1} - \frac{I+1}{\xi}\right)}{\frac{2KI}{\xi}} = \frac{\frac{K\xi}{I+1} + I + 1}{KI}. \tag{C.4}$$

To have a valid collusive equilibrium, we need

$$\frac{\frac{K\xi}{I+1} + I + 1}{KI} \geq \frac{\xi}{2I},$$

which implies

$$K\xi \leq \frac{2(I+1)^2}{I-1},$$

meaning that $\sigma_u/\sigma_v$ cannot be too small.

In summary, for given parameters $T$, $\rho$, $\omega$, and $I$, we have a range of $\sigma_u/\sigma_v$ to sustain the collusive equilibrium. That is, $\sigma_u/\sigma_v$ has to be sufficiently small (in order to rule out case 1) but cannot be too small (to ensure the existence of the collusive equilibrium in case 2). That is, $\sigma_u/\sigma_v$ should be determined such that

$$K\xi \in \left(\frac{(I+1)^2}{I-1}, \frac{2(I+1)^2}{I-1}\right]. \tag{C.5}$$

# D Proof of Proposition 3.6

We prove the proposition for the environment with $\theta = 0$, so the results derived in Appendix C can be directly used. More general environments with $\theta > 0$ can be proved similarly with more complex derivations.

Without loss of generality, we restrict the analysis to the parameter choices such that the collusive equilibrium exists, meaning that condition (C.5) is satisfied. Thus, $\chi^C$ is given by (C.4).

***Proof for the profit ratio*** $\Delta^C$**.** The expected profit associated with $\chi^C$ is

$$\pi^C = (1 - \xi^{-1} I \chi^C)\chi^C \sigma_v^2.$$

Thus, $\pi^C - \pi^N$ is as follows:

$$\pi^C - \pi^N = \left(\frac{I}{I+1} - \frac{I+1}{K\xi}\right)\left(\frac{\xi}{I(I+1)} + \frac{I+1}{KI}\right)\sigma_v^2 - \frac{\xi\sigma_v^2}{(I+1)^2}.$$

The expected profit associated with $\chi^M$ is

$$\pi^M = (1 - \xi^{-1} I\chi^M)\chi^M\sigma_v^2.$$

Thus, $\pi^M - \pi^N$ is as follows:

$$\pi^M - \pi^N = \frac{\xi\sigma_v^2}{4I} - \frac{\xi\sigma_v^2}{(I+1)^2} = \xi\frac{(I-1)^2}{4I(I+1)^2}\sigma_v^2.$$

Thus, $\Delta^C$ is

$$\Delta^C = \frac{4}{(I-1)^2}\left(I - \frac{(I+1)^2}{K\xi}\right)\left(1 + \frac{(I+1)^2}{K\xi}\right) - \frac{4I}{(I-1)^2}.$$

Because $K\xi \leq \frac{2(I+1)^2}{I-1}$, $\Delta^C$ is increasing in $\xi$ and $K$. Moreover, $K = \frac{\phi(\omega)}{\frac{1-\rho}{\rho-\rho^T}+1-\Phi(\omega)}\frac{\sigma_v}{\sigma_u}\sqrt{\frac{2}{\pi}}$ is increasing in $\rho$ and decreasing in $\sigma_u/\sigma_v$. Thus, $\Delta^C$ is increasing in $\rho$ and decreasing in $\sigma_u/\sigma_v$.

To show $K$ is increasing in $\rho$, it is sufficient to prove $\frac{1-\rho}{\rho-\rho^T}$ is decreasing in $\rho$, which is equivalent to show that $f(\rho) = \log(1-\rho) - \log(\rho - \rho^T)$ is decreasing in $\rho$. The first derivative is

$$f(\rho)' = -\frac{1}{1-\rho} - \frac{1 - T\rho^{T-1}}{\rho - \rho^T} = \frac{\rho^T - 1 + T\rho^{T-1}(1-\rho)}{(1-\rho)(\rho - \rho^T)}.$$

In order to have $f(\rho)' \leq 0$, we need $h(\rho, T) = \rho^T - 1 + T\rho^{T-1}(1-\rho) < 0$. Note that $h(\rho, 1) = 0$. Thus, it is sufficient to show that $h(\rho, T)$ is decreasing in $T$ for all $\rho$. The first derivative is

$$\begin{aligned}
\frac{\partial h(\rho, T)}{\partial T} =& \rho^{T-1}\left[\rho\log(\rho) + 1 - \rho + T(1-\rho)\log(\rho)\right] \\
\leq& \rho^{T-1}\left[\rho\log(\rho) + 1 - \rho + (1-\rho)\log(\rho)\right] \\
=& \rho^{T-1}\left[1 - \rho + \log(\rho)\right] \\
<& 0.
\end{aligned}$$

Next, we show that $\Delta^C$ is decreasing in $I$. We can rewrite $\Delta^C$ as follows

$$\Delta^C = \frac{4(I+1)^2}{K\xi(I-1)^2}\left[I - 1 - \frac{(I+1)^2}{K\xi}\right].$$

We have $\Delta^C > 0$ because $K\xi > \frac{(I+1)^2}{I-1}$. The first derivative is

$$\begin{aligned}\frac{\partial \Delta^C}{\partial I} =& \frac{4}{K\xi}\left[2\left(\frac{I+1}{I-1}\right)\left(-\frac{2}{(I-1)^2}\right)\left(I-1-\frac{(I+1)^2}{K\xi}\right)+\left(\frac{I+1}{I-1}\right)^2\left(1-\frac{2(I+1)}{K\xi}\right)\right] \\ =& \frac{4}{K\xi}\frac{(I+1)(I-3)}{(I-1)^2}\left[1-\frac{2(I+1)^2}{K\xi(I-1)}\right].\end{aligned}$$

The term $1-\frac{2(I+1)^2}{K\xi(I-1)} < 0$ because $K\xi \leq \frac{2(I+1)^2}{I-1}$. Thus, $\frac{\partial \Delta^C}{\partial I} \leq 0$ for $I \geq 3$.

***Proof for the price informativeness $\mathcal{I}^C$.*** The price informativeness $\mathcal{I}^C$ is

$$\mathcal{I}^C = \log\left[\left(I\chi^C\right)^2(\sigma_v/\sigma_u)^2\right] = 2\log\left(\frac{\xi}{I+1}+\frac{I+1}{K}\right)+2\log\left(\frac{\sigma_v}{\sigma_u}\right).$$

The price informativeness $\mathcal{I}^M$ is

$$\mathcal{I}^M = \log\left[\left(I\chi^M\right)^2(\sigma_v/\sigma_u)^2\right] = 2\log\left(\frac{\xi}{2}\frac{\sigma_v}{\sigma_u}\right).$$

Thus, the relative price informativeness $\mathcal{I}^C/\mathcal{I}^M$ is

$$\frac{\mathcal{I}^C}{\mathcal{I}^M} = \frac{\log\left(\frac{\xi}{I+1}\frac{\sigma_v}{\sigma_u}+\frac{I+1}{K}\frac{\sigma_v}{\sigma_u}\right)}{\log\left(\frac{\xi}{2}\frac{\sigma_v}{\sigma_u}\right)}. \tag{D.1}$$

According to (D.1), $\mathcal{I}^C/\mathcal{I}^M$ is increasing in $I$ if $K\xi < (I+1)^2$, which is satisfied for $I \geq 3$ because of condition (C.5) for the existence of the collusive equilibrium. Moreover, $\mathcal{I}^C/\mathcal{I}^M$ is decreasing in $K$. Thus, $\mathcal{I}^C/\mathcal{I}^M$ is decreasing in $\rho$ because $K$ is increasing in $\rho$.

To study the effect of $\sigma_u/\sigma_v$, substituting out $K$ using (C.3), equation (D.1) can be rewritten as

$$\frac{\mathcal{I}^C}{\mathcal{I}^M} = \frac{\log\left(\frac{\xi}{I+1}\frac{\sigma_v}{\sigma_u}+(I+1)\frac{\frac{1-\rho}{\rho-\rho^T}+1-\Phi(\omega)}{\phi(\omega)}\sqrt{\frac{\pi}{2}}\right)}{\log\left(\frac{\xi}{2}\frac{\sigma_v}{\sigma_u}\right)}.$$

Obviously, $\mathcal{I}^C/\mathcal{I}^M$ is increasing in $\sigma_u/\sigma_v$ and decreasing in $\xi$, because $\frac{\xi}{I+1}\frac{\sigma_v}{\sigma_u} < \frac{\xi}{2}\frac{\sigma_v}{\sigma_u}$ and $(I+1)\frac{\frac{1-\rho}{\rho-\rho^T}+1-\Phi(\omega)}{\phi(\omega)}\sqrt{\frac{\pi}{2}} > 0$.

***Proof for the mispricing $\mathcal{E}^C$.*** The mispricing $\mathcal{E}^C$ is

$$\mathcal{E}^C = \left|\frac{p^C(v_t)-\mathbb{E}^C[v_t|y_t]}{\mathbb{E}^C[v_t|y_t]-\overline{v}}\right| = \left|\frac{\lambda^C-\gamma^C}{\gamma^C}\right| = \left|\frac{1}{\gamma^C}\left(\frac{\theta\gamma^C+\xi}{\theta+\xi^2}-\gamma^C\right)\right| = \left|\frac{\xi(1-\xi\gamma^C)}{\gamma^C(\theta+\xi^2)}\right|.$$

Consider the case where $\theta = 0$ and $\xi$ is sufficiently large, i.e., $\xi > 1/\gamma^C$. Thus,

$$\mathcal{E}^C = 1 - \frac{1}{\gamma^C \xi} = 1 - \frac{I\chi^C}{\xi} - \frac{\sigma_u^2}{\xi\sigma_v^2}\frac{1}{I\chi^C}.$$

Substituting out $\chi^C$ using (C.4), we obtain

$$\mathcal{E}^C = 1 - \left[\frac{1}{I+1} + \frac{I+1}{K\xi} + \frac{\sigma_u^2}{\xi\sigma_v^2}\frac{K(I+1)}{K\xi + (I+1)^2}\right]. \tag{D.2}$$

Obviously, $\mathcal{E}^C$ increases as $\xi$ increases. Moreover,

$$\begin{aligned}\frac{\partial \mathcal{E}^C}{\partial I} &= -\left[\frac{1}{K\xi} - \frac{1}{(I+1)^2} + \frac{\sigma_u^2 K}{\sigma_v^2 \xi} \times \frac{K\xi - (I+1)^2}{[K\xi + (I+1)^2]^2}\right] \\ &= -\left[(I+1)^2 - K\xi\right]\left[\frac{1}{K\xi(I+1)^2} - \frac{\sigma_u^2 K}{\sigma_v^2 \xi}\frac{1}{[K\xi + (I+1)^2]^2}\right].\end{aligned}$$

Because $K\xi \leq \frac{2(I+1)^2}{I-1}$ (equation (C.5)), it is clear that $(I+1)^2 - K\xi > 0$ for $I \geq 3$. Moreover, to ensure that $\frac{1}{K\xi(I+1)^2} - \frac{\sigma_u^2 K}{\sigma_v^2 \xi}\frac{1}{[K\xi+(I+1)^2]^2} \geq 0$, we need

$$K\xi + (I+1)^2 \geq \frac{\sigma_u}{\sigma_v} K(I+1).$$

Because $K\xi > (I+1)^2/(I-1)$ (equation (C.5)), it is sufficient to have

$$(I+1)^2/(I-1) + (I+1)^2 \geq \frac{\sigma_u}{\sigma_v} K(I+1),$$

which implies $\vartheta < I(I+1)/(I-1)$, which is satisfied when $\omega$ is not too large or $\rho$ is not very close to 1 (see equation (B.12) for the dependence of $\vartheta$ on $\omega$ and $\rho$). Thus, $\frac{\partial \mathcal{E}^C}{\partial I} < 0$ if $I \geq 3$ and $\vartheta < I(I+1)/(I-1)$.

Substituting out $K$ using (C.3), equation (D.3) can be written as

$$\mathcal{E}^C = 1 - \left[\frac{1}{I+1} + \frac{I+1}{\vartheta\xi}\frac{\sigma_u}{\sigma_v} + \frac{\sigma_u}{\xi\sigma_v}\frac{\vartheta(I+1)}{\vartheta\xi\sigma_v/\sigma_u + (I+1)^2}\right].$$

Thus, $\mathcal{E}^C$ is decreasing in $\sigma_u/\sigma_v$. Moreover, we can further rewrite the above equation as follows:

$$\mathcal{E}^C = \frac{I}{I+1} - \frac{I+1}{\xi}\frac{\sigma_u}{\sigma_v}\left[\frac{1}{\vartheta} + \frac{\vartheta}{\vartheta\xi\sigma_v/\sigma_u + (I+1)^2}\right].$$

Obviously, $\mathcal{E}^C$ is increasing in $\vartheta$ if $\vartheta < I+1$. We have shown that $K = \frac{\sigma_v}{\sigma_u}\vartheta$ is increasing in $\rho$. Thus, $\mathcal{E}^C$ is increasing in $\rho$ if $\vartheta < I+1$, which is satisfied when $\omega$ is not too large or $\rho$ is not very close to 1 (see equation (B.12) for the dependence of $\vartheta$ on $\omega$ and $\rho$).

***Proof for the market liquidity $\mathcal{L}^C$.*** The market liquidity $\mathcal{L}^C$ is

$$\mathcal{L}^C = \frac{1}{\partial |z_t + y_t| / \partial u_t} = \frac{1}{1 - \xi \lambda^C}.$$

In the environment with $\theta = 0$, market liquidity is $\mathcal{L}^C = \frac{1}{|1 - \xi \frac{1}{\xi}|} = \infty$ because prices are determined by market clearing conditions, which are not affected by the noise order flow $u_t$ in expectation. Thus, to analyze how market liquidity depends on parameter values, we consider an environment with $\theta \approx 0$ rather than $\theta = 0$. In this environment, the market liquidity $\mathcal{L}^C$ is

$$\mathcal{L}^C = \frac{1}{\left|1 - \xi \frac{\theta \gamma^C + \xi}{\theta + \xi^2}\right|} \approx \frac{1}{\left|1 - \xi \frac{\theta \gamma^C + \xi}{\xi^2}\right|} = \frac{\xi}{\theta \gamma^C} = \frac{\xi}{\theta} \left( I \chi^C + \frac{\sigma_u^2}{\sigma_v^2} \frac{1}{I \chi^C} \right).$$

Substituting out $\chi^C$ using (C.4), we obtain

$$\mathcal{L}^C = \frac{\xi}{\theta} \left[ \frac{\xi}{I+1} + \frac{I+1}{K} + \frac{\sigma_u^2}{\sigma_v^2} \frac{K(I+1)}{K\xi + (I+1)^2} \right]. \tag{D.3}$$

The market liquidity in the perfect cartel equilibrium, $\mathcal{L}^M$, is

$$\mathcal{L}^M = \frac{\xi}{\theta} \left( I \chi^M + \frac{\sigma_u^2}{\sigma_v^2} \frac{1}{I \chi^M} \right) = \frac{\xi}{\theta} \left( I \frac{\xi}{2I} + \frac{\sigma_u^2}{\sigma_v^2} \frac{1}{I \frac{\xi}{2I}} \right) = \frac{1}{\theta} \left( \frac{\xi^2}{2} + \frac{2\sigma_u^2}{\sigma_v^2} \right).$$

Thus, the relative market liquidity $\mathcal{L}^C / \mathcal{L}^M$

$$\frac{\mathcal{L}^C}{\mathcal{L}^M} = \frac{\frac{\xi^2}{I+1} + \frac{\xi(I+1)}{K} + \frac{\sigma_u^2}{\sigma_v^2} \frac{K\xi(I+1)}{K\xi + (I+1)^2}}{\frac{\xi^2}{2} + \frac{2\sigma_u^2}{\sigma_v^2}}. \tag{D.4}$$

Clearly, $\mathcal{L}^C / \mathcal{L}^M$ is decreasing in $\xi$ if $\sigma_u / \sigma_v$ is sufficiently small; $\mathcal{L}^C / \mathcal{L}^M$ is increasing in $\sigma_u / \sigma_v$ if $\xi$ is sufficiently large.

In equation (D.4), the first derivative with respect to $K$ is

$$\begin{aligned} \frac{\partial \mathcal{L}^C / \mathcal{L}^M}{\partial K} =& \frac{\xi}{\frac{\xi^2}{2} + \frac{2\sigma_u^2}{\sigma_v^2}} \left[ -\frac{I+1}{K^2} + \frac{\sigma_u^2}{\sigma_v^2} \frac{(I+1)^3}{[K\xi + (I+1)^2]^2} \right] \\ =& \frac{\xi(I+1)}{\vartheta^2 \left( \frac{\xi^2}{2} + \frac{2\sigma_u^2}{\sigma_v^2} \right)} \frac{\sigma_u^2}{\sigma_v^2} \left[ -1 + \left[ \frac{(I+1)\vartheta}{K\xi + (I+1)^2} \right]^2 \right]. \end{aligned}$$

Thus, $\frac{\partial \mathcal{L}^C / \mathcal{L}^M}{\partial K} < 0$ if $\frac{(I+1)\vartheta}{K\xi + (I+1)^2} < 1$, which is achieved if

$$\vartheta < \frac{K\xi}{I+1} + I + 1 < \frac{2(I+1)^2/(I-1)}{I+1} + I + 1 = \frac{(I+1)^2}{I-1}.$$

where the second inequality is due to condition (C.5). We have shown that $K$ is increasing in $\rho$.

Thus, $\partial \mathcal{L}^C/\mathcal{L}^M$ is decreasing in $\rho$ if $\vartheta < (I+1)^2/(I-1)$, which is satisfied when $\omega$ is not too large or $\rho$ is not very close to 1 (see equation (B.12) for the dependence of $\vartheta$ on $\omega$ and $\rho$).

In equation (D.4), the first derivative with respect to $I$ is

$$\begin{aligned}\frac{\partial \mathcal{L}^C/\mathcal{L}^M}{\partial I} =& \frac{\xi}{\frac{\xi^2}{2}+\frac{2\sigma_u^2}{\sigma_v^2}}\left[-\frac{\xi}{(I+1)^2}+\frac{1}{K}+\frac{\sigma_u^2}{\sigma_v^2}\frac{K[K\xi+(I+1)^2]-2K(I+1)^2}{[K\xi+(I+1)^2]^2}\right] \\ =& \frac{\xi}{K(I+1)^2\left(\frac{\xi^2}{2}+\frac{2\sigma_u^2}{\sigma_v^2}\right)}\left[(I+1)^2-K\xi\right]\left[1-\left[\frac{(I+1)\vartheta}{K\xi+(I+1)^2}\right]^2\right].\end{aligned}$$

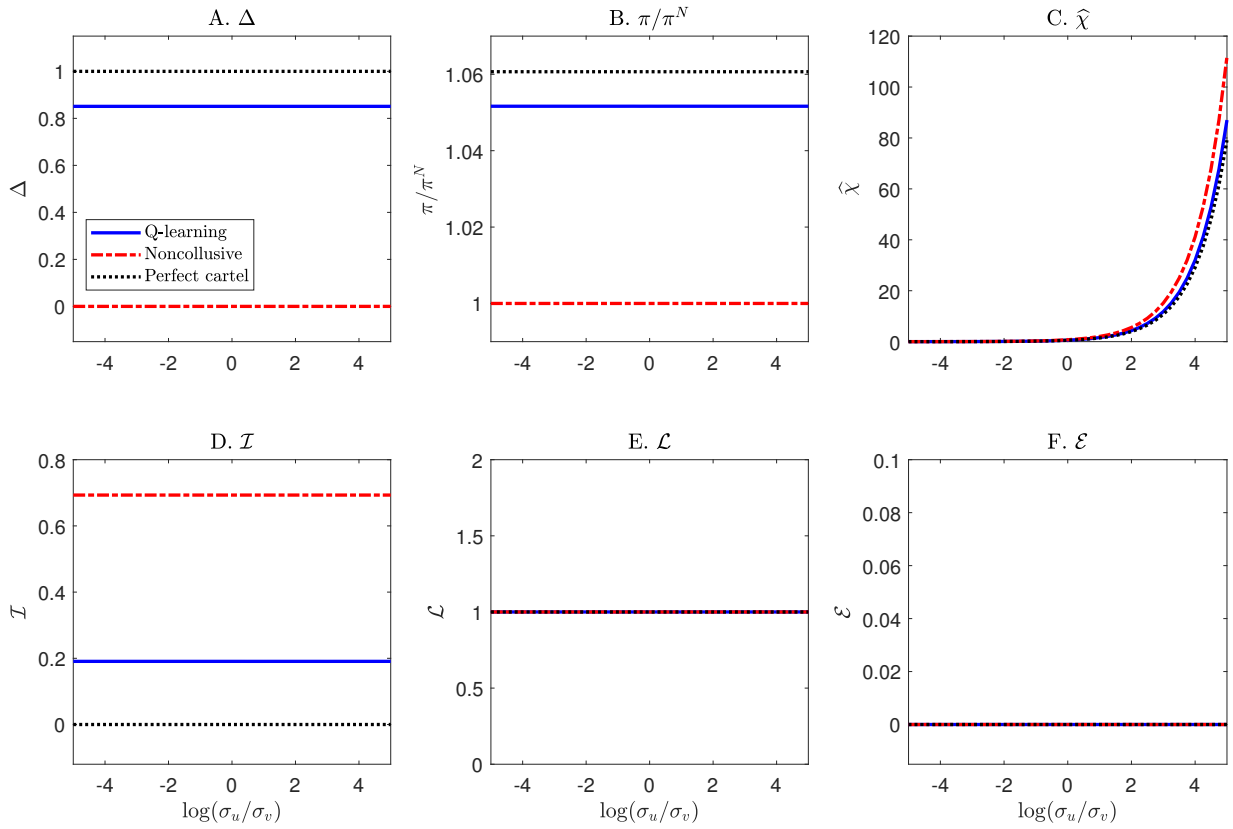
Thus, similarly, we can prove that if $\vartheta < (I+1)^2/(I-1)$, then $1-\left[\frac{(I+1)\vartheta}{K\xi+(I+1)^2}\right]^2 > 0$. Moreover, condition (C.5) implies that $(I+1)^2 - K\xi > 0$ for $I \geq 3$. Therefore, $\frac{\partial \mathcal{L}^C}{\partial I} > 0$ if $I \geq 3$ and $\vartheta < (I+1)^2/(I-1)$.

# E Environments with Efficient Prices

In this appendix section, we study informed AI speculators' behavior in the baseline economic environment except for setting $\xi = 0$, which essentially means that the preferred-habitat investor does not exist. Thus, the market maker sets prices purely for price discovery, i.e., $p_t = \mathbb{E}[v_t|y_t]$. This economic environment is similar to Kyle (1985) except for having $I = 2$ informed speculators. Proposition 3.3 in Section 3 indicates that implicit collusion cannot be sustained by any price-trigger strategies in this environment with efficient prices.

Figure A presents the average results across $N = 1,000$ simulation sessions with informed AI speculators. The blue solid lines in panels A and B show that informed AI speculators can attain an average $\Delta^C$ of 0.85 and their average profit is about 5% higher than that in the theoretical benchmark of the noncollusive equilibrium. As discussed in Section 5.2, collusion in this environment is achieved through homogenized learning biases. Similar to the property of the Kyle (1985) model, the profits of informed speculators in the theoretical benchmark of the noncollusive Nash equilibrium and the perfect cartel equilibrium are linear in the noise trading risk $\log(\sigma_u/\sigma_v)$. Thus, the red dash-dotted and black dotted lines in panels A and B are flat. Interestingly, the collusive equilibrium formed by informed AI speculators also has a constant $\Delta^C$ and $\pi^C/\pi^N$ as $\log(\sigma_u/\sigma_v)$ varies along the x-axis, exhibiting a similar scaling property with respect to $\log(\sigma_u/\sigma_v)$. Panel C shows that the informed AI speculators' order sensitivity to asset value $\widehat{\chi}^C$ increases exponentially with $\log(\sigma_u/\sigma_v)$ and linearly with $\sigma_u/\sigma_v$. This scaling property with respect to $\log(\sigma_u/\sigma_v)$ is similar to that in the theoretical benchmarks of the noncollusive Nash equilibrium and the perfect cartel equilibrium, a property that also holds in the Kyle (1985) model.

Panel D shows that due to collusion, price informativeness in the environment with informed AI speculators is lower than that in the theoretical benchmark of the noncollusive Nash equilibrium, but higher than that in the theoretical benchmark of the perfect cartel equilibrium. Moreover, as in




Note: We consider the economic environment with efficient prices as in Kyle (1985). That is, we set $\xi = 0$, implying that the asset's price $p_t$ is determined to minimize pricing errors, with $p_t = \mathbb{E}[v_t|y_t]$. The blue solid line plots the average values of $\Delta^C$, $\pi^C/\pi^N$, $\widehat{\chi}^C$, $\mathcal{I}^C/\mathcal{I}^M$, $\mathcal{L}^C/\mathcal{L}^M$, and $\mathcal{E}^C$ across $N = 1{,}000$ simulation sessions as $\log(\sigma_u/\sigma_v)$ varies. The red dash-dotted and black dotted lines represent the theoretical benchmarks of the noncollusive Nash equilibrium and perfect cartel equilibrium, respectively. The other parameters are set according to the baseline economic environment described in Section 4.7, except for $\xi = 0$.

Figure A: Implications of noise trading risks in the environment with $\xi = 0$.

the Kyle (1985) model, price informativeness remains unchanged as $\log(\sigma_u/\sigma_v)$ varies along the x-axis. Panel E shows that market liquidity is equal to 1 in this environment with efficient prices. This can be directly seen from equation (4.11). In the absence of the preferred-habitat investor, the market maker is the counterparty for informed speculators and the noise trader, and its inventory is equal to $-y_t \equiv -\sum_{i=1}^{I} x_{i,t} - u_t$. Thus, the sensitivity of the market maker's inventory to noise order flows is 1, which holds regardless of the level of noise trading risks or whether informed speculators collude. Panel F shows that mispricing in this environment is 0 because, by definition, prices are efficient, with $p_t = \mathbb{E}[v_t|y_t]$.

# F Q-Learning Market Maker

In the baseline economic environment, the market maker analyzes historical data to estimate the pricing rule (ese Section 4.2). In this appendix section, we consider the market maker adopting Q-learning algorithms to learn the pricing rule. All the results presented in the main text are similar; they do not depend on whether the market maker determines the pricing rule using statistical learning or Q-learning algorithms.

Below, we describe the Q-learning algorithm of the market maker. We consider the market maker adopting linear policies to price assets given the combined order flow $y_t$ from informed speculators and the noise trader:

$$p_t = v_t^{MM} + \lambda_t^{MM} y_t, \tag{F.1}$$

where $v_t^{MM}$ and $\lambda_t^{MM}$ are the market maker's decisions learned from its Q-learning algorithm. Specifically, the market maker's state variable is $s_t = \emptyset$ and action variables are $a_t = \{v_t^{MM}, \lambda_t^{MM}\} \in \mathcal{V} \times \Lambda$. The market maker updates its Q-matrix according to the following learning equation:

$$\begin{aligned}\widehat{Q}_{t+1}^{MM}(v_t^{MM}, \lambda_t^{MM}) =& (1-\alpha^{MM})\widehat{Q}_t^{MM}(s_t, a_t) + \alpha \Big[ (y_t - \xi(v_t^{MM} - \overline{v} + \lambda_t^{MM} y_t))^2 \\ &+ \theta(v_t^{MM} + \lambda_t^{MM} y_t - v_t)^2 + \rho^{MM} \min_{v' \in \mathcal{V}, \lambda' \in \Lambda} \widehat{Q}_t^{MM}(v', \lambda') \Big],\end{aligned} \tag{F.2}$$

where the reward in period $t$ is

$$\begin{aligned}(y_t + z_t)^2 + \theta(p_t - v_t)^2 =& (y_t - \xi(p_t - \overline{v}))^2 + \theta(p_t - v_t)^2 \\ =& (y_t - \xi(v_t^{MM} - \overline{v} + \lambda_t^{MM} y_t))^2 + \theta(v_t^{MM} + \lambda_t^{MM} y_t - v_t)^2.\end{aligned} \tag{F.3}$$

The optimal choices of $v_t^{MM}$ and $\lambda_t^{MM}$ are learned to minimize the Q-matrix. Similar to informed AI speculators' Q-learning algorithms, the market maker also conducts exploration with probability $\varepsilon_t^{MM}$ and exploitation with probability $1 - \varepsilon_t^{MM}$. In the exploration mode, the market maker randomly chooses actions $v$ and $\lambda$ over the set $\mathcal{V} \times \Lambda$.

To implement the Q-learning algorithm for the market maker, we construct discrete grid for $v_t^{MM}$ and $\lambda_t^{MM}$. Specifically, we discretize the intervals $[(1-\kappa)v^{MM}, (1+\kappa)v^{MM}]$ and $[(1-\kappa)\lambda^{MM}, (1+\kappa)\lambda^{MM}]$ into $n_v$ and $n_\lambda$ equally spaced grid points, i.e., $\mathbb{V} = \{v_1^{MM}, \cdots, v_{n_v}^{MM}\}$ and $\Lambda = \{\lambda_1^{MM}, \cdots, \lambda_{n_\lambda}^{MM}\}$. The parameters $v^{MM}$ and $\lambda^{MM}$ correspond to the optimal values in the theoretical benchmark of the noncollusive equilibrium. The parameter $\kappa > 0$ ensures that the values of $v_t$ and $\lambda_t$ chosen by the market maker can be different from the theoretical values, $v^{MM}$ and $\lambda^{MM}$.

For grid $(v_k^{MM}, \lambda_j^{MM}) \in \mathbb{V} \times \Lambda$, we initialize the market maker's Q-matrix as follows:

$$\widehat{Q}_0^{MM}(v_k^{MM}, \lambda_j^{MM}) = \frac{1}{1-\rho^{MM}} \mathbb{E}\left[ (y_t - \xi(v_k^{MM} - \overline{v} + \lambda_j^{MM} y_t))^2 + \theta(v_k^{MM} + \lambda_j^{MM} y_t - v_t)^2 \right]$$

Substituting out $y_t = I\chi^N(v_t - \overline{v}) + u_t$, we obtain

$$\begin{aligned}\widehat{Q}_0^{MM}(v_k^{MM}, \lambda_j^{MM}) =& \frac{1}{1-\rho^{MM}}\left[(1-\xi\lambda_j^{MM})^2((I\chi^N\sigma_v)^2+\sigma_u^2)+\xi^2(v_k^{MM}-\overline{v})^2\right] \\ &+ \frac{\theta}{1-\rho^{MM}}\left[(v_k^{MM}-\overline{v})^2+(\lambda_j^{MM}I\chi^N-1)^2\sigma_v^2+(\lambda_j^{MM}\sigma_u)^2\right]\end{aligned}$$

The exploration rate is $\varepsilon_t^{MM} = e^{-\beta^{MM}t}$, similar to equation (4.5). We set the parameters at $\beta^{MM} = 10^{-4}$, $\alpha^{MM} = 0.1$, $\rho^{MM} = 0.95$, $\kappa = 0.5$, and $n_v = n_\lambda = 31$. The results are similar if we choose different parameter values.

# G A Technical Appendix for Learning Biases

In this appendix, we explain why learning biases can lead informed AI speculators to exhibit collusive behavior from a technical perspective. We proceed in three steps. First, in Subsection G.1, we show that learning biases are significant when noise trading risks are high because in this case, the estimation of the Q-matrix cannot properly account for the distribution of the noise order flow $u_t$ due to the failure of the law of large numbers. Second, in Subsection G.2, we show that due to biased learning, the estimated Q-values associated with larger order flows have a larger unconditional variance. Third, in Subsection G.3, we show that large order flows are less likely to be included in the optimal strategies adopted by informed AI speculators after their Q-learning algorithms converge. In other words, biased learning would more likely lead informed AI speculators to optimally trade with small order flows, which coincide with the order flows adopted in the theoretical benchmark of the collusive Nash equilibrium. Taken together, we show that in the presence of high noise trading risks, collusive outcomes emerge due to informed AI speculators' homogenized learning biases.

## G.1 Biased Learning When Noise Trading Risks are High

First, we explain that when noise trading risks are high, there exist learning biases for the Q-matrix due to the failure of the law of large numbers.

Learning biases are caused by a generic feature of RL algorithms. As discussed in Section 2, Q-learning algorithms cannot take expectations due to the absence of knowledge about the underlying economic environment (e.g., the distribution of the noise order flow $u_t$). In each period $t$, the algorithm updates the value of one single state-action pair $(s, x_i)$ of the Q-matrix according to the currently realized profit $\pi_{i,t}$ (see equation (2.4)) rather than the expected profit $\mathbb{E}[\pi_{i,t}|s, x_i]$ as in a rational-expectations framework. Biases may exist in Q-value estimation because updating the Q-matrix sequentially based on past realized profits may not accurately reflect the expected profit, due to the failure of the law of large numbers.

To illustrate this point, we focus on a particular state-action pair $(s, x_i)$ that has been visited $T$ times in the past. Let $\tau = 1, 2, ..., T$ be the $\tau$-th visit to the state-action pair $(s, x_i)$. Let $t(\tau)$ be the

period for the $\tau$-th time that the Q-learning algorithm visits the state-action pair $(s, x_i)$. According to Equation (2.4), in each period $t$, the Q-learning algorithm only updates the state-action pair of the Q-matrix that the algorithm visits. Thus, the state-action pair $(s, x_i)$ has been updated $T$ times in the past, and these updates occur in periods $t(\tau)$ for $\tau = 1, 2, ..., T$. In other words, for each $\tau = 1, 2, ..., T$, the value of $\widehat{Q}_{i,t}(s, x_i)$ is a constant and equal to $\widehat{Q}_{i,t(\tau)+1}(s, x_i)$ from period $t = t(\tau) + 1$ to period $t = t(\tau + 1)$ and gets updated with a new value, $\widehat{Q}_{i,t(\tau+1)+1}(s, x_i)$, in period $t(\tau + 1) + 1$.

Based on equation (2.4), for the $T$-th visit to the state-action pair $(s, x_i)$, we have

$$\widehat{Q}_{i,t(T)+1}(s, x_i) = (1 - \alpha)\widehat{Q}_{i,t(T)}(s, x_i) + \alpha \left[ (v_{t(T)} - p_{t(T)})x_i + \rho \max_{x' \in \mathcal{X}} \widehat{Q}_{i,t(T)}(s_{t(T)+1}, x') \right] \tag{G.1}$$

For the $(T - 1)$-th visit to the state-action pair $(s, x_i)$, we have

$$\widehat{Q}_{i,t(T-1)+1}(s, x_i) = (1 - \alpha)\widehat{Q}_{i,t(T-1)}(s, x_i) + \alpha \left[ (v_{t(T-1)} - p_{t(T-1)})x_i + \rho \max_{x' \in \mathcal{X}} \widehat{Q}_{i,t(T-1)}(s_{t(T-1)+1}, x') \right] \tag{G.2}$$

..., and for the 1-st visit to the state-action pair $(s, x_i)$, we have

$$\widehat{Q}_{i,t(1)+1}(s, x_i) = (1 - \alpha)\widehat{Q}_{i,t(1)}(s, x_i) + \alpha \left[ (v_{t(1)} - p_{t(1)})x_i + \rho \max_{x' \in \mathcal{X}} \widehat{Q}_{i,t(1)}(s_{t(1)+1}, x') \right] \tag{G.3}$$

Because the Q-value for the state-action pair $(s, x_i)$ does not change from $t = t(\tau) + 1$ to $t = t(\tau + 1)$, we have $\widehat{Q}_{i,t(\tau)+1}(s, x_i) = \widehat{Q}_{i,t(\tau+1)}(s, x_i)$, for $\tau = 1, 2, ..., T - 1$. Thus, combining above equations, we derive

$$\begin{aligned} \widehat{Q}_{i,t(T)+1}(s, x_i) = &\sum_{\tau=0}^{T-1} \alpha(1 - \alpha)^\tau \left[ (v_{t(T-\tau)} - p_{t(T-\tau)})x + \rho \max_{x' \in \mathcal{X}} \widehat{Q}_{i,t(T-\tau)}(s_{t(T-\tau)+1}, x') \right] \\ &+ (1 - \alpha)^T \widehat{Q}_{i,0}(s, x_i). \end{aligned} \tag{G.4}$$

As $T \to \infty$, we can omit the last term and rewrite the above equation as

$$\widehat{Q}_{i,t(T)+1}(s, x_i) = \sum_{\tau=0}^{T} \alpha(1 - \alpha)^\tau \left[ (v_{t(T-\tau)} - p_{t(T-\tau)})x_i + \rho \max_{x' \in \mathcal{X}} \widehat{Q}_{i,t(T-\tau)}(s_{t(T-\tau)+1}, x') \right]. \tag{G.5}$$

By substituting out $p_{t(T-\tau)}$, the above equation becomes

$$\begin{aligned} \widehat{Q}_{i,t(T)+1}(s, x_i) = &\sum_{\tau=0}^{T} \alpha(1 - \alpha)^\tau [v_{t(T-\tau)} - \overline{v} - \lambda(y_{t(T-\tau)} - u_{t(T-\tau)})]x_i \\ &- \alpha\lambda x_i \sum_{\tau=0}^{T} (1 - \alpha)^\tau u_{t(T-\tau)} + \rho \sum_{\tau=0}^{T} \max_{x' \in \mathcal{X}} \widehat{Q}_{i,t(T-\tau)}(s_{t(T-\tau)+1}, x'). \end{aligned} \tag{G.6}$$

The term $\alpha\lambda x_i \sum_{\tau=0}^{T} (1 - \alpha)^\tau u_{t(T-\tau)}$ represents a stochastic term that depends on the noise order flow $u_{t(T-\tau)}$. With $\mathbb{E}[u_t] = 0$, the estimation for the limit value of $\widehat{Q}_{i,t(T)+1}(s, x_i)$ is unbiased only

if $\alpha\lambda x_i \sum_{\tau=0}^{T}(1-\alpha)^{\tau} u_{t(T-\tau)} = 0$ as $T \to \infty$[20], which occurs if $\alpha \to 0$. Thus, for any $\alpha > 0$, the term $\alpha\lambda x_i \sum_{\tau=0}^{T}(1-\alpha)^{\tau} u_{t(T-\tau)}$ would bias the estimate of $\widehat{Q}_{i,t(T)+1}(s, x_i)$. This is due to the failure of the law of large numbers because in general, as $T \to \infty$, we have $\alpha\lambda x_i \sum_{\tau=0}^{T}(1-\alpha)^{\tau} u_{t(T-\tau)} \neq \alpha\lambda x_i \mathbb{E}[u_t]$ unless $\alpha \to 0$.

The magnitude of learning biases depends on the importance of the term $\alpha\lambda x_i \sum_{\tau=0}^{T}(1-\alpha)^{\tau} u_{t(T-\tau)}$ relative to other terms in equation (G.6), as $T \to \infty$. Specifically, learning biases are absent when there is no noise trading risk (i.e., $\sigma_u/\sigma_v = 0$) or when $\alpha \approx 0$. Learning biases become more significant when $\sigma_u/\sigma_v$ is higher, $\lambda$ is higher, $\rho$ is lower, or $\alpha$ is higher.

## G.2 Complementarity Between Informed AI Speculators' Order and Noise Order

Second, we show that due to biased learning, the estimated Q-values associated with larger order flows have larger unconditional variances.

To begin with, we decompose the per-period profit $(v_t - p_t)x_i$ that an informed speculator $i$ receives when choosing order flow $x_i \in \mathcal{X}$ in period $t$ into two parts:

$$(v_t - p_t)x_i = \left[v_t - \overline{v} - \lambda(y_t - u_t)\right] x_i - \lambda x_i u_t. \tag{G.7}$$

The term $\left[v_t - \overline{v} - \lambda(y_t - u_t)\right] x_i$ captures the profit determined by the asset's fundamental value $v_t$ and the term $\lambda x_i u_t$ captures the profit determined by the noise order flow $u_t$. Through the term $\lambda x_i u_t$ in equation (G.7), there exists complementarity between the informed speculator's order flow $x_i$ and the noise order flow $u_t$ in determining per-period profits. This complementarity implies that, choosing larger order flows (i.e., a larger absolute value $|x_i|$) would amplify the impact of the noise order flow $u_t$ on per-period profits.

Because the estimated Q-value is the accumulated discounted per-period profits realized in the past, the complementarity between $x_i$ and $u_t$ in equation (G.7) would propagate to equation (G.6), captured by the term $\alpha\lambda x_i \sum_{\tau=0}^{T}(1-\alpha)^{\tau} u_{t(T-\tau)}$. In the absence of learning biases (i.e., when $\alpha \to 0$), we have $\alpha\lambda x_i \sum_{\tau=0}^{T}(1-\alpha)^{\tau} u_{t(T-\tau)} \approx \alpha\lambda x_i \mathbb{E}[u_t] = 0$ as $T \to \infty$, so that the unbiased estimate of the Q-value is not affected by the complementarity. However, as long as $\alpha > 0$, we would have $\alpha\lambda x_i \sum_{\tau=0}^{T}(1-\alpha)^{\tau} u_{t(T-\tau)} \neq 0$, and thus, the estimated limit Q-value is biased, due to the failure of the law of large numbers. The biased learning implies that the estimated Q-value of an informed AI speculator's particular order flow $x_i$ is path dependent, crucially depending on the realized noise order flow $u_t$ in the past periods when the informed AI speculator chose $x_i$.

Thus, in the presence of learning biases, there exists complementarity between $x_i$ and $u_t$ in determining the estimated Q-value. This complementarity implies that the estimated Q-values associated with larger order flows have larger unconditional variances.

[20] To see why unbiasedness requires $\alpha\lambda x_i \sum_{\tau=0}^{T}(1-\alpha)^{\tau} u_{t(T-\tau)} = 0$ as $T \to \infty$, note that the Q-matrix is essentially a precursor of the value function (i.e., $V_i(s) \equiv \max_{x' \in \mathcal{X}} Q_i(s, x')$, see equation (2.1)), which represents the discounted "expected" profits. In our model, the noise order $u_t$ should have no direct effect on an informed speculator's "expected" profits except for affecting its order flow $x_{i,t}$.

### G.3 Impacts of Biased Learning on Optimal Strategies

Third, we show that large order flows are less likely to be the optimal strategies adopted by informed AI speculators after their Q-learning algorithms converge. In other words, learning biases would more likely lead informed AI speculators to optimally choose small order flows, which coincide with those order flows in the theoretical benchmark of the collusive Nash equilibrium.

Before discussing why learning biases make the choice of large order flows less likely, it is useful to clarify that although informed AI speculators start their Q-learning algorithms with a mix of the exploration mode and the exploitation mode, it must be the case that the exploration rate drops to zero at some point before Q-learning algorithms to converge. In other words, in a long period of time right before Q-learning algorithms converge, informed AI speculators must be in pure exploitation mode, choosing the order flows that maximize their Q-values rather than choosing order flows randomly. Therefore, without loss of generality, we focus on the exploitation mode in our discussions below.

To fix the idea, consider a simple setting in which there is a single state $s$ and each informed AI speculator $i$'s order flow $x_i$ can take two different values, $x_i = x_S, x_L$, with $0 < x_S < x_L$, meaning that $x_L$ is a large order flow and $x_S$ is a small order flow. As discussed above, in the presence of learning biases caused by noise trading risks, there is complementarity between $x_i$ and $u_t$ in determining the estimated Q-value. Thus, relative to the estimated Q-value associated with the small order flow $x_S$, the estimated Q-value associated with the large order flow $x_L$ has a large unconditional variance (see equation (G.6)). Let $[\underline{Q}(s, x_S), \overline{Q}(s, x_S)]$ and $[\underline{Q}(s, x_L), \overline{Q}(s, x_L)]$ be the 99.9% confidence interval of the estimated Q-value for order flows $x_S$ and $x_L$, respectively. Thus, we have $[\underline{Q}(x_S), \overline{Q}(s, x_S)] \subset [\underline{Q}(s, x_L), \overline{Q}(s, x_L)]$.

Because the informed AI speculator is purely in the exploitation mode, in any period $t$, its order flow is determined according to $\text{argmax}_{x_S, x_L} \left\{ \widehat{Q}_{i,t}(s, x_S), \widehat{Q}_{i,t}(s, x_L) \right\}$. There are two cases, either $\widehat{Q}_{i,t}(s, x_L) > \widehat{Q}_{i,t}(s, x_S)$ or $\widehat{Q}_{i,t}(s, x_L) <= \widehat{Q}_{i,t}(s, x_S)$. In the first case, for $\tau > [t, t']$, the informed AI speculator would keep choosing $x_L$ to update $\widehat{Q}_{i,\tau}(s, x_L)$ while $\widehat{Q}_{i,\tau}(s, x_S)$ remains unchanged at $\widehat{Q}_{i,t}(s, x_S)$. The period $t' > t$ is the first passage time for $\widehat{Q}_{i,t'}(s, x_L) <= \widehat{Q}_{i,t'}(s, x_S)$. From period $t'$ on, the informed AI speculator switches from choosing the large order flow $x_L$ to choosing the small order flow $x_S$, and fall into the second case as described below.

In the second case, for $\tau > [t, t']$, the informed AI speculator would keep choosing $x_S$ to update $\widehat{Q}_{i,\tau}(s, x_S)$ while $\widehat{Q}_{i,\tau}(s, x_L)$ remains unchanged at $\widehat{Q}_{i,t}(s, x_L)$. The period $t' > t$ is the first passage time for $\widehat{Q}_{i,t'}(s, x_L) > \widehat{Q}_{i,t'}(s, x_S)$. From period $t'$ on, the informed AI speculator switches from choosing the small order flow $x_S$ to choosing the large order flow $x_L$, and fall into the first case as described above.

These two cases alternate over time. In one simulation session, given our convergence criterion specified in Section 4.8 (i.e., stability of optimal strategy for $T = 100,000$ consecutive periods), eventually, the optimal strategy will converge to $x_S$ with probability $\mathcal{P}$ and $x_L$ with probability $1 - \mathcal{P}$. We have $p > 0.5$ because $\underline{Q}(s, x_L) < \underline{Q}(s, x_S)$. The probability $\mathcal{P}$ is higher if the estimated $Q$-value associated with the order flow $x_L$ has a larger probability to be in the

interval $[\underline{Q}(s, x_L), \underline{Q}(s, x_S)]$, which happens when noise trading risks are higher (i.e., higher $\sigma_u/\sigma_v$ so the magnitude of learning biases is larger) or the difference in order flows is larger (i.e., larger $x_L - x_S$). This explains why learning biases make the choice of large order flows less likely.

According to our model in Section 3, the sensitivity of informed speculators' order flow to the asset's value $v_t$ is lower under collusion, i.e., $\chi^M \leq \chi^C < \chi^N$. Because informed speculator $i$'s order $x_{i,t}$ is $x_{i,t} = \chi(v_t - \overline{v})$, the absolute value of its order flows satisfies $|x_{i,t}^M| \leq |x_{i,t}^C| < |x_{i,t}^N|$ for any $v_t$, indicating that informed speculators would collude if they adopt more conservative (i.e., choosing order flows with smaller magnitude), rather than more aggressive, trading strategies. Taken together, it is clear that in the presence of high noise trading risks, homogenized learning biases lead to collusive outcomes.