AI-Powered Trading, Algorithmic Collusion, and Price Efficiency

Winston Wei Dou Itay Goldstein Yan Ji [*]

October 16, 2023

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, and market liquidity. 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 profits by strategically under-reacting to information, even in the absence of explicit communication or coordination that might breach conventional antitrust regulations. Algorithmic collusion emerges from two distinct mechanisms. The first mechanism is collusion via price-trigger strategies ("artificial intelligence"), while the second stems from learning biases ("artificial stupidity") and homogenization. The former is evident only when there is limited price efficiency and information asymmetry. In contrast, the latter persists even under conditions of high price efficiency or severe information asymmetry. 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: Asymmetric information, Price informativeness, Reinforcement learning, Market liquidity, Collusion, Homogenization.

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 Cary Frydman, Tim Johnson, Chris Jones, Larry Harris, Zhiguo He, David Hirshleifer, Jerry Hoberg, Pete Kyle, Tse-Chun Lin, Ye Luo, Semyon Malamud, Lasse Pedersen, Josh Pollet, Tom Sargent, Daniel Sokol, Liyan Yang, Neng Wang, and seminar and conference participants at Fudan, HKU, PKU/PHBS Sargent Institute Macro-Finance Workshop, Nordic Fintech Symposium, UIUC, and USC for their comments.

1 Introduction

The integration of algorithmic trading and reinforcement-learning (RL) algorithms, commonly known as AI-powered trading, has the potential to reshape capital markets fundamentally and presents new regulatory challenges. Notably, AI-powered trading bots have consistently delivered remarkable profits in the equity and forex markets, showcasing their prowess and effectiveness through established track records.[1] Additionally, supported by compelling survey evidence and industry studies,[2] AI has proven highly effective in portfolio management, with the emergence of AI advisors surpassing human advisors in actively managed equity funds. This noteworthy trend is not confined to quantitative hedge funds; it also finds manifestation among industry behemoths like BlackRock and JPMorgan, further underlining the significance and widespread adoption of AI-powered trading strategies in the investment management arena.

Consequently, policymakers, regulators, and financial market supervisors worldwide have recognized AI as a regulatory priority, directing their attention to how AI techniques are applied in financial markets to comprehend the associated implications and assess potential systemic risks.[3] Security and Exchange Commission (SEC) Chair Gary Gensler, in particular, has 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 challenge for the SEC lies in promoting competitive and efficient markets amid the rapid 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. Moreover, while many AI proponents argue that algorithms can be designed without the unconscious biases present in human decision-making, regulators acknowledge the biases inherent in reinforcement learning processes due to factors like artificial stupidity. They have repeatedly highlighted the potential for AI to inadvertently amplify biases that could lurk in their designers, further jeopardizing competition and market efficiency.

This paper aims to analyze the behavior of AI-powered trading algorithms that possess

[1] The Meta Trade Bot (https://metatradebot.com) serves as a recent example, widely covered by the media. This sophisticated, cloud-hosted AI trading system has undergone meticulous development and testing over several years, evidencing its capabilities with a commendable track record.

[2] According to BarclayHedge Poll, 56% of hedge fund respondents stated they employed AI or machine learning in their investment processes. Moreover, the JPMorgan Chase Survey found that more than 50% of the 835 institutional and professional traders surveyed believed AI technologies would exert the most significant influence on trading in the next three years.

[3] 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).

private information, investigating the significant effects they have on the market power of informed AI traders and the overall price efficiency of capital markets. It is crucial to note that AI algorithms do not merely imitate human behavior. In a similar vein to how decision theory and psychology literature have provided insights into modeling human behavior in an economic context, laying the foundation for modern finance research, comprehending 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).

Specifically, we extend the influential framework introduced by Kyle (1985) by incorporating three novel dimensions. First, it considers the involvement of multiple informed speculators within 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. Third, the model introduces a market maker who takes into account both inventory and pricing error, going beyond the limited focus on price error alone, as seen in Kyle (1985). By combining theoretical rigor with practical relevance, our model serves as a valuable laboratory for exploring the profound implications of AI-powered trading strategies on the market power of informed traders and price informativeness. Our main focus is to utilize Q-learning algorithms as a proof-of-concept illustration of algorithmic collusion and its consequent effects on price informativeness. Q-learning algorithms, known for their simplicity, transparency, and economic interpretability, have provided the foundation for various variants of reinforcement learning procedures that have driven significant advancements in the field of AI.

In our experimental framework, informed AI speculators utilize Q-learning algorithms to drive their trading decisions. Our study includes multiple informed AI speculators, a representative preferred-habitat investor, a continuum of atomistic and homogeneous noise traders, and a market maker. The market maker updates its belief about the asset's fundamental value by closely monitoring the total order flows generated by both informed AI speculators and noise traders. This belief formation process relies on "historical data" encompassing past total order flows and corresponding asset values. Furthermore, the market maker employs a statistical learning approach to understand the demand curve of the representative preferred-habitat investor. This understanding is achieved by analyzing historical data that includes past order flows of the preferred-habitat investor and corresponding market prices of the asset. Consequently, the market maker utilizes a data-driven procedure to adaptively construct its conditional expectation of the asset's value and its estimate of the preferred-habitat demand curve. Remarkably, our findings indicate that this data-driven pricing rule converges autonomously to a pricing rule that closely resembles the hypothetical scenario where the market maker possesses rational expectations, is knowledgeable about the preferred-habitat demand curve, and comprehends the collusive behavior among informed AI speculators in the market. This observation highlights the effectiveness of the data-driven approach in achieving pricing consistency despite the presence of complex market dynamics involving informed AI speculators and the preferred-habitat investor.

To ascertain whether informed AI speculators' behaviors exhibit collusion due to the intelligence of the algorithms, we begin by analyzing the fundamental theoretical properties of tacit collusion. This analysis assumes that both the informed speculators and the market maker possess rational expectations and have a comprehensive understanding of the preferred-habitat demand curve. We highlight how tacit collusion changes across diverse market structures and information environments. This theoretical investigation enables us to establish a baseline understanding of collusive behavior in the presence of asymmetric information and the market maker's endogenous strategic pricing rules. Furthermore, 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 rational expectations and perfect knowledge of the preferred-habitat demand curve. As a particularly noteworthy contribution, we establish a novel theory on the impossibility of collusion under information asymmetry. This theory presents a distinctive and intuitive perspective, emphasizing that informed speculators cannot exploit pricing errors to achieve collusive outcomes, given the already high level of efficiency in prices that accurately reflects the fundamental value. The value of this theory lies in its theoretical insights and novelty, as it illuminates a distinct mechanism separate from existing theories on the impossibility of collusion under information asymmetry in the context of product market competition, as previously posited by Abreu, Milgrom and Pearce (1991) and Sannikov and Skrzypacz (2007).

Furthermore, as another theoretical contribution, our research demonstrates that in scenarios where preferred-habitat investors play a substantial role in price formation, resulting in prices that are not highly efficient, tacit collusion among informed speculators can be sustained through the use of price-trigger strategies. The effectiveness of these strategies is contingent upon the level of information asymmetry in the market, which should not be overly severe, and the number of informed speculators, which should not be excessively large. In addition, we show that collusion capacity increases and price informativeness reduces, when the number of informed speculators drops, information asymmetry reduces, the subjective rate of time preference ("impatience") declines, or preferred-habitat demand elasticity rises.

Our numerical findings provide compelling evidence that informed AI speculators can collude and achieve supra-competitive profits by strategically manipulating excessively low order flows, even in the absence of explicit coordination that would constitute an antitrust infringement. The significance of information exchange in collusion among multiple firms operating within a market has been well-established in existing research in experimental economics and game theory. To demonstrate this key idea, we intentionally focus on relatively naive Q-learning algorithms that solely rely on one-period-lagged asset prices, without incorporating more extensive lagged data or their own order flow information. Remarkably, our study illustrates that these algorithms can intelligently communicate and collaborate using just one period of historical prices, when the trading environment is excessively complex relative to the AI algorithms. These algorithmic collusion behaves exactly like what the theory would predict across diverse market structures and information environments. Even more strikingly, in the scenarios where the trading environment is too challenging or complex for the AI algorithms, informed AI speculators can still collude and achieve supra-competitive profits by manipulating excessively low order flows, as long as the algorithms are equally naive. Therefore, the emergence of algorithmic collusion can be attributed to two distinct sources or mechanisms.

The first mechanism, known as algorithmic collusion through price-trigger strategies or collusion due to "artificial intelligence," bears resemblance to its theoretical counterpart – collusion through price-trigger strategies – when both the informed speculators and the market maker possess rational expectations and have a comprehensive understanding of the preferred-habitat demand curve. When one informed AI speculator deviates from the agreed collusive order flow level by increasing its magnitude intentionally or randomly, the asset price reacts unfavorably for the other informed AI speculator. Consequently, they seek to optimize their own performance by selecting a different order flow level, often leading to a more aggressive approach. This, in turn, negatively impacts the deviating informed AI speculator. While the underlying mechanisms between the algorithmic collusion and the economic collusion may differ, despite that both are through price-trigger strategies, the resulting patterns exhibit notable similarities. At the heart of both, the punishment threat effectively serves as a deterrent to discourage individual speculators from breaking the collusion and pursuing higher profits.

Algorithmic collusion through price-trigger strategies introduces a paradoxical situation regarding price informativeness. This paradox arises because algorithmic collusion through price-trigger strategies relies on the informativeness of prices, specifically the ability of an informed AI speculator to deduce the order flows of other informed AI speculators from observed prices. When price informativeness is high, it becomes easier for an informed AI speculator to accurately infer the order flows of others, thus facilitating algorithmic collusion. The paradox emerges because the presence of strong price informativeness, where prices are sensitive to new information and are not primarily driven by noise trading flows, makes it simpler for informed AI traders to discern each other's order flows. This heightened ability to deduce others' actions strengthens collusion among the speculators. However, as collusion becomes stronger, it compromises the overall price informativeness of the market. The collusion among informed AI speculators distorts the information content of prices, reducing their ability to accurately reflect underlying fundamentals and impeding the efficiency of price formation. Consequently, in a capital market where AI-powered trading is prevalent and algorithmic collusion through price-trigger strategies exists, perfect price informativeness or perfect price efficiency becomes unattainable.

The second mechanism, referred to as algorithmic collusion through learning bias (sometimes termed "artificial stupidity")[4] and homogenization, relies upon a hub-and-spoke conspiracy.[5] Despite the learning bias originating from the algorithms' intrinsic imperfections, informed speculators, even while ostensibly competing, may exploit these shared biased algorithms to sustain supra-competitive profits, as a form of this hub-and-spoke conspiracy. Johnson and Sokol (2021) underscore the prevalence of this "hub-and-spoke" AI-driven algorithmic collusion in the context of e-commerce platforms. This conspiracy tends to surface when informed speculators base their AI-driven trading systems on the same foundational models, potentially leading to a high level of homogenization as noted by Bommasani et al. (2022), among others. In the context of the Q-learning process, the emergence of learning bias is directly tied to the inconsistency in statistical learning, which results from exploitation. This inherently biased algorithm prompts the informed speculator to under-react to its private information in the trading, relative to the optimal trading strategy in the non-collusive competitive setting. Such an under-reaction can lead to the realization of supra-competitive profits, a scenario more likely to occur if there's a widespread homogenization in the use of algorithms among speculators. This situation is further compounded when no speculator seeks to gain an advantage by utilizing superior algorithms in contrast to others.

[4] Learning bias, also known as algorithm bias or AI bias, manifests when an algorithm produces results that are systemically skewed due to erroneous assumptions in the learning process.

[5] In the setting of product market competition, a hub-and-spoke conspiracy is a metaphor used to describe a cartel that includes a firm at one level of a supply chain, such as a buyer or supplier, who acts like the "hub" of a wheel. Vertical agreements up or down the supply chain act as the "spokes." Anti-competitive effects can occur, when multiple competitors use the same AI pricing algorithm supplied by a common service provider who acts as a hub (e.g., Johnson and Sokol, 2021).

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, uncovering not only 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 an 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, but that achieving these gains may require policies that condition on past behavior and treat sellers in a non-neutral fashion. Harrington (2019) delves into critical policy issues surrounding the definition of collusion. Harrington (2019) provides discussions on policy issues, 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 one of the first few that study how the widespread adoption of AI-powered trading strategies would affect capital markets. The work of Colliard, Foucault and Lovo (2022) is closely related to our research as it also explores the emergence of algorithmic collusion in capital markets through the interactions of Q-learning algorithms.

However, there are notable differences in their focus compared to our paper. Specifically, Colliard, Foucault and Lovo (2022) concentrate on AI-powered oligopolistic market makers, whereas our study centers on AI-powered oligopolistic informed traders who face perfectly competitive market makers. Colliard, Foucault and Lovo (2022) delve into how AI-powered market makers strategically mitigate adverse selection by leveraging their market power, which is sustained through algorithmic collusion. Their research sheds light on the strategies employed by market makers to cope with the challenges posed by private information and to optimize their outcomes within an oligopolistic environment. In contrast, our paper complements the aforementioned works by examining how AI-powered informed traders exploit their private information and exert their market power through algorithmic collusion. We investigate the dynamics and implications of collusion among informed traders in the presence of perfectly competitive market makers. By focusing on the perspective of informed traders, we provide additional insights into the strategies employed by these participants to leverage their private information and maximize their profits through collusion.

2 Model

This model extends the influential framework introduced by Kyle (1985) by incorporating three novel dimensions. First, it considers the involvement of multiple informed speculators within 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, the model introduces a market maker who takes into account both inventory and pricing error, going beyond the limited focus on price error alone, as seen in Kyle (1985).

By blending theoretical rigor with practical relevance, this model offers a valuable laboratory for exploring the implications of AI-powered trading behaviors 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 behaviors in simulated experiments.

2.1 Economic Environment

Time is discrete, indexed by $t = 1, 2, \cdots$, and it runs forever. There are $I \geq 2$ risk-neutral informed speculators, a representative noise traders, 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 that 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, the fundamental value of the asset, v_t, becomes observed by all agents. It is drawn from a normal distribution $N(\overline{v}, \sigma_v^2)$. Here, $\overline{v}$ represents the mean and σ_v^2 the variance of the distribution, with $\overline{v}$ set to 1 for convenience. After the realization of the fundamental value v_t, trading profits for all agents in period t are determined.

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, at the beginning of the period, 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 σ_u^2 is the variance. The informed speculators perfectly know the value v_t, but they are unaware of u_t when submitting their order flows. The informed speculators are indexed by $i \in \{1, \cdots, I\}$. Each speculator i, whose order flow is $x_{i,t}$, understands that its choice of $x_{i,t}$ 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:

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \rho^t (v_t - p_t) x_{i,t}\right], \tag{2.1}$$

where $\rho \in (0, 1)$ is the subject 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 demand curve for the net trading flow z_t that slopes downward:

$$z_t = -\xi(p_t - \overline{v}), \quad \text{with } \xi > 0. \tag{2.2}$$

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 undervalued. This demand curve is proportional to the spread between the ex-ante expected fundamental value and the market price. Graham (1973) calls this spread a safety margin.

The average holding of the preferred-habitat investor in this type of asset, denoted as $\overline{z}$, is often substantial. Consequently, this leads to an approximately very small price elasticity of demand, represented as $\varepsilon \approx \xi/\overline{z}$. 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 demand curve of the preferred-habitat investor, as specified in equation (2.2), 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 fundamental value, $\overline{v}$, in the market, the preferred-habitat investor will buy more of the asset. Analogous to Kyle and Xiong (2001), we can justify the demand curve, as outlined in (2.2), through 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 η, then the demand curve has the functional form of (2.2), where the slope ξ 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:

$$\widetilde{z}_t = -u_t + \xi(p_t - \overline{v}), \tag{2.3}$$

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

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 noise traders, represented by

$y_t = \sum_{i=1}^{I} x_{i,t} + u_t$, as well as the order flow of the preferred-habitat investor, denoted by z_t. However, the market maker cannot distinguish between order flows from informed speculators and noise traders. Instead, they can only make statistical inferences about the fundamental value v_t based on the combined order flow $y_t = \sum_{i=1}^{I} x_{i,t} + u_t$ and not on individual order flows. The market maker sets the price p_t to jointly minimize inventory and pricing errors according to the following objective function:

$$\min_{p_t} \mathbb{E}\left[(y_t + z_t)^2 + \theta (p_t - v_t)^2 \middle| y_t \right], \tag{2.4}$$

where $\theta > 0$ represents the weight the market maker places on minimizing pricing errors. Here, $\mathbb{E}\left[\cdot | y_t\right]$ 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, as described in (2.4), captures both the inventory cost and asymmetric information faced by the market maker. The term $(y_t + z_t)^2$ represents the inventory-holding costs borne by the market maker. Its 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. Assigning a weight, represented by θ, to the pricing error serves as a reduced-form method to encapsulate factors such as the benefits of dynamically increasing the trading flows from a growing client base or competing with other trading platforms.[6] As θ approaches zero, the price 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 θ increases towards infinity, the price p_t is primarily determined by the pricing-error minimization condition, $p_t = \mathbb{E}\left[v_t | y_t\right]$, as in the model of Kyle (1985).

Given the repeated-game nature of this framework involving multiple informed speculators, various equilibria with tacit collusion 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 this 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

[6] Similarly, in the context of e-commerce platforms, it's often assumed that the platform aims to maximize a weighted average of per-unit fee revenues and consumer surplus (see, e.g., Johnson, Rhodes and Wildenbeest, 2023). Assigning a weight to the consumer surplus in this context acts as a reduced-form method. This captures aspects such as the benefits of dynamically expanding the consumer base over time and competing with rival platforms.

the trading strategy of the informed speculators is characterized by

$$x_{i,t} = \chi(v_t - \overline{v}), \text{ for any } i = 1, \cdots, I. \tag{2.5}$$

The first-order condition of the minimization problem (2.4) leads to

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

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

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

Therefore, the pricing rule of the market maker is

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

2.2 Noncollusive Nash Equilibrium

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

$$x^N(v_t) = \underset{x_i}{\operatorname{argmax}} \mathbb{E}\left[\left(v_t - p_t\right) x_i \middle| v_t\right], \tag{2.6}$$

where $\mathbb{E}\left[\cdot | v_t\right]$ 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 $p_t = p^N(y_t)$. Here, $p^N(\cdot)$ is a pricing function that is determined in the equilibrium characterized as follows:

$$p^N(y_t) = \overline{v} + \lambda^N y_t, \text{ with } \lambda^N = \frac{\theta\gamma^N + \xi}{\theta + \xi^2} \text{ and } \gamma^N = \frac{I\chi^N}{(I\chi^N)^2 + (\sigma_u/\sigma_v)^2}. \tag{2.7}$$

And, y_t is the combined order flow of informed speculators and noise traders, characterized by

$$y_t = x_i + (I-1)x^N(v_t) + u_t. \tag{2.8}$$

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

Proposition 2.1. *The order flow and price in the non-collusive Nash equilibrium are*

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

where χ^N and λ^N satisfy

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

The expected profit 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, that is, $\mathcal{I}^N = \log\left[\left(I\chi^N\right)^2 (\sigma_v/\sigma_u)^2\right]$.

2.3 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

$$x^M(v_t) = \underset{x_i}{\operatorname{argmax}}\, \mathbb{E}\left[(v_t - p_t)\, x_i \middle| v_t\right], \tag{2.9}$$

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 $p_t = p^M(y_t)$. Here, $p^M(\cdot)$ is a pricing function that is determined in the equilibrium characterized as follows:

$$p^M(y_t) = \overline{v} + \lambda^M y_t, \text{ with } \lambda^M = \frac{\theta\gamma^M + \xi}{\theta + \xi^2} \text{ and } \gamma^M = \frac{I\chi^M}{(I\chi^M)^2 + (\sigma_u/\sigma_v)^2}. \tag{2.10}$$

And, y_t is the combined order flow of informed speculators and noise traders, characterized by

$$y_t = Ix_{i,t} + u_t. \tag{2.11}$$

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

Proposition 2.2. *The order flow and price in the perfect cartel equilibrium are*

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

where χ^M and λ^M satisfy

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

The expected profit 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, that is, $\mathcal{I}^M = \log\left[\left(I\chi^M\right)^2(\sigma_v/\sigma_u)^2\right]$.

2.4 Collusive Nash Equilibrium

Information asymmetry is a significant characteristic of capital markets, rendering standard grim trigger strategies less viable due to the challenges in accurately observing and monitoring each other's actions.[7] 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 non-collusive competition to occur with positive probability. 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 non-collusive reversion. In the collusive regime, informed speculators implicitly coordinate on order flows less aggressive than the order flows in the static non-collusive Nash equilibrium. If the price breaches a critical level, suspicion of cheating arises, leading to a non-collusion reversion. In the punishment regime, informed speculators trade non-collusively with 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 traders choose the same collusive order flow, denoted by $x^C(v_t)$. Such

[7] Tacit collusion sustained by grim trigger strategies has been a subject of extensive research, with pioneering work by Fudenberg and Maskin (1986) and Rotemberg and Saloner (1986), among other notable contributors. Recent studies have delved 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).

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 non-collusively for the duration (we will specify this to be $T-1$ periods) of a reversionary episode. At time 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, at time 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 price trigger is violated at $t-T$ and $s_{t-T} = 0$; otherwise, $s_t = 1$.

Similar to Green and Porter (1984), we assume that the state variable s_t is a common knowledge to all agents. When $s_t = 1$, the equilibrium order flows and price are characterized in Section 2.2. We now focus on characterizing the equilibrium order flow $x^C(v_t)$ and price p_t^C for the case of $s_t = 0$.

We focus on linear policy functions for the case of $s_t = 0$:

$$x^C(v) \equiv \chi^C(v - \overline{v}), \tag{2.12}$$

$$p^C(y) = \overline{v} + \lambda^C y. \tag{2.13}$$

We specify the price-trigger function $q(v)$ using the expected price under the coordinated trading conditional on v, denoted by $\overline{p}^C(v) \equiv \mathbb{E}\left[p^C(y)|v\right]$. Specifically, plugging (2.12) into (2.13) and taking expectation over u, we obtain that $\overline{p}^C(v) \equiv \overline{v} + \lambda^C I \chi^C(v - \overline{v})$. The trigger price is specified as follows:

$$q(v) \equiv \begin{cases} \overline{p}^C(v) + \lambda^C \sigma_u \omega, & \text{if } v > \overline{v} \\ \overline{p}^C(v) - \lambda^C \sigma_u \omega, & \text{if } v < \overline{v}, \end{cases} \tag{2.14}$$

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

Equation (2.14) warrants further in-depth discussion on several important points. 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 non-collusive Nash equilibrium once the market price goes above 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 non-collusive Nash equilibrium once the market price falls below 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 different ways of specifying the functional form of the price threshold $q(v)$, we focus on a specification that ensures a linear model solution as in Kyle (1985) and statistically meaningful. Each informed investor can infer from the price $p_t = p^C(y_t)$ that the noise trading order should be $\hat{u}_t = [p_t - q(v_t)]/\lambda^C$. If $\hat{u}_t$ is excessively positive when $v_t > \overline{v}$, say $\hat{u}_t > \omega\sigma_u$ for certain constant $\omega > 0$, the informed investor would suspect that some other informed investors might have deviated from the implicit agreement. Analogously, if $\hat{u}_t$ is excessively negative when $v_t < \overline{v}$, say $\hat{u}_t < -\omega\sigma_u$ for certain constant $\omega > 0$, the informed investor would suspect that some other informed investors might have deviated from the implicit agreement. Fourth, the multiplier σ_u ensures that the probability of price-trigger violation is independent of the magnitude of noisy trading, σ_u, in the collusive Nash equilibrium.

Given that $s_t = 0$, let $J^C(\chi_i)$ denote each informed trader 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 $x^C(v_t)$. That is,

$$\begin{aligned} J^C(\chi_i) = {} & \mathbb{E}\left[\left(v_t - p^C(y_t)\right)\chi_i(v_t - \overline{v})\right] \\ & + \rho J^C(\chi_i)\mathbb{P}\left\{\text{Price trigger is not violated in period } t \middle| \chi_i, \chi^C\right\} \\ & + \mathbb{E}\left[\sum_{\tau=1}^{T-1}\rho^\tau \pi^N(v_{t+\tau}) + \rho^T J^C(\chi_i)\right]\mathbb{P}\left\{\text{Price trigger is violated in period } t \middle| \chi_i, \chi^C\right\}, \end{aligned} \tag{2.15}$$

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

$$y_t = \chi_i(v_t - \overline{v}) + (I-1)x^C(v_t) + u_t. \tag{2.16}$$

The probability of price trigger violation is

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

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, prices are determined by the market maker, who sets them approximately at the expectation of the fundamental value, conditional on the observed total order flow. In other words, efficient prices in this context are unbiased estimates of the fundamental asset value, and they minimize pricing errors. The proof can be found in Appendix B.

Proposition 2.3 (Impossibility of Collusion When Efficient Prices Prevail)**.** *If θ is large or ξ 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) price impact of informed speculators' order flows needs to be sufficiently low to ensure that there is sufficient room for significant informational rents.

However, in cases where θ is large or ξ is small, the environment closely resembles the standard Kyle benchmark (Kyle, 1985). In this scenario, it is important to note that λ^C becomes approximately equal to γ^C. Importantly, in this case, low price impact endogenously reflects a proportionally high information asymmetry, captured by σ_u/σ_v. Despite the aggressive trading by informed speculators induced by low price impact, the negative effect of information asymmetry and the positive effect of informed order flows on price informativeness balance each other out in this environment. As a result, the two necessary conditions (i) and (ii) cannot coexist simultaneously in an environment close to the standard Kyle benchmark environment, where efficient prices prevail.

Proposition 2.3 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 informativeness, "false positive" errors, made by triggering punishments, occur on the equilibrium path disproportionately often, erasing all benefits from collusion. In contrast, Proposition 2.3 offers a distinctive intuitive perspective, highlighting that informed speculators cannot exploit pricing errors to achieve collusive outcomes due to the already high level of efficiency in prices, which accurately reflect the 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 2.3 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 findings contribute to 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 plays an important role in price formation (i.e., when prices are not very efficient). But, when information asymmetry, captured by σ_u/σ_v, is too large, no collusion can be sustained through price-trigger strategies even though prices are not very efficient. Moreover, when the number of informed speculators, denoted by I, is too large, no collusion can be sustained through price-trigger strategies even though prices are not very efficient. The proof is in Appendix C.

Proposition 2.4 (Existence of Collusion with a Significant Preferred-Habitat Investor)**.** *If θ is sufficiently small or if ξ is sufficiently large, there exists a collusive Nash equilibrium that can be sustained by price-trigger strategies for σ_u/σ_v and I that are not too large*

When σ_u/σ_v is too large, price informativeness is low, and thus price-trigger strategies are difficult to sustain. This is because when prices are not informativeness, agents to make "false positive" errors by triggering punishments on the equilibrium path disproportionately often, erasing all benefits from collusion. The key idea is exactly the same as that of Abreu, Milgrom and Pearce (1991) and Sannikov and Skrzypacz (2007).

If θ is small or if ξ is large, the price is primarily determined by the market clearing condition, which is probably not an unbiased estimate of the fundamental value with minimum pricing errors. If market clearing condition dominates, low price impact does not reflect a proportionally high information asymmetry; as a result, it allows informed speculators trade aggressively, thereby leading to higher price informativeness. Consequently, the necessary conditions (i) and (ii) can hold simultaneously when the preferred-habitat investor plays an important role in price formation.

Properties of Collusion Sustained by Price-Trigger Strategies. To characterize whether informed speculators trade in a tacitly collusive manner based on observable outcomes, it is necessary to derive the testable properties of collusion.

Proposition 2.5 (Supra-competitive nature of collusion)**.** *In the price-trigger collusive equilibrium, it holds that*

$$\pi^M \geq \pi^C > \pi^N, \tag{2.17}$$

If we define $\Delta^C \equiv \dfrac{\pi^C - \pi^N}{\pi^M - \pi^N}$*, inequalities in (2.17) can be summarized as* $\Delta^C \in (0, 1]$.

Clearly, a greater Δ^C signifies a higher collusion capacity. We use Δ^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). Below, we derive how collusion capacity, Δ^C, and price informativeness, $\mathcal{I}^C$, change across various market structures and information environments. The proof of the following proposition can be found in Appendix D.

Proposition 2.6 (Effects of Market Structures and Information Environments)**.** *If θ is sufficiently small or if ξ is sufficiently large, the price-trigger collusive Nash equilibrium satisfies the following properties:*

(i) $I \uparrow \quad \Longrightarrow \quad \Delta^C \downarrow \quad \& \quad \mathcal{I}^C \uparrow$

(ii) $\sigma_u/\sigma_v \uparrow \quad \Longrightarrow \quad \Delta^C \downarrow \quad \& \quad \mathcal{I}^C \uparrow$

(iii) $\rho \uparrow \Longrightarrow \Delta^C \uparrow \quad \& \quad \mathcal{I}^C \downarrow$

(iv) $\xi \uparrow \Longrightarrow \Delta^C \uparrow \quad \& \quad \mathcal{I}^C \downarrow$

3 AI-Powered Trading Algorithms

The theoretical results above hinge on the assumption that the informed speculators and the market maker have rational expectations in the sense that they can perfectly figure out (i) the order flows of other informed speculators (known by informed speculators but not the market maker due to information asymmetry), (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 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 (2.12) to (2.14). Meanwhile, this also allows the market maker perfectly understands 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 above based on stringent, and at times, unrealistic assumptions. Specifically, in this section, we investigate the capability of RL algorithms to attain tacit collusion and generate supercompetitive trading profits when the machines have no direct knowledge of order flows from their counterparts or are oblivious to the distribution of noisy trading flows and the fundamental values of assets. If these algorithms demonstrate such capability, our study further delves into the mechanisms driving these algorithmic collusive behaviors. RL is the type of machine learning in which the algorithm learns by itself through autonomous trial-and-error experimentation.

3.1 Q-Learning

We examine Q-learning algorithms, exploring whether AI-powered trading algorithms can autonomously achieve tacit collusion under asymmetric information, without the overt acts of communication or agreements typically seen in competition law infringements (Harrington, 2018). Our experimental design and methodology are similar to the studies of Calvano et al. (2020) and Asker, Fershtman and Pakes (2022). They explored product market competition without the complexities of asymmetric information or endogenous pricing rules.

Our main objective is to employ Q-learning algorithms as a proof-of-concept illustration, shedding light on the potential of algorithmic collusion and its consequential effects on the informativeness of prices. While reinforcement learning encompasses different variants (e.g., Watkins and Dayan, 1992; Sutton and Barto, 2018), our choice to focus on Q-learning is motivated by several reasons. First, Q-learning serves as a foundational framework for numerous reinforcement learning 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 reinforcement learning algorithms.

The fundamental rationale behind the Q-learning algorithm, akin to all reinforcement learning approaches, rests on the principle that actions leading to higher past payoffs are prioritized for future occurrences compared to actions generating lower profits. Consequently, through multiple rounds of exploration and experimentation, Q-learning algorithms can adapt their actions towards achieving optimal outcomes, even in the absence of prior knowledge concerning the problem at hand. Below, we outline the Q-learning algorithm employed by a generic informed speculator $i \in \{1, \cdots, I\}$.

Bellman Equation and Q-Function. Informed speculator i's intertemporal optimization problem, specified in (2.1), is usually solved recursively using the dynamic programming approach and the associated Bellman equation:

$$V_i(s) = \max_{x \in \mathcal{X}} \left\{ \mathbb{E}\left[(v-p)x|s,x\right] + \rho \mathbb{E}\left[V_i(s')|s,x\right] \right\}, \tag{3.1}$$

where $\mathcal{X}$ is the set of available actions, s is the current state, s' represents the state in the next period, the first term on the right-hand side, $\mathbb{E}\left[(v-p)x|s,x\right]$, is the expected payoff of the current period, and the second term, $\rho \mathbb{E}\left[V_i(s')|s,x\right]$, is the continuation value.

The Bellman equation (3.1) reflects the recursive formulation of dynamic control problems, as described by Bellman (1954) and Ljungqvist and Sargent (2012), among others. The value function $V_i(s)$, a function of the state s, and its associated Bellman equation focus on the equilibrium path. However, instead of focusing solely on the optimal value of each state $V_i(s)$ along the equilibrium path, we can extend our analysis to the counterfactual value of each state-action pair, denoted as $Q_i(s,x)$, which captures scenarios even off the equilibrium path. By definition, $Q_i(s,x)$ is the same value as what's in the curly brackets of the Bellman equation (3.1):

$$Q_i(s,x) = \mathbb{E}\left[(v-p)x|s,x\right] + \rho \mathbb{E}\left[V_i(s')|s,x\right]. \tag{3.2}$$

Intuitively, the Q-function value, $Q_i(s,x)$, can be interpreted as the quality of action x at 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 (3.2), we can establish a recursive formula for the Q-function as follows:

$$Q_i(s,x) = \mathbb{E}\left[(v-p)x|s,x\right] + \rho \mathbb{E}\left[\max_{x' \in \mathcal{X}} Q_i(s',x') \middle| s,x \right]. \tag{3.3}$$

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

State Variables. State variables, s_t, are essential for characterizing the recursive relation presented in equation (3.3). While 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 and the private information available to the speculator. We utilize the smallest possible set of state variables in s_t that can generate tacit collusion sustained by price-trigger strategies. Drawing from the insights in Section 2.4, we include the market price of the asset from the preceding period $t-1$, denoted by p_{t-1}, as part of s_t. We incorporate v_t instead of v_{t-1} in the state variable s_t because informed AI speculators engage in trading activities in period t after observing v_t at the beginning of period t, thereby necessitating the inclusion of v_t as part of the state variable in period t. Consequently, 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 history for decision-making, similar to the approach in Calvano et al. (2020). One could also include the informed AI speculator's own lagged order flow $x_{i,t-1}$, a piece of private information only known by informed AI speculator i, and even more lagged asset prices and order flows, as a state variable. In our simulation experiments, we observed 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 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 $s_t \equiv \{p_{t-1}, v_t\}$ exhibits a convergence speed significantly faster than those incorporating a more extensive list of state variables.[8]

Q-Learning Algorithm. If informed AI speculators possessed knowledge of their Q-matrices, determining the optimal actions for any given state would be straightforward. In essence, Q-learning algorithms serve as methods to estimate this Q-matrix without knowing the underlying distribution $\mathbb{E}\left[\cdot|s,a\right]$ or observing sufficient off-equilibrium pairs (s,x) in the data. These algorithms address both challenges concurrently: They employ Monte Carlo methods, backed by the law of large numbers, to estimate the underlying distribution $\mathbb{E}[\cdot|s,x]$, while simultaneously conducting trial-and-error experiments to produce off-equilibrium counterfactuals.

The iterative experimentation starts from an arbitrary initial Q-matrix of informed AI speculator i, denoted by $\widehat{Q}_{i,0}$, and updates the estimated Q-matrix $\widehat{Q}_{i,t}$ recursively. Observing $s_t \equiv \{p_{t-1}, v_t\}$, informed AI speculator i chooses its order flow $x_{i,t}$, following one of two experimentation modes, which we describe in detail below. After receiving the total quantity of market orders, the market maker determines the price p_t according to its own pricing rules described in Subsection 3.2.

The evolution of informed AI speculator i's state variable $s_{i,t}$ is given by $s_{i,t+1} \equiv \{p_t, v_{t+1}\}$, where v_{t+1} is randomly drawn from the distribution $N(\overline{v}, \sigma_v^2)$. The price p_t depends on the noise trading flow, which remains unknown to informed AI speculators when they make decisions.

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

The Q-learning algorithm employs a recursive update process for informed AI speculator i to refine its estimated Q-matrix. The learning equation governing this update is as follows:

$$\widehat{Q}_{i,t+1}(s_t, x_{i,t}) = (1-\alpha)\underbrace{\widehat{Q}_{i,t}(s_t, x_{i,t})}_{\text{Past knowledge}} + \underbrace{\alpha\left[(v_t - p_t)x_{i,t} + \rho \max_{x\in\mathcal{X}} \widehat{Q}_{i,t}(s_{t+1}, x)\right]}_{\text{Present learning based on a new experiment}}, \tag{3.4}$$

where $\alpha \in [0,1]$ captures the learning rate, s_t is the state that the iteration t concentrates on, s_{t+1} is randomly drawn from the Markovian transition probability conditional on s_t, and the action variable $x_{i,t}$ is chosen as follows:

$$x_{i,t} = \begin{cases} \operatorname{argmax}_{x\in\mathcal{X}} \widehat{Q}_{i,t}(s_t, x), & \text{with prob. } 1-\varepsilon_t, \quad \text{(exploitation)} \\ \widetilde{x} \sim \text{uniform distribution on } \mathcal{X}, & \text{with prob. } \varepsilon_t. \quad \text{(exploration)} \end{cases} \tag{3.5}$$

Here, $\widehat{Q}_{i,t}(s,x)$ is the estimated Q-matrix of informed AI speculator i in the t-th iteration, and $(v_t - p_t)x_{i,t}$ is the trading profit in iteration t if the order flow of informed AI speculator i is $x_{i,t}$. With probability $1-\varepsilon_t$, the Q-learning is in the exploitation mode with $x_{i,t}$ to be set as the maximizer of the estimated Q-matrix, $\widehat{Q}_{i,t}(s_t, x)$. On the other hand, with probability ε_t, the Q-learning is in the exploration model with $x_{i,t}$ to be randomly drawn from the uniform distribution on $\mathcal{X}$.[9] As t approaches infinity, the pre-specified exploration probability ε_t monotonically decreases to zero.

In equation (3.4), we see that during iteration t, the estimated Q-matrix for informed AI speculator i, denoted as $\widehat{Q}_{i,t}(s,x)$, undergoes an update exclusively at the state-action pair $(s_t, x_{i,t})$. The new value is updated to $\widehat{Q}_{i,t+1}(s_t, x_{i,t})$. However, 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 $x \neq x_{i,t}$. This updated value 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, $(v_t - p_t)x_{i,t} + \rho \max_{x\in\mathcal{X}} \widehat{Q}_{i,t}(s_{t+1}, x)$. A key distinction between the Q-learning recursive algorithm (3.4) and the Bellman recursive relation (3.1) lies in how they handle expectations. Q-learning algorithm (3.4) does not form expectations about the continuation value due to the unknown Markovian transition probability of s_{t+1}. Instead, it directly discounts the continuation value based on the randomly realized state s_{t+1} in the $t+1$ iteration.

It is crucial to note that the learning rate, denoted by the weight α, plays a significant role in the Q-learning algorithm, balancing past knowledge against present learning

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

based on a new experiment. A higher value of α not only indicates a greater impact of present learning on the Q-matrix value update but also implies a quicker forgetting of past knowledge, potentially leading to biased learning. This can be seen from the following expression:

$$\widehat{Q}_{i,t}(s_t, x_{i,t}) \approx \sum_{\tau=0}^{\infty} \alpha(1-\alpha)^{\tau} \underbrace{\left[(v_{t-\tau} - p_{t-\tau})x_{i,t-\tau} + \rho \max_{x \in \mathcal{X}} \widehat{Q}_{i,t-\tau}(s_{t+1-\tau}, x)\right]}_{\text{Learning based on the experiment in iteration } t-\tau}, \tag{3.6}$$

when t is large and ε_t has decayed almost to 0. Specifically, when α is not close to 0, the weights given by $\alpha(1-\alpha)^{\tau}$ decay so rapidly with τ that it jeopardizes the applicability of the LLN.

In the presence of randomness in the underlying environment, such as the noise traders' order flow u_t and asset value v_t in our model, a sufficiently small value of α is crucial for ensuring low bias in learning. However, a smaller value of α requires more iterations, and thus it incurs a greater computational cost. In contrast, for a relatively large α, it may cause the LLN to fail, thereby leading to biased learning. Moreover, if α is excessively small relative to the decay speed of the exploration rate ε_t, biased learning may arise from the insufficient exploration.

Experimentation. Conditional on the state variable s_t, informed trader i selects its order flow $x_{i,t}$ in two experimentation modes: exploitation and exploration. To determine the mode, we employ the simple ε-greedy method, which governs the decision-making process of the Q-learning algorithm. Specifically, as outlined in equation (3.5), informed trader i engages in the exploration mode with an exogenous probability ε_t during iteration t, whereas with a probability of $1-\varepsilon_t$, the trader operates in the exploitation mode. In the exploitation mode, informed trader i selects its order flow to maximize the current state's Q-value, given by $x_{i,t} = \text{argmax}_{x \in \mathcal{X}} \widehat{Q}_{i,t}(s_t, x)$. Conversely, in the exploration mode, informed trader i randomly chooses an order flow level $\tilde{x}$ from the set of all possible values in $\mathcal{X}$, each with equal probability. Essentially, the exploration mode guides the Q-learning algorithm to experiment with suboptimal actions based on the current Q-matrix approximation, $\widehat{Q}_{i,t}$.

Given that informed trader i lacks prior knowledge about its Q-matrix, it becomes evident that sufficient exploration is crucial to increase the accuracy of approximating the true Q-matrix, even when starting from an arbitrary initial value $\widehat{Q}_{i,0}$. 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. The random selection of actions by one informed trader introduces noise to the other traders, impeding their learning processes.

Exploitation, as a defining characteristic of reinforcement learning algorithms, plays a vital role in generating collusion among trading algorithms by biasing the estimation of the Q-matrix away from its true values. This bias leads to excessive overestimation of Q-values for certain choices that can sustain collusive profits, while simultaneously underestimating Q-values for other choices in $\mathcal{X}$. Termed as "collusion through biased learning," this phenomenon 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 traders 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.

Exploration is not only critical for approximating the true Q-matrix but also for informed traders to learn and sustain "collusion through punishment threat." In each iteration t, the randomly selected choice $\tilde{x}$ typically differs significantly from the exploited choice that generates collusion profits. Thus, such deviation, triggered by exploration, provides the only opportunity for the algorithms to learn strategies related to collusion through punishment threat.

3.2 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 value, asset price, order flow from the preferred-habitat investor, and total order flow: $\mathcal{D}_t \equiv \{(v_{t-\tau}, p_{t-\tau}, z_{t-\tau}, y_{t-\tau})\}_{\tau=1}^{T_m}$, where T_m is a large integer. The market maker estimates the demand curve of the preferred-habitat investor and the conditional expectation $\mathbb{E}[v_t|y_t]$ using the following linear regression models:

$$z_{t-\tau} = \xi_0 - \xi_1 p_{t-\tau}, \tag{3.7}$$

$$v_{t-\tau} = \gamma_0 + \gamma_1 y_{t-\tau} + \epsilon_{t-\tau}, \tag{3.8}$$

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 data set $\mathcal{D}_t$ in period t. The pricing rule adaptively adheres to the theoretical optimal policy using a plug-in procedure:

$$p_t(y) = \widehat{\gamma}_{0,t} + \frac{\theta \widehat{\gamma}_{1,t} + \widehat{\xi}_{1,t}}{\theta + \widehat{\xi}_{1,t}^2} y, \tag{3.9}$$

where θ is market maker's own choice. Therefore, the market maker is adaptive using simple statistical models.

3.3 Repeated Games of Machines

At $t = 0$, each informed trader $i \in \{1, \cdots, I\}$ 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 speculator i draws a random value that determines whether it will be in the exploration mode with probability ε_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 $x_{i,t}$.

(2) Noise traders, as a group, submit their order flow u_t, which is randomly drawn from a normal distribution $N(0, \sigma_u^2)$.

(3) Market makers observe the "historical data" $\mathcal{D}_t \equiv \{v_{t-\tau}, p_{t-\tau}, z_{t-\tau}, y_{t-\tau}\}_{\tau=1}^{T_m}$ and estimate the optimal pricing rule according to (3.7) – (3.9).

(4) Each informed AI speculator i realizes its profits $(v_t - p_t)x_{i,t}$ and updates its Q-matrix according to equation (3.4).

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

The interactions of informed AI speculators and an adaptive market maker, together with the randomness caused by noise traders and stochastic asset values in the background, make the stationary equilibrium difficult to achieve. The underlying economic environment we 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. 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 Design of Simulation Experiments

Theoretical analysis of the Q-learning programs 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.

Motivated by our theoretical framework, we focus on the experimental economic environment that consists of a group of $I \geq 2$ symmetric informed traders, a representative preferred-habitat investor, a market maker, and a representative noise trader.

4.1 Discretization of State and Action Space

Because Q-learning requires a finite 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 Φ^{-1} is the inverse cumulative density function of a standard normal distribution. The mathematical property of Φ^{-1} implies that grid points around the mean $\overline{v}$ are closer to each other than those far away from the mean. Because the probabilities of all n_v grid points of v_t are the same, the speed of convergence is significantly increased.[10]

[10] 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 different, 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 programs.

Following the guidelines offered by Calvano et al. (2020), we construct the discrete grid points for informed traders' order $x_{i,t}$ based on their optimal actions in the noncollusive Nash equilibrium and perfect cartel equilibrium. According to our model in Section 2, 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 traders' 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 firms can choose quantities beyond the theoretical levels corresponding to the noncollusive Nash equilibrium and perfect cartel equilibrium. As the action space is discrete, the exact quantities corresponding to the perfect cartel equilibrium may not be feasible. Despite this, our simulations show that firms 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 impact of the representative noise trader 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 2, market makers set the price according to the total order flow y_t, which is the sum of informed traders' 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., $\mathbb{P} = \{p_1, \cdots, p_{n_p}\}$.

4.2 Initial Q-Matrix and States

We adopt the initialization method of Calvano et al. (2020) by setting the initial Q-matrix at $t = 0$ using the discounted payoff that would accrue to informed trader i if the other informed traders randomize their actions uniformly over the grid points defined by $\mathbb{X}$.[11] Moreover, we consider zero trading orders from the representative noise trader,

[11] In reinforcement learning 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, resulting in early improvement in estimated action values. Thus, setting optimistic initial values are in some sense equivalent to adopting a thorough exploration over the entire action space early in the learning phase and then exploitation later on. Following this heuristic argument, we verify that adopting higher initial values for the Q-matrix has little effect on the quantitative results after informed traders' Q programs fully converge.

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

$$\widehat{Q}_{i,0}(p_m, v_k, x_n) = \frac{\sum_{x_{-i} \in \mathbb{X}} \left[v_k - (\overline{v} + \lambda(x_n + (I-1)x_{-i}))\right] x_n}{(1-\rho)n_x}, \tag{4.1}$$

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 randomly chosen. Specifically, the value of v_0 is drawn randomly from the discretized distribution of asset values, $\mathbb{V}$. The variable p_{-1} is randomized uniformly over the grids points of price, $\mathbb{P}$.

4.3 Specification of Learning Modes

Following Calvano et al. (2020), we adopt an exponentially time-declining state-dependent exploration rate for informed traders,

$$\varepsilon_{t(v_k)} = e^{-\beta t(v_k)}, \tag{4.2}$$

where the parameter $\beta > 0$ governs the speed that informed traders' exploration 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.[12] The specification of $t(v_k)$ implies that the exploration rate is state dependent, which ensures that informed traders can sufficiently explore their actions for all grid points of the exogenous state variable v_t.

The specification (4.2) implies that initially, Q-learning programs are almost always in the exploration mode, choosing actions randomly. However, as time passes, Q-learning programs gradually switch to the exploitation mode.

4.4 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 traders nor to the market maker.[13] They instead adopt Q-learning algorithms to learn how to make decisions in an

[12] In principle, we can allow informed traders 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 $s_t = \{p_{t-1}, v_t\}$.

[13] An exception is ρ and θ. The parameter ρ is known to informed traders as this parameter captures their own discount rates. The parameter θ is known to the representative market maker as this is their own choice.

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 σ_v plays a similar role as σ_u because what matters in our theoretical framework in Section 2 is the ratio σ_u/σ_v. We thus normalize the value of σ_v to unity. The parameter θ determines the extent to which the market maker focuses on price discovery. We find that the implications of different values of θ can be analyzed similarly by varying the value of ξ. Thus, for simplicity, we fix the value of θ 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 traders 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 ξ ranging from $\xi = 0$ to $\xi = 500$.

Simulation Parameters. Following Calvano et al. (2020), we set $\iota = 0.1$ so that informed traders can go beyond the theoretical bounds of actions 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 trader'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 traders will not be able to detect small deviations of peers even in the absence of noise, which in turn significantly lowers the possibility of algorithmic collusion through punishment threats.

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 of this paper are robust if we choose a larger n_v, n_x, n_p, or ι, as long as the hyperparameters, α and β, are adjusted accordingly to ensure sufficiently good learning outcomes. However, the cost of using denser grids is that significantly longer time would be required for Q-learning algorithms to fully converge to limit strategies.

We set $T_m = 10,000$ so that market makers store sufficiently long time-series data to estimate the linear regressions (3.7) and (3.8). 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 and the conditional expectation $\mathbb{E}\left[v_t|y_t\right]$. 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-programs by $N = 1,000$ times. All the random initial states and shocks (i.e., v_t, u_t, and exploration status of each informed trader 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 experiments in Sections 5 and 6.

Hyperparameters. The hyperparameters that control the learning process of Q-programs are set at $\alpha = 0.01$ and $\beta = 10^{-5}$. All results are robust to choosing different values of α and β so long as they are in the reasonable range that ensures sufficiently good learning outcomes. The implications of α and β for achieving collusive outcomes are discussed extensively by Calvano et al. (2020). Our baseline choice of β implies that any action $x_k \in \mathbb{X}$ is visited purely by random exploration by $1/[(1-\exp(-10^{-5}))n_x] = 6,666$ times on average before exploration completes.[14]

4.5 Convergence

Strategic games played by Q-learning algorithms do not have general convergence results. To verify convergence, Calvano et al. (2020) adopt a practical criterion by checking whether each player's optimal strategy does not change for 100,000 consecutive periods. 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

[14] 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 noise traders' random order and the pricing rule adopted by market makers.

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 programs playing repeated games.

In general, setting a smaller value of α or β requires longer time for the program to reach convergence. For example, with $\beta = 10^{-5}$, informed traders' Q-learning programs are still doing exploration with 36.8% probability after 100,000 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 programs to reach stationary optimal strategies is that exploration rate is virtually zero, say, after 1,000,000 periods. Moreover, with a small α, the Q-matrix is updated slowly when new information arrives. As a result, informed traders 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 more stringent criterion than the one used by Calvano et al. (2020) by requiring all informed traders' 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 on the realized values of random variables. Among all the experiments we study, the number of periods to reach convergence ranges from 2 million to 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 cluster that consists of 9 high-powered-computing servers, with 376 CPU cores in total. It takes about 1 to 30 mins to finish all N simulation sessions in one experiment, depending on the number of iterations needed to reach convergence.

4.6 Metrics Reflecting Collusive Behavior

Motivated by our theoretical framework in Section 2, we calculate three simple metrics that can be indicative of potential collusive behavior among informed traders. The values of all three metrics are computed in each simulation session over $T = 100,000$ periods, after informed traders' optimal strategies fully converge to the limit strategies according to the convergence criterion in Section 4.5.

Collusion Capacity. As in Calvano et al. (2020), the degree of collusion can be reflected by the Delta metric defined as follows:

$$\Delta^C \equiv \frac{\overline{\pi} - \overline{\pi}^N}{\overline{\pi}^M - \overline{\pi}^N}, \tag{4.3}$$

where $\overline{\pi} \equiv \sum_{t=T_c}^{T_c+T} \sum_{i=1}^{I} \pi_{i,t}(v_t, u_t)$ is the average profits of all informed traders over T periods after Q-learning programs reach convergence at T_c.[15] The values of $\overline{\pi}^N = \sum_{t=T_c}^{T_c+T} \sum_{i=1}^{I} \pi_i^N(v_t, u_t)$ and $\overline{\pi}^M = \sum_{t=T_c}^{T_c+T} \sum_{i=1}^{I} \pi_i^M(v_t, u_t)$ are the average profit that each informed trader would obtain, theoretically, in the noncollusive Nash equilibrium or perfect cartel equilibrium, respectively. Specifically, according to the formulas in Section 2.2, conditional on the realized values of v_t and u_t in period t, an informed trader's profit in the noncollusive Nash equilibrium is

$$\pi_i^N(v_t, u_t) = \left[v_t - p^N(Ix^N(v_t) + u_t)\right] x^N(v_t), \text{ for } i = 1, \cdots, I, \tag{4.4}$$

where $x^N(v_t) = \chi^N(v_t - \overline{v})$ and $p(Ix^N(v_t) + u_t) \equiv \overline{v} + \lambda^N(Ix^N(v_t) + u_t)$. Similarly, according to the formulas in Section 2.3, conditional on the realized values of v_t and u_t in period t, an informed trader's profit in the perfect cartel equilibrium is

$$\pi_i^M(v_t, u_t) = \left[v_t - p(Ix^M(v_t) + u_t)\right] x^M(v_t), \text{ for } i = 1, \cdots, I, \tag{4.5}$$

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

In principle, the value of Δ^C should range from 0 to 1. A larger Δ^C implies that informed traders attains more supra-competitive profits. The value of Δ^C can never be larger than 1 because $\overline{\pi}^M$ is the highest theoretically possible average profit. In fact, because informed traders 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 Δ^C below 0 under the limit strategies of informed traders. This outcome implies that informed traders 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.

[15] We average over $T = 100,000$ periods to smooth out the stochastic underlying economic environment, caused by the randomness in noise traders' order u_t and the stochastic variation of the asset value v_t over time. In fact, even if the underlying economic environment is stationary, as in the experiments of Calvano et al. (2020), Q-learning programs' optimal limit strategies may not be time invariant. Calvano et al. (2020) show that a large fraction of sessions displays cycles in AI agents' behavior even after convergence. We also find such cyclical patterns in our setting if we consider a setting without noise traders and with a constant asset value.

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 traders would obtain in the noncollusive Nash equilibrium theoretically. Specifically, the relative profit gain is $\overline{\pi}/\overline{\pi}^N$, where $\overline{\pi}$ and $\overline{\pi}^N$ are calculated similarly as those in equation (4.3).

Order Sensitivity to Asset Value. Our theoretical framework indicates that informed traders tend to be more conservative in placing their orders if there is implicit collusion. That is, the sensitivity of trading order $x_{i,t}$ to the asset value $v_t - \overline{v}$ is lower when informed traders collude more. Theoretically, informed traders' trading order $x_{i,t}$ is linear, as captured by $x_{i,t} = \chi^C(v_t - \overline{v})$. However, in our simulation experiments with Q-learning programs, such linearity restriction is not imposed at all. Despite this, we find that informed traders learn roughly linear strategies (see Figure 9). Therefore, we estimate $\widehat{\chi}^C$ based on the recorded asset value and order flow $\{v_t, x_{i,t}\}_{t=T_c}^{T_c+T}$ for each informed trader $i = 1, \cdots, I$, by running the following linear regression:

$$x_{i,t} = \chi_{i,0}^C + \chi_{i,1}^C v_t + \epsilon_t. \tag{4.6}$$

Consistent with our theoretical framework, we find that the estimates satisfy $\widehat{\chi}_{i,0}^C \approx -\overline{v}\widehat{\chi}_{i,1}^C$ in the unrestricted regression (4.6). The estimate $\widehat{\chi}_{i,1}^C$ captures the sensitivity of informed trader i's order $x_{i,t}$ to the asset's value v_t under the optimal trading strategies informed by their Q-learning programs. We further compute the average sensitivity of informed traders as $\widehat{\chi}^C = \frac{1}{I}\sum_{i=1}^{I} \widehat{\chi}_{i,1}^C$.

In our theoretical framework, it should be the case that $\chi^M \leq \chi^C \leq \chi^N$. Although no restriction is imposed on the Q-learning programs, we show in Section 6 that the estimated $\widehat{\chi}^C$ also satisfies $\chi^M \leq \widehat{\chi}^C \leq \chi^N$.

5 AI Collusion under Information Asymmetry

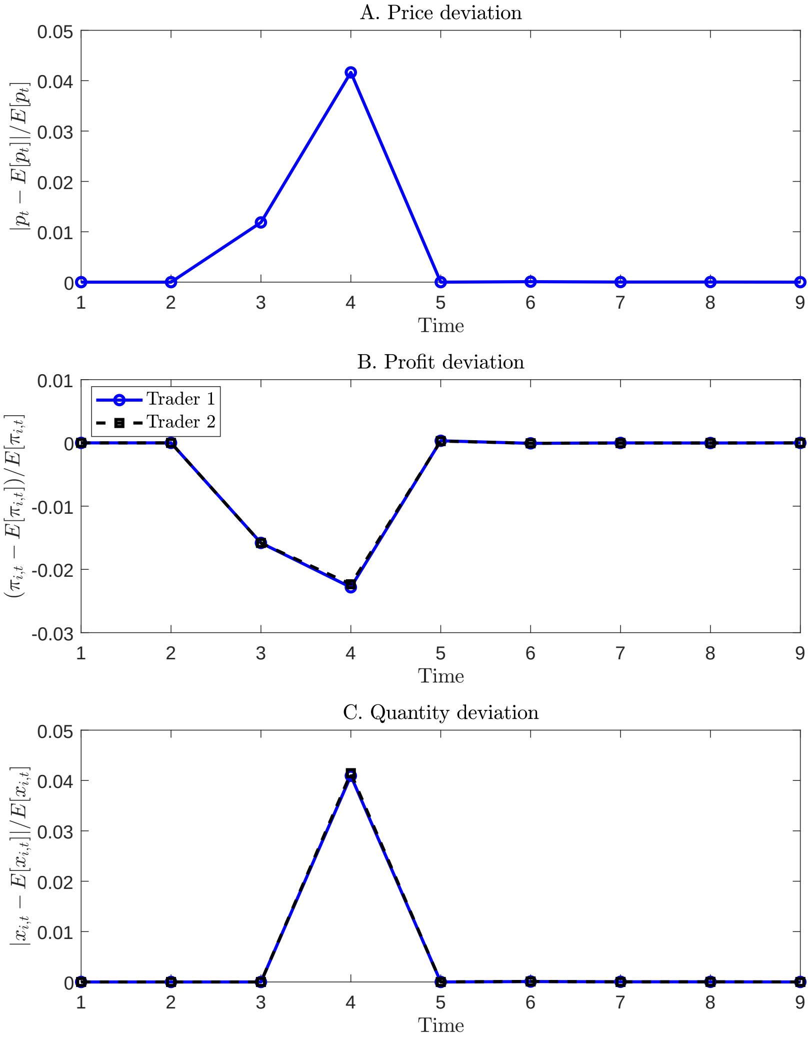
Our model suggests that under certain conditions, informed traders can achieve supra-competitive profits through implicit collusion when information asymmetry is small. In this section, we conduct simulation experiments with AI traders whose trading is powered by Q-learning programs. We are mainly interested in four questions. First, can AI traders learn to collude through the adoption of Q-learning programs, even if they do not communicate with each other or possess any information about the underlying economic environment? Second, if collusion exists, what are the mechanisms that generate such collusive behavior among AI traders? Third, how the pricing rule adopted by market makers affects the trading patterns of AI traders. Fourth, what are the implications of AI-powered trading for the price informativeness of financial markets?

In Subsection 5.1, we show that when information asymmetry is small, AI traders achieve supra-competitive profits through implicit collusion sustained by punishment threat. The price-trigger strategies learned by AI traders are quite similar to the those characterized by our model. In Subsection 5.2, we show that when information asymmetry is large, AI traders achieve supra-competitive profits due to their biased learning of the economic environment. In fact, because collusion is achieved through biased learning, rather than punishment threat, this result is also consistent with the model's prediction that implicit collusion cannot be sustained by price-trigger strategies in the presence of large information asymmetry. In Subsection 5.3, we study the role of information asymmetry and market makers in determining AI traders' profits and collusive behavior. Finally, in Subsection 5.4, we study the price informativeness of financial markets and show that perfect price informativeness is not achievable in the presence of AI traders.

5.1 Artificial Intelligence: Collusion through Punishment Threat

In this subsection, we study AI traders' behavior when information asymmetry is small, with $\sigma_u/\sigma_v = 0.1$. We focus on the baseline economic environment described in Section 4.4. The implications of alternative values of σ_u/σ_v are studied in Subsection 5.3. In the presence of small information asymmetry, we find that AI traders are able to achieve supra-competitive profits. Across $N = 1,000$ simulation sessions, the average value of Δ^C is about 0.73 and the average profit of AI traders is about 9% higher than the profit in the noncollusive equilibrium. Below, we examine the mechanism that leads to supra-competitive profits.

Price-Trigger Strategy. Motivated by our model, we examine whether the optimal strategies learned by AI traders are consistent with the price-trigger strategy illustrated in Section 2. To this end, we study the impulse response function (IRF) after an exogenous shock to the asset's price, which could be caused by the realization of random trading flows from noise traders. Specifically, in each simulation session, based on the economic environment that the session has converged to, we consider an exogenous shock to the asset's price p_t in period $t = 3$, which changes the value of p_t marginally by one grid point of price in $\mathbb{P}$. Though the exogenous change in the asset's price p_t in period $t = 3$ is caused by the trading flows from noise traders, to investigate whether price-trigger strategies are adopted, the direction of the price change is made to mimic the price impact of a profitable deviation from AI traders. That is, the exogenous change in p_t is positive if $v_t > \overline{v}$ and negative if $v_t < \overline{v}$. Both AI traders play their learned optimal strategies and the asset's price is determined endogenously by market makers according to their learned pricing rule in the subsequent periods, $t \geq 4$.



Note: In each simulation session, we consider an exogenous shock to the asset's price p_t in period $t = 3$, which changes the value of p_t marginally by one grid point of price in $\mathbb{P}$. The direction of the price change is made to mimic the price impact of a profitable deviation from AI traders. That is, the exogenous change in p_t is positive if $v_t > \overline{v}$ and negative if $v_t < \overline{v}$. Both AI traders play their learned optimal strategies and the asset's price is determined endogenously by market makers according to their learned pricing rule in the subsequent periods. Panel A plots the absolute percentage price deviation from the long-run mean, i.e., $|p_t - \mathbb{E}[p_t]|/\mathbb{E}[p_t]$. Panel B plots the two AI traders' per-period profit deviations from the long-run mean, i.e., $(\pi_{i,t} - \mathbb{E}[\pi_{i,t}])/\mathbb{E}[\pi_{i,t}]$ for $i = 1, 2$. Panel C plots the two AI traders' absolute percentage quantity deviations from the long-run mean, i.e., $|x_{i,t} - \mathbb{E}[x_{i,t}]|/\mathbb{E}[x_{i,t}]$ for $i = 1, 2$. All curves are average values across $N = 1,000$ 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 4.4.

Figure 1: IRF of an exogenous price change ($\sigma_u/\sigma_v = 10^{-1}$).

In Figure 1, panel A plots the evolution of the percentage deviation of the asset's price p_t from its long-run mean, i.e., $|p_t - \mathbb{E}[p_t]|/\mathbb{E}[p_t]$. Panel B plots the per-period profit deviations from the long-run mean for each AI trader, i.e., $(\pi_{i,t} - \mathbb{E}[\pi_{i,t}])/\mathbb{E}[\pi_{i,t}]$ for $i = 1, 2$. Panel C plots the evolution of the absolute percentage quantity deviations from the long-run mean for each AI trader, i.e., $|x_{i,t} - \mathbb{E}[x_{i,t}]|/\mathbb{E}[x_{i,t}]$ for $i = 1, 2$.[16]

In period $t = 3$, panel A shows that the asset's price p_t deviates from its long-run mean by 1.2% due to the exogenous shock. Panel B shows that this exogenous price change reduces both AI traders' profits by 1.6% of the long-run mean. Panel C shows that the trading quantities of both AI traders' remain at the long-run mean because informed traders submit their orders in period t before observing p_t.

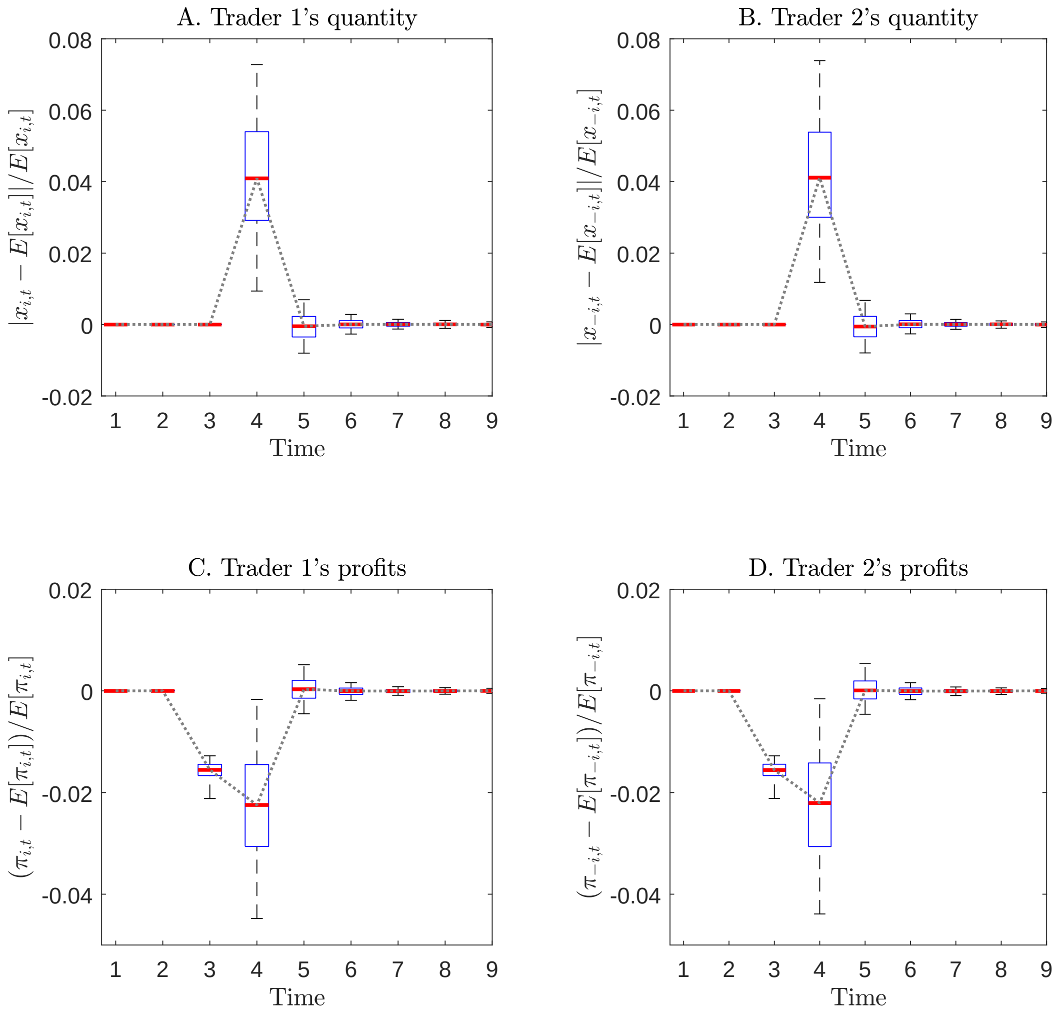
In period $t = 4$, panel C shows that in response to the exogenous price change in the previous period, both traders' orders significantly deviate from the long-run mean by 4.2%. The AI traders' aggressive behavior are similar to the price-trigger strategies described in Section 2. As a result of increased trading flows from AI traders, the percentage deviation of the asset's price continues to increase to 4.2% of the long-run mean (see panel A), which further enlarges both AI traders' profit losses to -2.4% of the long-run mean (see panel B).

In periods $t \geq 5$, panel C shows that both AI traders abruptly return to the predeviation level of quantities. As a result, both the price and profit deviation abruptly return to zero.

The patterns illustrated in Figure 1 are observed not because we take the average over $N = 1,000$ simulation sessions. In fact, we find that AI traders adopt similar price-trigger strategies in most simulation sessions. Figure 2 plots the distribution of the impulse responses. Although the magnitudes of quantity and price deviations differ significantly across sessions, the $[25\%, 75\%]$ and $[5\%, 95\%]$ confidence intervals indicate that price-trigger strategies are consistently adopted by AI traders.

Punishment for Deviation. According to our model in Section 2, price-trigger strategies are implemented based on whether the price in the preceding period deviates from the long-run mean, which could be caused by either the random orders submitted by noise traders or the orders submitted by informed traders. Informed traders cannot distinguish

[16] In our model and simulation experiments, informed (AI) traders may take long ($x_{i,t} > 0$) and short ($x_{i,t} < 0$) decisions depending on the sign of $v_t - \overline{v}$ (see Figure 9). As v_t is randomized independently across periods and sessions, the quantities of long and short positions and the prices determined by these positions will offset each other after taking the average. Thus, we focus on the average absolute percentage deviations from the long-run mean when plotting the IRF for order quantities and prices. Moreover, even after convergence, the economic environment is stochastic due to the random shocks to v_t and u_t. To clearly illustrate the IRF corresponding to the optimal trading strategies of AI traders, we smooth out these random shocks by taking the average across 10,000 independently simulated IRF for each of the $N = 1,000$ simulation sessions.



Note: The experiment is similar to that described for Figure 1. Panels A and B plot the two traders' quantity deviation from the long-run mean, and panels C and D plot their profit 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 $N = 1,000$ sessions. Parameters are set as in Figure 1.

Figure 2: Confidence intervals for the IRF of an exogenous price change ($\sigma_u/\sigma_v = 10^{-1}$).

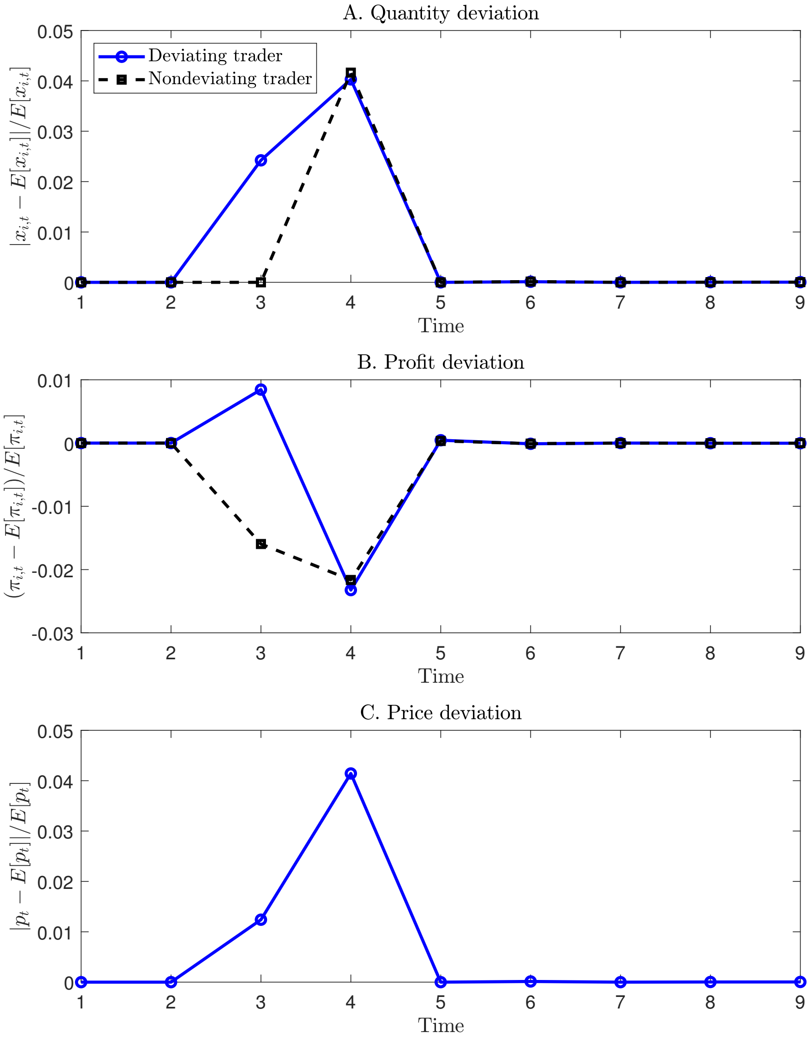
the causes of price deviation under information asymmetry.

Complementary to the impulse responses to an exogenous price change caused by the trading flows of noise traders (see Figure 1), we further study the impulse responses to a unilateral deviation by one of the AI traders. Specifically, in each simulation session, based on the economic environment that the session has converged to, we exogenously force one AI trader to have a one-time deviation from its learned limit strategy in period $t = 3$. The other AI trader does not detect this defect in period $t = 3$, and thus plays its learned optimal strategy in period $t = 3$. Starting from $t = 4$, both AI traders continue to play their learned optimal strategies in the subsequent periods. The one-time deviation in period $t = 3$ is made to the direction that increases the contemporaneous profits of the deviating trader (i.e., the deviating trader increases its order if $v_t > \overline{v}$ and reduces its order if $v_t < \overline{v}$). We consider a marginal deviation by one grid point of quantity in $\mathbb{X}$, which ensures that the resulting price deviation is similar to that in panel A of Figure 1 for comparison purposes.

Panel A of Figure 3 plots the evolution of the absolute percentage quantity deviations from the long-run mean for the deviating trader and nondeviating trader, respectively, i.e., $|x_{i,t} - \mathbb{E}[x_{i,t}]|/\mathbb{E}[x_{i,t}]$ and $|x_{-i,t} - \mathbb{E}[x_{-i,t}]|/\mathbb{E}[x_{-i,t}]$. In period $t = 3$, on average, the deviating trader's order deviates from the long-run mean by 2.5% while the nondeviating trader's order remains unchanged. In period $t = 4$, the deviation gets punished as the nondeviating trader behaves more aggressively, deviating its quantity from the long-run mean by 4.2%. Note that the behavior of the nondeviating trader in panel A of Figure 3 is almost identical to that in panel C of Figure 2 because the nondeivating trader only observes the asset's price in the previous period rather than its peer's trading order.

Rather than reducing the deviation amount, the deviating trader further increases its deviation amount to 4.1% of the long-run mean in period $t = 4$, slightly below that of the nondeviating trader. This form of overshooting exists for small deviations. As shown in panel A of Figure 5, if we consider a large deviation by three grid points of quantity, the deviating trader would reduce its quantity deviation in period $t = 4$. Regardless of whether its a small or a large deviation, both AI traders abruptly return to the predeviation level of quantities thereafter.

Panel B of Figure 5 plots the per-period profit deviations from the long-run mean for each AI trader, i.e., $(\pi_{i,t} - \mathbb{E}[\pi_{i,t}])/\mathbb{E}[\pi_{i,t}]$ and $(\pi_{-i,t} - \mathbb{E}[\pi_{-i,t}])/\mathbb{E}[\pi_{i,t}]$. In period $t = 3$, the forced deviation increases the deviating trader's profit by 0.8% of the long-run mean while reduces the nondeviating trader's profit by 1.6%. In period $t = 4$, due to the punishment strategy implemented by the nondeviating trader, the profit of the deviating trader drops substantially from 0.8% to -2.4% of the long-run mean. The expected



Note: In each simulation session, we exogenously force one AI trader to have a one-time deviation from its learned optimal strategy in period $t = 3$ while the other AI trader continues to play its learned optimal strategy in period $t = 3$. Starting from $t = 4$, both AI traders continue to play their learned optimal strategies in the subsequent periods. The one-time deviation in period $t = 3$ is made towards the direction that increases the contemporaneous profits of the deviating trader (i.e., the trader increases its quantity if $v_t > \overline{v}$ and reduces its quantity if $v_t < \overline{v}$). We consider a marginal deviation by one grid point of quantity in $\mathbb{X}$. Panel A plots the two AI traders' absolute percentage quantity deviations from the long-run mean, i.e., $|x_{i,t} - \mathbb{E}[x_{i,t}]|/\mathbb{E}[x_{i,t}]$ and $|x_{-i,t} - \mathbb{E}[x_{-i,t}]|/\mathbb{E}[x_{-i,t}]$. Panel B plots the per-period profit deviations from the long-run mean $(\pi_{i,t} - \mathbb{E}[\pi_{i,t}])/\mathbb{E}[\pi_{i,t}]$ and $(\pi_{-i,t} - \mathbb{E}[\pi_{-i,t}])/\mathbb{E}[\pi_{i,t}]$. Panel C plots the absolute percentage price deviation from the long-run mean, i.e., $|p_t - \mathbb{E}[p_t]|/\mathbb{E}[p_t]$. All curves are average values across $N = 1,000$ 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 4.4.

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

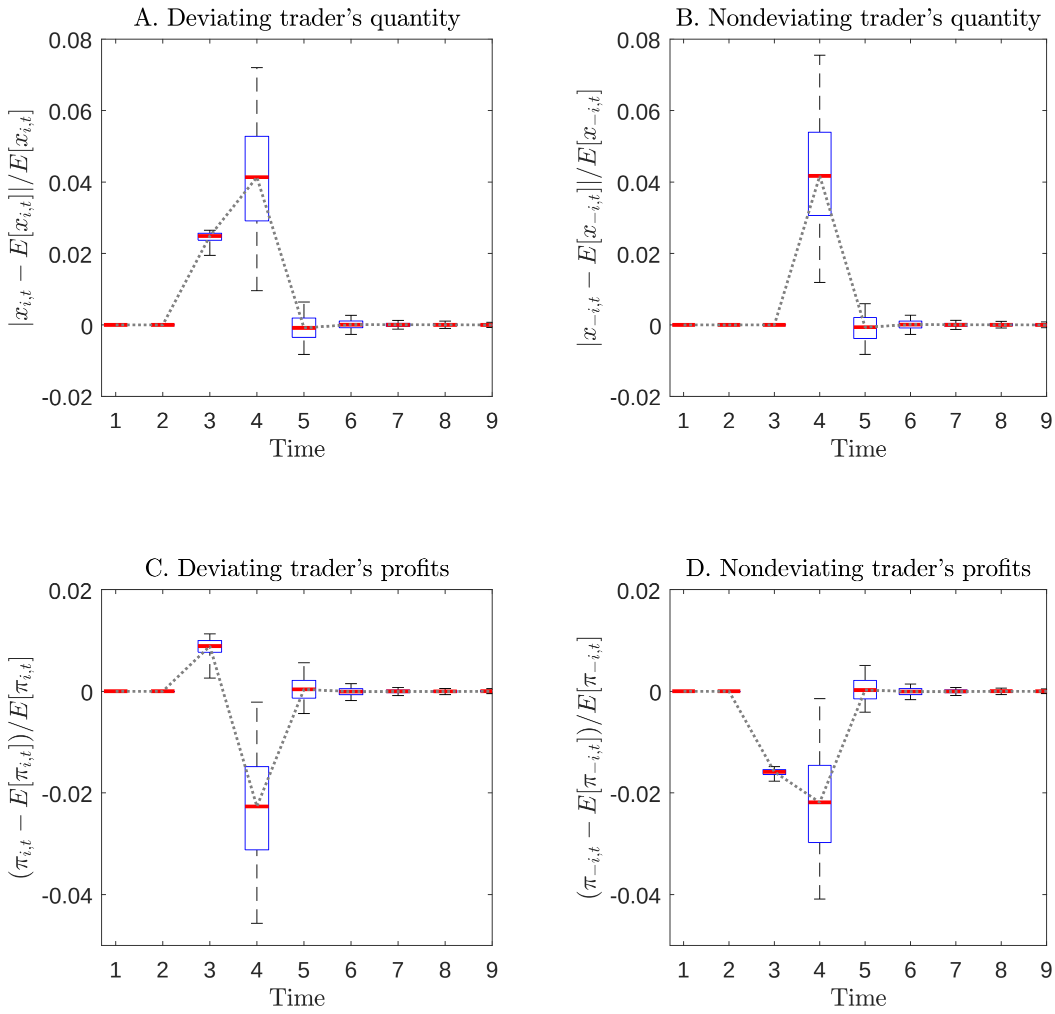
discounted profit of deviation is about -1.6% of the long-run mean for the deviating trader, indicating that the forced deviation is not a profitable strategy.

Panel C of Figure 3 plots the evolution of $|p_t - \mathbb{E}[p_t]|/\mathbb{E}[p_t]$, the percentage deviation of the asset's price from its long-run mean. In period $t = 3$, due to the forced deviation, the asset's price deviates from its long-run mean by 1.2%. In fact, this is the force that triggered both AI traders to change their decisions (i.e., order quantities) in period $t = 4$ because p_{t-1} is the only state variable that records the forced deviation in the last period $t = 3$. The asset's price continues to increase to 4.2% in period $t = 4$ because of the overshooting in the deviating trader's quantity, and then abruptly returns to the long-run mean in period $t = 5$ as the two AI traders revert to their predeviation behavior.

Figure 4 plots the distribution of the impulse responses and shows that the deviating trader 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 unilateral large deviation by three grid points of quantity in $\mathbb{X}$ in the experiment with $\sigma_u/\sigma_v = 10^{-1}$. The nondeviating trader still implements a punishment strategy by substantially increase its order in period $t = 4$ to punish the deviating trader's defect in period $t = 3$. The expected discounted profit of deviation is negative for the deviating trader. In panels D to F of Figure 5, we present the IRF of a unilateral marginal deviation by one grid point of quantity in $\mathbb{X}$ in the experiment with $\sigma_u/\sigma_v = 1$, in which the two AI traders 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 trader implements price-trigger strategies to deter deviations. However, the quantitative 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 Δ^C and the theoretical insight that collusive behavior becomes more difficult to achieve when informed traders are less able to monitor peers' deviations in the presence of larger information asymmetry.

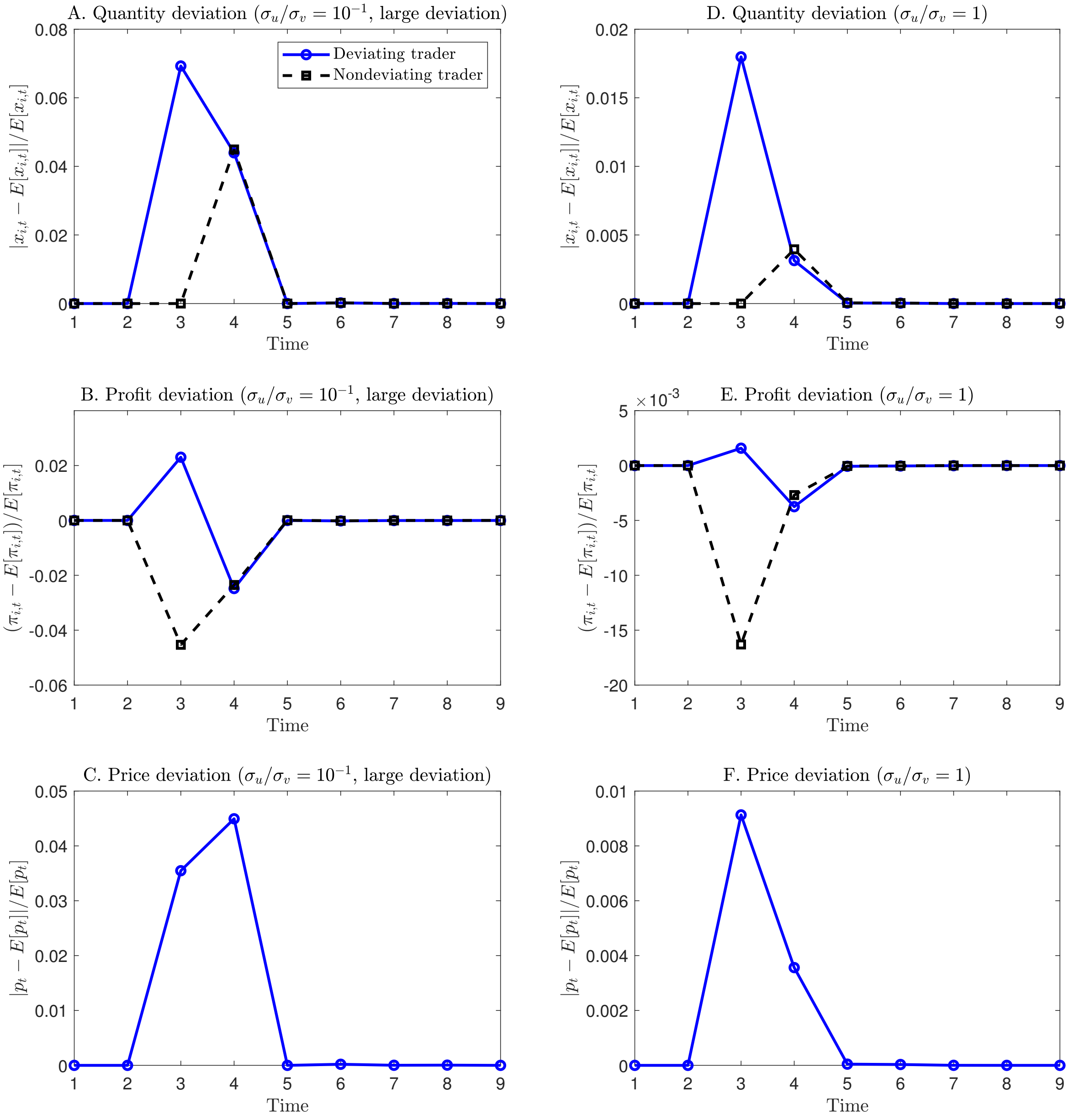
Further Discussions. Except for the duration of punishment, the impulse responses presented in Figures 1, 3 and 5 are quite consistent with the price-trigger strategies described in our model in Section 2. The patterns observed in our experiments coincide with our theoretical predictions that when information asymmetry is sufficiently small, informed traders are able to collude with each other by adopting price-trigger strategies to deter deviations. Moreover, collusion is more difficult to attain as information asymmetry becomes large.

Q-learning programs can learn price-trigger strategies because of experimentations. When one AI trader switches to the exploration mode in the process of learning, it would



Note: The experiment is similar to that described for Figure 3. Panels A and B plot the two traders' quantity deviation from the long-run mean, and panels C and D plot their profit 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 $N = 1,000$ sessions. Parameters are set as in Figure 3.

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



Note: The experiment is similar to that described for Figure 3. The left three panels consider a unilateral large deviation by three grid points of quantity in $\mathbb{X}$ in the experiment with $\sigma_u/\sigma_v = 10^{-1}$. The right three panels consider a unilateral marginal deviation by one grid point of quantity in $\mathbb{X}$ in the experiment with $\sigma_u/\sigma_v = 1$. The other parameters are set according to the baseline economic environment described in Section 4.4.

Figure 5: Robustness of IRF: large deviation or high information asymmetry ($\sigma_u/\sigma_v = 1$).

choose actions randomly. Such behavior is effectively similar to defect from an implicit collusive agreement, if any. When this occurs, the two AI traders would be trapped in the punishment phase until further explorations by one or both AI traders occur. AI traders are able to learn coordination strategies because exploration modes will eventually stop, a necessary condition for the simulation session to converge.

Our finding that AI traders are able to learn price-trigger strategies is similar to the finding of Calvano et al. (2020) that AI traders 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 $t = 4$, rather than gradually returning to predeviation behavior, the AI traders in our experiments abruptly return to their predeviation behavior. This difference is mainly due to the information asymmetry introduced by noise traders (i.e., $\sigma_u > 0$) and the stochastic asset's value (i.e., $\sigma_v > 0$). Both model ingredients make informed traders more difficult to achieve the collusion sustained by punishment threat, not just in the simulation experiments with AI traders, but also in the model in Section 2.

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 value of assets v_t in each period is drawn from an i.i.d. distribution. In this stochastic setting, it becomes more difficult for the two AI traders to learn punishment strategies to sustain collusion than in the deterministic setting with a constant v_t.[17] Second, noise traders' random actions generate information asymmetry to informed traders, which makes grim trigger strategies infeasible. As a result, informed traders have to adopt price-trigger strategies to collude. In both the model with rational-expectation informed traders and the simulation experiments with AI traders, the ratio σ_u/σ_v plays a crucial role in determining the level of collusion in financial markets.

The information asymmetry in our economic environment implies that peer AI traders' lagged actions are unobservable and thus cannot be included as state variables. Thus, as described in Section 3.1, 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 AI traders. 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 AI traders' can perfectly observe peers' actions, which is close to the perfect-information setting of

[17] In one of the robustness checks, Calvano et al. (2020) consider stochastic demand and show that the average Δ^C is lower when aggregate demand can take two values randomly. We also find that with stochastic v_t, the average Δ^C declines because it is more difficult for Q-learning programs to learn strong punishment strategies. The decline in Δ^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.

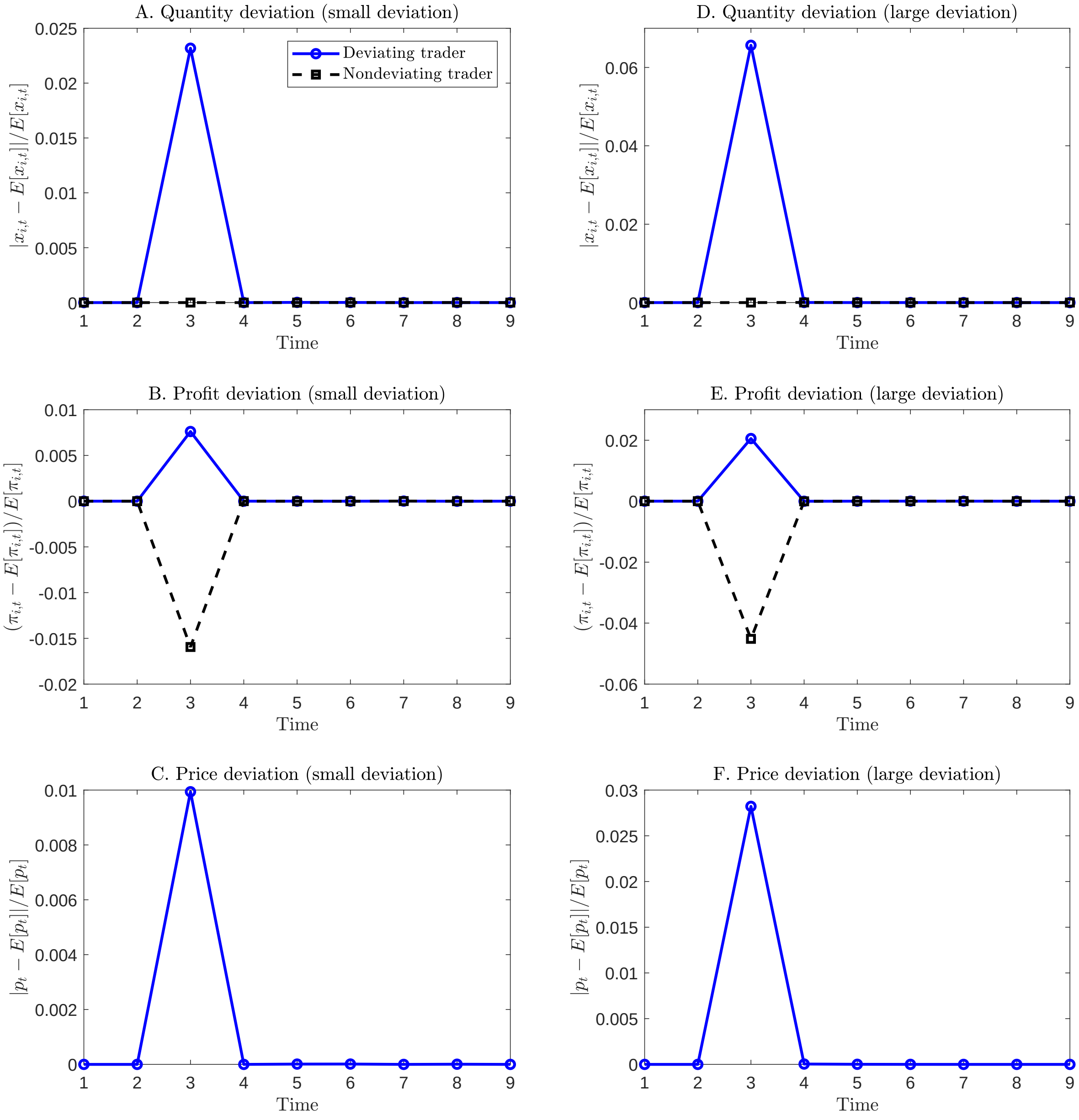
Calvano et al. (2020) except for including v_t as an additional state variable. Second, we consider the setting where state variables are $s_t = \{p_{t-1}, x_{i,t-1}, v_t\}$. We find that under the perfect information benchmark (i.e., $\sigma_u/\sigma_v = 0$) with two AI traders $I = 2$, these two alternative settings have almost the same average Δ^C. This is not surprising because under the perfect information benchmark, recording $x_{i,t-1}$ and p_{t-1} allows each AI trader to back out its peer's action $x_{-i,t-1}$. 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 Δ^C than the other setting with $s_t = \{p_{t-1}, x_{i,t-1}, v_t\}$. In addition, we find that the average Δ^C in these two alternative settings is higher than that in our baseline setting. Thus, incorporating AI traders' lagged actions as additional state variables indeed helps AI traders 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 traders and simulation experiments with AI traders, including lagged price p_{t-1} alone can already result in a significant degree of collusion.

5.2 Artificial Stupidity: Collusion through Biased Learning

In this subsection, we study AI traders' behavior when information asymmetry is large, with $\sigma_u/\sigma_v = 10^2$. Similar to Section 5.2, we focus on the baseline economic environment.

According to our model in Section 2, informed traders should find it impossible to collude with each other in this setting with large information asymmetry. However, in simulation our experiments, AI traders are able to achieve supra-competitive profits. Across $N = 1,000$ simulation sessions, the average value of Δ^C is about 0.6 and the average profit of AI traders is about 7.5% higher than the profit in the noncollusive equilibrium. The profits become even higher as information asymmetry increases. Below, we examine the mechanism that leads to such supra-competitive profits.

To begin with, we study the impulse responses to a unilateral deviation in Figure 6. Clearly, 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 trader. Instead, panels A and D of Figure 6 show that the nondeviating trader's order is virtually unchanged and the deviating trader 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 trader obtains an extra amount of one-period profit in period $t = 3$, which causes a one-period profit loss for the nondeviating trader. 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 trader.



Note: The experiment parameters are similar to those described for Figure 3, except for setting $\sigma_u/\sigma_v = 10^2$. The left three panels consider a unilateral marginal deviation by one grid point of quantity in $\mathbb{X}$. The right three panels consider a unilateral large deviation by three grid points of quantity in $\mathbb{X}$.

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

The collusive behavior of the two AI traders is clearly not sustained by price-trigger strategies when σ_u/σ_v is large, which is consistent with the prediction of our model (Proposition 2.4). We find that the seemingly collusive behavior under large information asymmetry is caused by AI traders' biased learning. Although deviation seems to be profitable in terms of increasing the discounted profits, both AI traders choose not to do this according to their learned optimal trading strategies. The reason is that AI traders' actions are governed by their learned Q-matrix, which suggests that the (no-deviation) strategies they are playing are already optimal and any deviations cannot be profitable. Such behavior constitutes a unique character of AI algorithms, which is intrinsically different from how human traders would behave.

We now explain how biased learning can lead AI traders to exhibit collusive behavior in three steps. First, in Subsection 5.2.1, we show that biased learning is significant when information asymmetry is large because in this case, the estimation of the Q-matrix cannot properly accounts for the distribution of noise trader's order u_t due to the failure of the law of large numbers. This is a generic issue of reinforcement learning algorithms. Second, in Subsection 5.2.2, we show that due to biased learning, actions with larger order amounts would be associated with larger unconditional variances of the estimated Q-values. Third, in Subsection 5.2.3, we show that these actions are less likely to be the optimal strategies adopted by AI traders after Q-learning programs converge. In other words, biased learning would more likely lead AI traders to optimally take actions with small order amounts, which coincide with those actions played in the collusive Nash equilibrium. Taken together, we argue that in the presence of large information asymmetry, collusive outcomes emerge due to AI traders' biased learning.

The magnitude of biased learning increases with the degree of information asymmetry (i.e., $\sigma_u/\sigma_v = 0$) in financial markets, along with other parameters. In Subsection 5.2.4, we further discuss the theoretical properties of biased learning, which provide unique predictions for us to test the relationship between biased learning and collusive outcomes in simulation experiments.

5.2.1 Biased Learning When Information Asymmetry is Large

First, we explain that when information asymmetry is large, there is biased learning for the Q-matrix due to the failure of the law of large numbers.

Biased learning is caused by a generic feature of reinforcement learning algorithms. As discussed in Section 3.1, Q-learning programs cannot take expectations due to the absence of knowledge about the underlying economic environment (e.g., the distribution of noise u_t). In each period t, the algorithm updates the value of one single cell (s, x)

(which includes state s and action x) of the Q-matrix according to the currently realized profit $(v_t - p_t)x_{i,t}$ (see equation (3.4)) rather than the expected profit $\mathbb{E}[(v-p)x|s,x]$ as in a rational-expectation 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, consider a simple setting in which s is the only state of the economy and x is the only action that trader i can play. Thus, the Q-matrix contains exactly one cell.[18]. The learning process updates the Q-matrix according to Equation (3.4). Thus, starting from the initial Q-matrix $\widehat{Q}_0(s,x)$, after T updates, the Q-matrix's value becomes

$$\begin{aligned}\widehat{Q}_{i,T}(s,x) =&\alpha \sum_{t=0}^{T-1} \delta^{T-1-t}(v_t - p_t)x + \delta^T \widehat{Q}_0(s,x) \\ =&\alpha \sum_{t=0}^{T-1} \delta^{T-1-t}\left[v_t - \overline{v} - \lambda(y_t - u_t)\right] x - \alpha\lambda x \sum_{t=0}^{T-1} \delta^t u_t + \delta^T \widehat{Q}_0(s,x). \qquad (5.1)\end{aligned}$$

where $\delta = 1 - \alpha + \alpha\rho$. The term $\alpha\lambda x \sum_{t=0}^{T-1} \delta^t u_t$ represents a stochastic term that depends on the noise order u_t, and it becomes relatively more important in determining $\widehat{Q}_{i,T}(s,x)$ when information asymmetry is large, i.e., σ_u/σ_v is large. With $\mathbb{E}[u_t] = 0$, the estimation for the limit value of $\widehat{Q}_{i,T}(s,x)$ is unbiased only if $\alpha\lambda x \sum_{t=0}^{T-1} \delta^t u_t = 0$ as $T \to \infty$[19], which occurs if $\delta \to 1$. Given $\rho \in (0,1)$, a necessary condition for $\delta \to 1$ is $\alpha \to 0$.[20] Thus, for any $\alpha > 0$, the term $\alpha\lambda x \sum_{t=0}^{T-1} \delta^t u_t$ would significantly bias the estimate of $\widehat{Q}_{i,T}(s,x)$ when σ_u/σ_v is sufficiently large. This is due to the failure of the law of large numbers because in general, as $T \to \infty$, we have $\alpha\lambda x \sum_{t=0}^{T-1} \delta^t u_t \neq \alpha\lambda x \mathbb{E}[u_t]$ unless $\delta \to 0$.

The magnitude of biased learning depends on the importance of the term $\alpha\lambda x \sum_{t=0}^{T-1} \delta^t u_t$ relative to the term $\alpha \sum_{t=0}^{T-1} \delta^{T-1-t}\left[v_t - \overline{v} - \lambda(y_t - u_t)\right] x$ in equation (5.1). Obviously, biased learning is absent when three is no information asymmetry (i.e., $\sigma_u/\sigma_v = 0$), and biased learning becomes more significant when σ_u/σ_v is larger. In general, the magnitude of biased learning also depends on the parameters α, λ, and ρ. These theoretical properties provide unique predictions for us to test the relationship between biased learning

[18] In the more general case with many values of s and x, the logic of our explanations still applies. However, equation (5.1) needs to be modified because the Q-learning programs do not necessarily visit and update the same cell (s,x) in every period.

[19] To see why unbiasedness requires $\alpha\lambda x \sum_{t=0}^{T-1} \delta^{T-1-t} u_t = 0$ 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 Section 3.1), which represents the discounted "expected" profits. In our model, the noise order u_t should have no direct effect on informed traders' "expected" profits except for affecting their trading order $x_{i,t}$.

[20] By setting $\rho = 1$, we also have $\delta = 1$. However, the choice of $\rho = 1$ is not feasible because the limit value of $\widehat{Q}_{i,T}(s,x)$ will explode. Moreover, unlike the hyperparameter α, the parameter ρ cannot be freely adjusted because it has an economic meaning and captures informed traders' patience.

and collusive outcomes in simulation experiments. We discuss them in Subsection 5.2.4.

5.2.2 Complementarity Between Informed Traders' Order and Noise Order

Second, we show that due to biased learning, actions with larger order amounts would be associated with larger unconditional variances of the estimated Q-values.

To begin with, we decompose the per-period profit $(v_t - p_t)x$ that an informed trader i receives when playing action $x \in \mathcal{X}$ in period t into two parts:

$$(v_t - p_t)x = \left[v_t - \overline{v} - \lambda(y_t - u_t)\right] x - \lambda x u_t. \tag{5.2}$$

The term $\left[v_t - \overline{v} - \lambda(y_t - u_t)\right] x$ captures the profit determined by the asset's fundamental value v_t and the term $\lambda x u_t$ captures the profit determined by the noise order u_t. Through the term $\lambda x u_t$ in equation (5.2), there exists complementarity between the informed trader's action x and the noise order u_t in determining per-period profits. This complementarity implies that, actions with larger order amounts (i.e., a larger absolute value $|x|$) would amplify the impact of the noise order u_t.

Because the estimated Q-value is the discounted value of per-period profits realized in the past, the complementarity between x and u_t in equation (5.2) would propagate to equation (5.1), captured by the term $\alpha\lambda x \sum_{t=0}^{T-1} \delta^t u_t$. In the absence of biased learning (i.e., when $\delta \to 1$ as $\alpha \to 0$), for a sufficiently large T, we should have $\alpha\lambda x \sum_{t=0}^{T-1} \delta^t u_t \approx \alpha\lambda x \mathbb{E}[u_t] = 0$, so that the unbiased estimate of the Q-value is not affected by the complementarity. However, as long as $\alpha > 0$, we would have $\alpha\lambda x \sum_{t=0}^{T-1} \delta^t u_t \neq 0$ for a sufficiently large T, 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 AI trader's particular action is path dependent, crucially depending on the realized noise order u_t when the AI trader plays this action in the past.

Thus, in the presence of biased learning, there exists complementarity between x and u_t in determining the estimated Q-value. This complementarity implies that the action that a larger order amount would be associated with a larger unconditional variance of its estimated Q-value, which consequently affects AI traders' optimal trading strategies in a way that makes the choice of large order amounts less likely.

5.2.3 Impacts of Biased Learning on Optimal Strategies

Third, we show that actions with large order amounts are less likely to be the optimal strategies adopted by AI traders after Q-learning programs converge. In other words, biased learning would more likely lead AI traders to optimally take actions with small order amounts, which coincide with those actions played in the collusive Nash equilibrium.

Before discussing why biased learning makes the choice of large order amounts less likely, it is useful to clarify although AI traders start their Q-learning programs 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 in time before Q-learning programs to converge. In other words, in a long period of time right before Q-learning programs converge, AI traders must be in pure exploitation mode, choosing the action that maximizes the Q-value rather than choosing the action 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 the AI trader can take two actions $0 < x_S < x_L$ in state s, with x_L being the action with a large order amount. As discussed above, in the presence of biased learning caused by information asymmetry, there is complementarity between x and u_t in determining the estimated Q-value. Thus, relative to the action x_S, the action x_L generates a large unconditional variance of the estimated Q-value (see equation (5.1)). Let $[\underline{Q}(x_S), \overline{Q}(x_S)]$ and $[\underline{Q}(x_L), \overline{Q}(x_L)]$ be the 99.9% confidence interval of the estimated Q-value for actions x_S and x_L, respectively. Thus, we have $[\underline{Q}(x_S), \overline{Q}(x_S)] \subset [\underline{Q}(x_L), \overline{Q}(x_L)]$. Because the AI trader is purely in the exploitation mode, at any time t, its action follows $\text{argmax}_{x_S, x_L} \left\{ \widehat{Q}_{i,t}(x_S), \widehat{Q}_{i,t}(x_L) \right\}$.

At any time t, there are two cases, either $\widehat{Q}_{i,t}(x_L) > \widehat{Q}_{i,t}(x_S)$ or $\widehat{Q}_{i,t}(x_L) <= \widehat{Q}_{i,t}(x_S)$. In the first case, for $\tau > [t, t']$, the AI trader would keep choosing x_L to update $\widehat{Q}_{i,\tau}(x_L)$ while $\widehat{Q}_{i,\tau}(x_S)$ remains unchanged at $\widehat{Q}_{i,t}(x_S)$. The time $t' > t$ is the first passage time for $\widehat{Q}_{i,t'}(x_L) <= \widehat{Q}_{i,t'}(x_S)$. From time t' on, the AI trader switches from playing x_L to playing x_S, and fall into the second case as described below.

In the second case, for $\tau > [t, t']$, the AI trader would keep choosing x_S to update $\widehat{Q}_{i,\tau}(x_S)$ while $\widehat{Q}_{i,\tau}(x_L)$ remains unchanged at $\widehat{Q}_{i,t}(x_L)$. The time $t' > t$ is the first passage time for $\widehat{Q}_{i,t'}(x_L) > \widehat{Q}_{i,t'}(x_S)$. From time t' on, the AI trader switches from playing x_S to playing 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 4.5 (i.e., stability of optimal strategy for $T = 100,000$ consecutive periods), eventually, the optimal strategy will converge to x_S with probability p and x_L with probability $1 - p$. We have $p > 0.5$ because $\underline{Q}(x_L) < \underline{Q}(x_S)$. The probability p is higher if the action x_L's estimated Q-value has a larger probability to be in the interval $[\underline{Q}(x_L), \underline{Q}(x_S)]$, which happens when information asymmetry is larger (i.e., larger σ_u / σ_v so there is more significant biased learning) or the difference in order amounts is larger (i.e., larger $x_L - x_S$). This explains why biased learning makes the choice of large order amounts less likely.

According to our model in Section 2, the sensitivity of informed traders' order flow to the asset's value v_t is lower under collusion, i.e., $\chi^M \leq \chi^C \leq \chi^N$. Because informed trader i's order $x_{i,t}$ is $x_{i,t} = \chi(v_t - \overline{v})$, its absolute order amount satisfies $|x_{i,t}^M| \leq |x_{i,t}^C| \leq |x_{i,t}^N|$ for any v_t, indicating that informed traders would collude if they adopt more conservative (i.e., trading smaller order $|x_{i,t}|$), rather than more aggressive, trading strategies. Taken together, it is clear that in the presence of large information asymmetry, biased learning leads to collusive outcomes.

5.2.4 Testable Predictions of the Biased Learning Mechanism

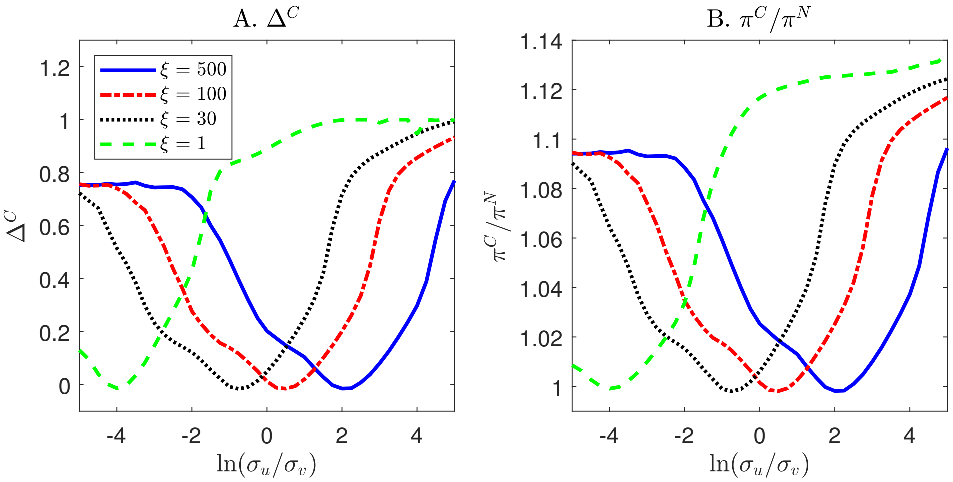
Per our discussions above, more collusive outcomes are generated as a result of biased learning. As discussed in the end of Subsection 5.2.1, the magnitude of biased learning depends on the relative importance of the term $\alpha\lambda x \sum_{t=0}^{T-1}(1-\alpha+\alpha\rho)^t u_t$ in equation (5.1). Obviously, this term becomes more important in determining the estimated Q-value when σ_u/σ_v is larger, λ is larger, ρ is smaller, or α is larger. These theoretical properties predict that AI traders can attain higher supra-competitive profits due to a larger degree of biased learning when σ_u/σ_v is larger, λ is larger, ρ is smaller, or α is larger.

The above unique predictions allow us to further test and understand the impacts of biased learning on AI traders' collusive outcomes in simulation experiments. We briefly summarize the findings below. Consistent these predictions, first, we show that the average Δ^C across $N = 1,000$ simulation sessions increases with $\log(\sigma_u/\sigma_v)$ in the region with large information asymmetry in panel A of Figure 7. Second, we show that conditional on large information asymmetry (e.g., $\log(\sigma_u/\sigma_v) = 2$), reducing ξ from 500 to 1 (which results in a larger λ) leads to a larger average Δ^C in panel A of Figure 7. Third, we show that with large information asymmetry, reducing the value of ρ will lead to a larger average Δ^C in panel C of Figure 12. Finally, we show that with large information asymmetry, a higher α would result in a lower average Δ^C in panel B of Figure 14.

5.3 Role of Information Asymmetry and Market Efficiency

In this subsection, we study the role of information asymmetry and market efficiency.

Role of Information Asymmetry. Consider the baseline economic environment described in Section 4.4. The blue solid line in Panel A of Figure 7 plots the average Δ^C across $N = 1,000$ simulation sessions as $\log(\sigma_u/\sigma_v)$ varies from -5 to 5 along the x-axis, holding all other parameters unchanged. It shows that as $\log(\sigma_u/\sigma_v)$ increases along the x-axis, the average Δ^C first decreases and then increases. This U-shape pattern is an



Note: This figure plots the average Δ^C and profit gain relative to noncollusion (π^C/π^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 4.4.

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

outcome of the interaction of the two mechanisms discussed in Subsections 5.1 and 5.2. Panel B of Figure 7 plots the profit gain relative to noncollusion (π^C/π^N), the pattern is similar to that in panel A.

Specifically, in the region of small information asymmetry, i.e., $\log(\sigma_u/\sigma_v) < 2$, the average Δ^C is decreasing in $\log(\sigma_u/\sigma_v)$. In this region, AI traders adopt price-trigger strategies to attain supra-competitive profits, as discussed in Section 5.1. The negative relationship between the average Δ^C and $\log(\sigma_u/\sigma_v)$ observed in our simulation experiments is consistent with the prediction of our model (see Proposition 2.6.(ii)).

In the region of large information asymmetry, i.e., $\log(\sigma_u/\sigma_v) \geq 2$, the average Δ^C is increasing in $\log(\sigma_u/\sigma_v)$. In this region, AI traders attain supra-competitive profits because of biased learning, as discussed in Section 5.2. The positive relationship between the average Δ^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 Subsection 5.2.4).

Role of Market Efficiency. According to our model in Section 2, the market maker focuses more on minimizing pricing errors when ξ is small or θ is large. In this case, market is efficient and there is no collusive Nash equilibrium that can be sustained by price-trigger strategies for any $\sigma_u/\sigma_v > 0$ (Proposition 2.3). By contrast, when ξ is large or θ is small, the market maker focuses more on minimizing inventory costs. In this case, market is inefficient and there exists a collusive Nash equilibrium that can be sustained by price-trigger strategies for small σ_u/σ_v and I (Proposition 2.4).

By varying the value of ξ in our simulation experiments, we study how market efficiency affects AI traders' trading profits. We do not conduct experiments with different θ because a smaller θ has similar impacts as a larger ξ on market efficiency.

Specifically, the four curves in panel A of Figure 7 represent the experiments with $\xi = 500, 100, 30$ and 1, representing different weights in the market maker's pricing objective in terms of minimizing pricing errors or inventory costs. The overall U-shaped relationship between the average Δ^C and $\log(\sigma_u/\sigma_v)$ is not peculiar to the choice of ξ. All four curves display U-shape patterns.

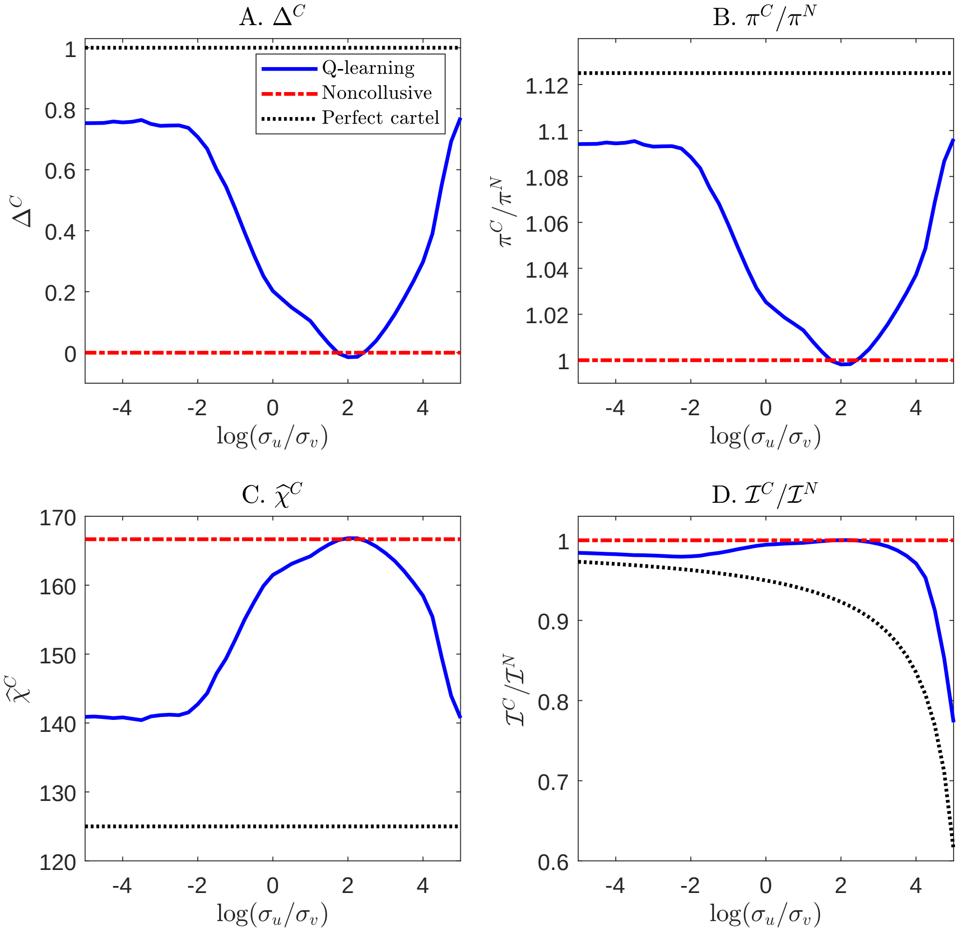
As we compare the four curves in panel A of Figure 7, one salient feature is that the trough of the U-shape shifts to the left as ξ decreases. This suggests that with a smaller ξ, a lower level of information asymmetry is needed for AI traders to adopt price-trigger strategies. A similar point can be made if we focus on the region with small information asymmetry, in which price-trigger strategies are adopted for AI traders. For example, holding $\ln(\sigma_u/\sigma_v) = -4$ unchanged, it is clear that the average Δ^C declines monotonically as ξ decreases from 500 to 1. Thus, collusion becomes more difficult to achieve through price-trigger strategies as ξ decreases, as predicted by our model (see Proposition 2.6.(iv)).

By contrast, let us switch the focus to the region with large information asymmetry, in which AI traders' trading strategies are dominantly affected by biased learning. For example, holding $\ln(\sigma_u/\sigma_v) = 2$ unchanged, it is clear that the average Δ^C increases monotonically as ξ decreases from 500 to 1. This is consistent with the theoretical property of biased learning discussed in Subsection 5.2.4, that is, the magnitude of biased learning increases with λ (i.e., decreases with ξ). Thus, a lower ξ leads to more biased learning, allowing AI traders to achieve higher supra-competitive profits.

5.4 Price Informativeness

In this subsection, we study the implication of information asymmetry for price informativeness in a financial market with AI traders.

Consider the baseline economic environment described in Section 4.4. The blue solid line in panel A of Figure 8 is similar to that in panel A of Figure 7, which plots the average Δ^C across $N = 1,000$ simulation sessions as $\log(\sigma_u/\sigma_v)$ varies. 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. Panel B of Figure 8 plots the average π^C/π^N, the profit gain relative to noncollusion, across $N = 1,000$ simulation sessions. The blue solid line in panel B exhibits a similar U-shape pattern as panel A. When $\log(\sigma_u/\sigma_v)$ is very small or very large, AI traders can increase their profits by about 9.5% relative to what they would obtain in the noncollusive Nash equilibrium

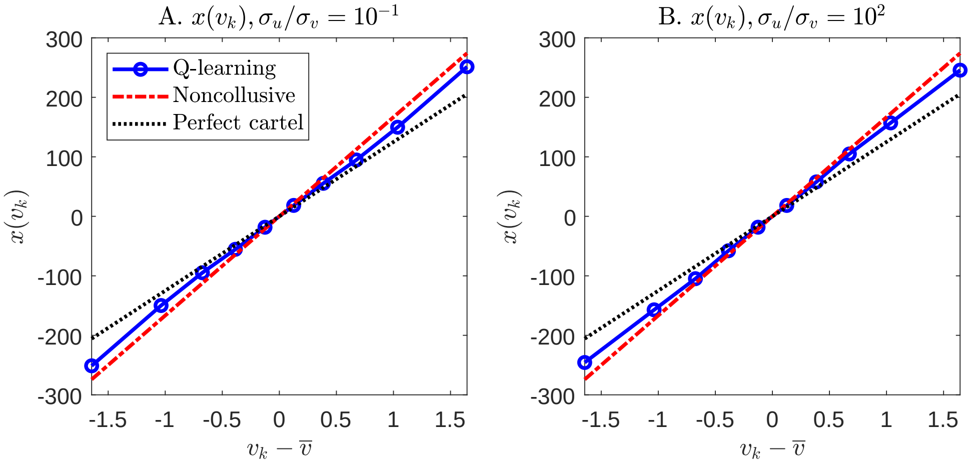


Note: This figure plots the average values of different metrics across $N = 1,000$ simulation sessions as $\log(\sigma_u/\sigma_v)$ varies. Panels A, B, and C plot the average Δ^C, profit gain relative to noncollusion (π^C/π^N), and informed traders' order sensitivity to asset value $(\widehat{\chi}^C)$. These metrics are defined in Section 4.6. Panel D plots the price informativeness relative to the theoretical benchmark of the noncollusive Nash equilibrium, i.e., $\mathcal{I}^C/\mathcal{I}^N$, where the price informativeness in the simulation experiment is calculated by $\mathcal{I}^C = \log\left[(I\widehat{\chi}^C)^2(\widehat{\sigma}_v/\sigma_u)^2\right]$. The blue solid line represents the simulation experiments with AI traders; the red dash-dotted and black dotted lines represent the theoretical benchmarks in the noncollusive Nash equilibrium and perfect cartel equilibrium, respectively. The other parameters are set according to the baseline economic environment described in Section 4.4.

Figure 8: Price informativeness for $\log(\sigma_u/\sigma_v) \in [-5, 5]$.

theoretically.

Trading Strategy of AI Traders. In panel C of Figure 8, we plot the sensitivity of AI traders' order to the asset's value, $\widehat{\chi}^C$ estimated based on equation (4.6). Consistent with panel A of Figure 8, $\widehat{\chi}^C$ displays an inverted U-shape as $\log(\sigma_u/\sigma_v)$ increases along the x-axis. By contrast, the theoretical benchmarks χ^N and χ^M stay roughly unchanged as $\log(\sigma_u/\sigma_v)$ increases.



Note: The blue solid line plots the average trading strategy of AI traders across $N = 1,000$ simulation sessions. Panels A and B represent the experiments with small ($\sigma_u/\sigma_v = 10^{-1}$) and large ($\sigma_u/\sigma_v = 10^{2}$) information asymmetry. 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 trading strategy of I AI traders across all grid points of $\mathbb{P}$, after Q-learning programs converge. The dots on the blue solid lines represent the discrete grid points of $\mathbb{V}$. The other parameters are set according to the baseline economic environment described in Section 4.4.

Figure 9: The trading strategy of AI traders.

In fact, the estimated $\widehat{\chi}^C$ almost sufficiently describes AI traders' 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. To show this, in Figure 9, we present the average trading strategy of AI traders across $N = 1,000$ simulation sessions. Panel A is for the experiment with small information asymmetry ($\sigma_u/\sigma_v = 10^{-1}$) and panel B is for the experiment with large information asymmetry ($\sigma_u/\sigma_v = 10^{2}$). 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 trading strategy of I informed traders across all grid points of $\mathbb{P}$, after Q-learning programs converge. The dots on the blue solid lines represent 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 AI traders learn a trading strategy that is roughly linear in the asset's value, even though the linearity restriction is never imposed on the Q-learning programs. Moreover, the slope of a linear fit for the trading strategy of AI traders lies between χ^M and χ^C in both panels A and B of Figure 9. Thus, the trading strategy learned by AI traders is more conservatively than that in the noncollusive Nash equilibrium, which explains why AI traders are able to attain supra-competitive profits.

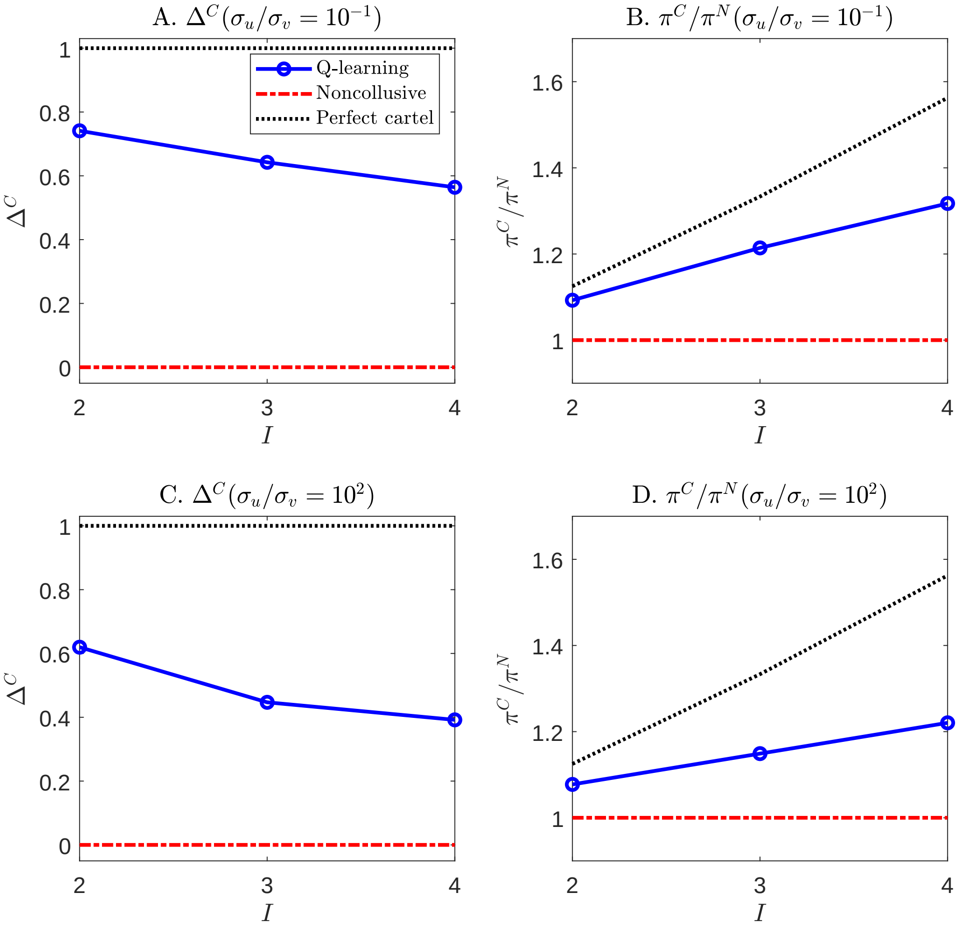
Price Informativeness in Markets with AI Trading. In panel D of Figure 8, we present the relative price informativeness. Specifically, price informativeness is measured by the natural log of the signal-noise ratio of price, which are $\mathcal{I}^N = \log\left[(I\chi^N)^2(\widehat{\sigma}_v/\sigma_u)^2\right]$ and $\mathcal{I}^M = \log\left[(I\chi^M)^2(\widehat{\sigma}_v/\sigma_u)^2\right]$ in the theoretical benchmarks of the noncollusive Nash equilibrium and perfect cartel equilibrium, respectively. The price informativeness in our numerical experiment with AI-powered trading is $\mathcal{I}^C = \log\left[(I\widehat{\chi}^C)^2(\widehat{\sigma}_v/\sigma_u)^2\right]$, where $\widehat{\chi}^C$ is estimated based on equation (4.6) and $\widehat{\sigma}_v$ is the standard deviation of v_t under our discrete grid points in $\mathbb{V}$. The relative price informativeness is measured by $\mathcal{I}^C/\mathcal{I}^N$ (the blue solid line), $\mathcal{I}^N/\mathcal{I}^N \equiv 1$ (the red dash-dotted line), and $\mathcal{I}^M/\mathcal{I}^N$ (the black dotted line).

Consistent with panel C of Figure 8, the blue solid line displays an inverted U-shape. The relative price informativeness is close to one when $\log(\sigma_u/\sigma_v)$ is around 2, which is also the region when $\widehat{\chi}^C \approx \chi^N$. When $\log(\sigma_u/\sigma_v)$ is very small or very large, the price informativeness in our numerical experiments with AI traders is significantly lower than that in the theoretical benchmark of the noncollusive Nash equilibrium. The reason is that AI traders place orders in a more conservative manner, with $\widehat{\chi}^C < \chi^N$, as shown in panel C of Figure 8.

Our findings suggest that perfect price informativeness is not achievable in the presence of AI traders. In our simulation environments, when information asymmetry declines (i.e., σ_u/σ_v decreases), AI traders would withhold their information and collude more through price-trigger strategies, placing orders more conservatively than what they would do in the noncollusive Nash equilibrium (see panel C of Figure 8). This reduces price informativeness. Crucially, AI traders never need to communicate with each other, whether explicitly or implicitly, the adoption of Q-learning programs automatically leads to such collusive behavior.

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 AI traders affects their trading strategies. In Subsection 6.2, we study the implication of AI traders' discount rates. Finally, in Subsection 6.3, we study the impacts of hyperparameters α and β on AI traders' learning outcomes.



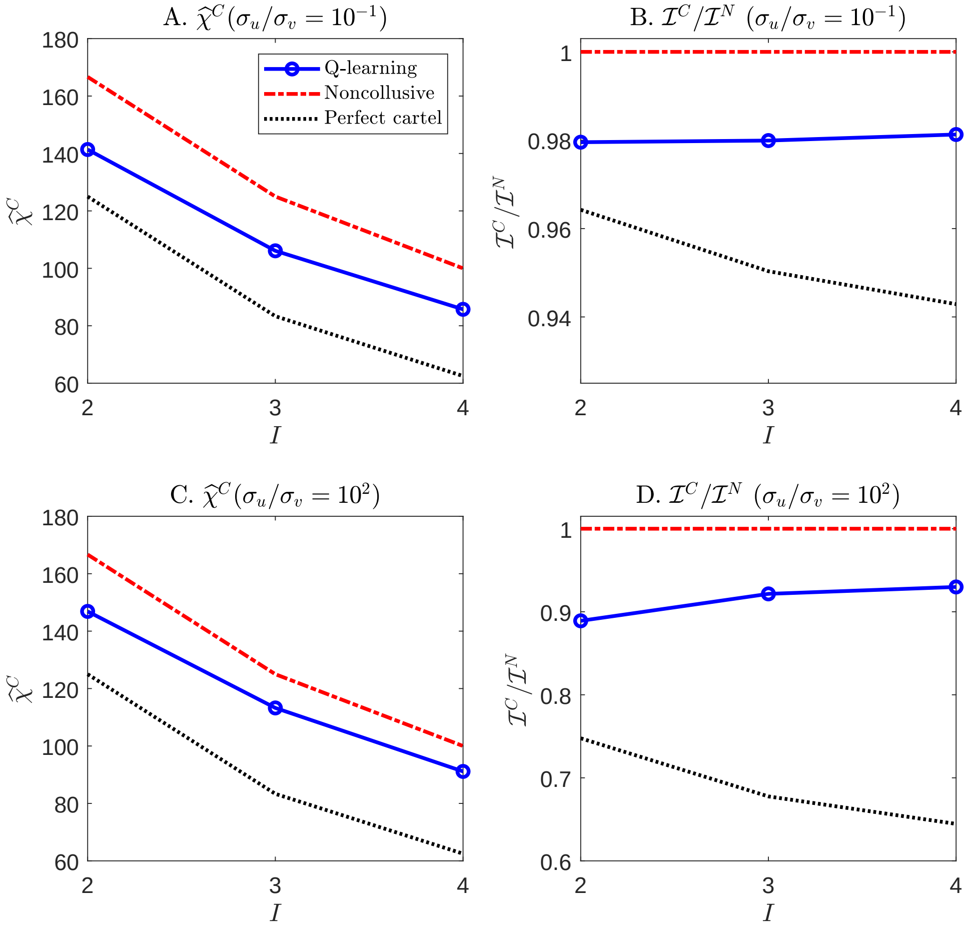
Note: This figure plots the average values of different metrics across $N = 1,000$ simulation sessions as the number of AI traders I varies. Panels A and B plot the average Δ^C and profit gain relative to noncollusion (π^C/π^N) under small information asymmetry ($\sigma_u/\sigma_v = 10^{-1}$). Panels C and D plot these metrics under large information asymmetry ($\sigma_u/\sigma_v = 10^2$). The blue solid line represents the simulation experiments with AI traders; the red dash-dotted and black dotted lines represent the theoretical benchmarks in the noncollusive Nash equilibrium and perfect cartel equilibrium, respectively. The other parameters are set according to the baseline economic environment described in Section 4.4.

Figure 10: Implications of the number of AI traders on Δ^C and π^C/π^N.

6.1 Number of AI Traders

Our model in Section 2 predicts that when ξ is sufficiently large (or θ is sufficiently small) and information asymmetry is small (i.e., small σ_u/σ_v), informed traders are less able to collude through price-trigger strategies when there are more informed traders. That is, the average Δ^C decreases with I and price informativeness $\mathcal{I}^C$ increases with I (see Proposition 2.6.(i)).

In the simulation experiments with AI traders, we find similar patterns. Specifically,



Note: This figure plots the average values of different metrics across $N = 1,000$ simulation sessions as the number of AI traders I varies. Panels A and B plot informed traders' order sensitivity to asset value ($\widehat{\chi}^C$) and the relative price informativeness ($\mathcal{I}^C/\mathcal{I}^N$) under small information asymmetry ($\sigma_u/\sigma_v = 10^{-1}$). Panels C and D plot these metrics under large information asymmetry ($\sigma_u/\sigma_v = 10^2$). The blue solid line represents the simulation experiments with AI traders; the red dash-dotted and black dotted lines represent the theoretical benchmarks in the noncollusive Nash equilibrium and perfect cartel equilibrium, respectively. The other parameters are set according to the baseline economic environment described in Section 4.4.

Figure 11: Implications of the number of AI traders on $\widehat{\chi}^C$ and $\mathcal{I}^C/\mathcal{I}^N$.

consider the baseline economic environment described in Section 4.4. In panels A and B of Figures 10 and 11, we conduct simulation experiments under small information asymmetry ($\sigma_u/\sigma_v = 10^{-1}$). Consistent with the model prediction, panels A and B of Figure 10 show that as the number of AI traders I increases from 2 to 4, the average Δ^C decreases from 0.74 to 0.56 and the average π^C/π^N increases from 1.09% to 1.32%, respectively. Moreover, panels A and B of Figure 11 show that as I increases, $\widehat{\chi}^C$ declines due to a congestion effect and the relative price informativeness $\mathcal{I}^C/\mathcal{I}^N$ increases. Because in our model, $\mathcal{I}^N$ increases with I, the positive relationship between $\mathcal{I}^C/\mathcal{I}^N$ and I implies that $\mathcal{I}^C$ is increasing in I, which is consistent with the model's prediction.

For comparisons, in panels C and D of Figures 10 and 11, we conduct simulation experiments under large information asymmetry ($\sigma_u/\sigma_v = 10^2$). In this case, AI traders achieve supra-competitive profits due to biased learning, as discussed in Subsection 5.2. We find that the implications of I for AI traders' strategies are similar to the experiments with small information asymmetry. Specifically, in panels C and D of Figure 10, the blue solid lines show that as I increases from 2 to 4, the average Δ^C decreases from 0.62 to 0.39 and the average π^C/π^N increases from 1.08% to 1.22%, respectively. Moreover, the blue solid lines in panels C and D of Figure 11 show that as I increases, $\widehat{\chi}^C$ decreases and the relative price informativeness $\mathcal{I}^C/\mathcal{I}^N$ increases. These results seem to suggest that the coordination through biased learning becomes more difficult to achieve when there are more AI traders in the market.

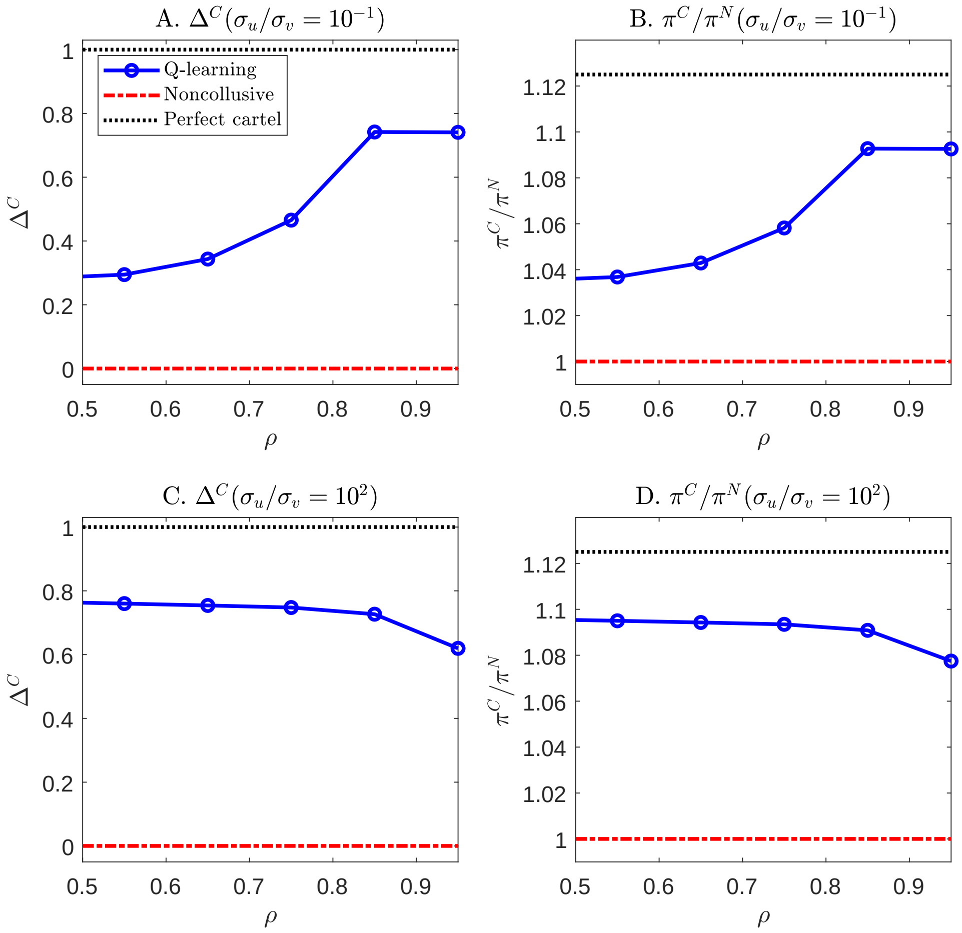
6.2 Discount Rates

Our model in Section 2 predicts that when ξ is sufficiently large (or θ is sufficiently small) and information asymmetry is small (i.e., small σ_u/σ_v), informed traders are more able to collude through price-trigger strategies as the discount rate ρ increases. That is, the average Δ^C increases with ρ and price informativeness $\mathcal{I}^C$ decreases with ρ (see Proposition 2.6.(iii)).

In the simulation experiments with AI traders, we find similar patterns. Specifically, consider the baseline economic environment described in Section 4.4. In panels A and B of Figures 12 and 13, we conduct simulation experiments under small information asymmetry ($\sigma_u/\sigma_v = 10^{-1}$). Consistent with the model prediction, panels A and B of Figure 12 show that as ρ increases from 0.5 to 0.95, the average Δ^C decreases from 0.29 to 0.74 and the average π^C/π^N increases from 1.04% to 1.09%, respectively. Moreover, panels A and B of Figure 13 show that as ρ increases, both $\widehat{\chi}^C$ and relative price informativeness $\mathcal{I}^C/\mathcal{I}^N$ decrease, which is consistent with the model's prediction.

Turning to the economic environment with large information asymmetry, the theoretical properties discussed in Subsection 5.2.4 imply that as the discount rate ρ increases, the magnitude of biased learning declines, and as a result, the supra-competitive profits that AI traders are able to achieve would decrease. The patterns observed in our simulation experiments are consistent with these predictions.

In particular, in panels C and D of Figures 12 and 13, we conduct simulation experiments under large information asymmetry ($\sigma_u/\sigma_v = 10^2$). Panels C and D of Figure 12 show that as ρ increases from 0.5 to 0.95, the average Δ^C decreases from 0.76 to 0.62



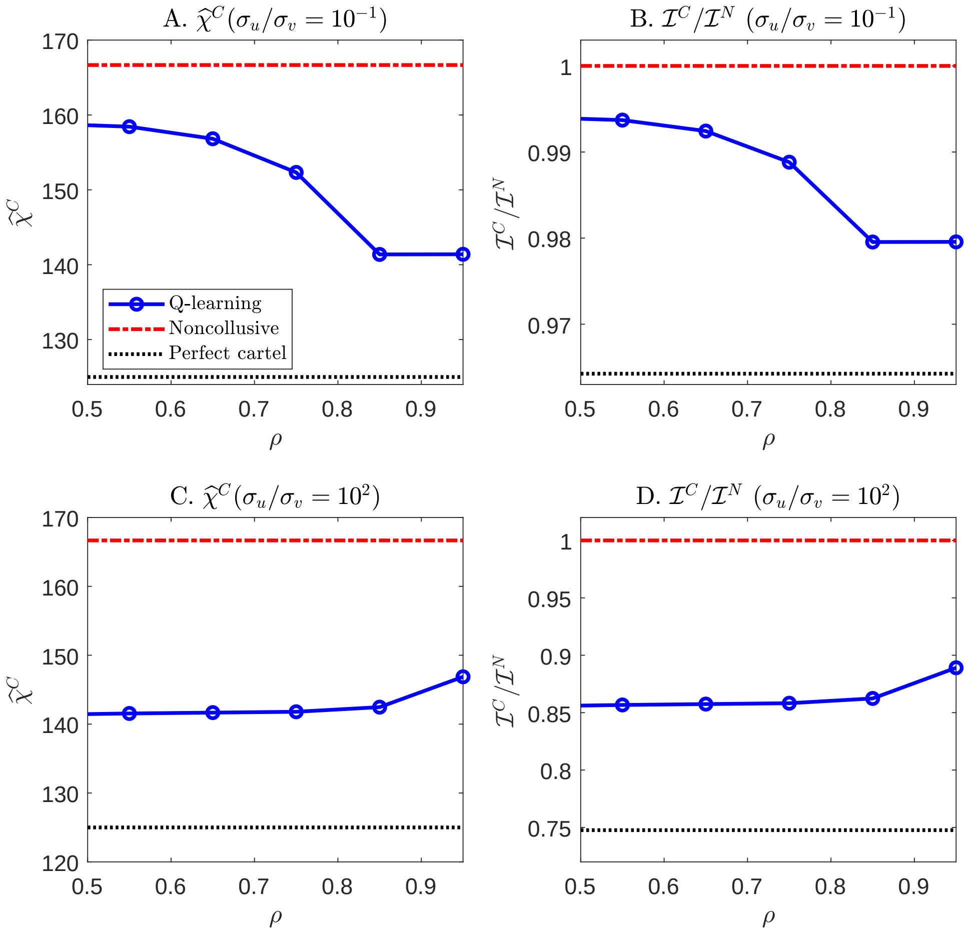
Note: This figure plots the average values of different metrics across $N = 1,000$ simulation sessions as the discount rate ρ varies. Panels A and B plot the average Δ^C and profit gain relative to noncollusion (π^C/π^N) under small information asymmetry ($\sigma_u/\sigma_v = 10^{-1}$). Panels C and D plot these metrics under large information asymmetry ($\sigma_u/\sigma_v = 10^2$). The blue solid line represents the simulation experiments with AI traders; the red dash-dotted and black dotted lines represent the theoretical benchmarks in the noncollusive Nash equilibrium and perfect cartel equilibrium, respectively. The other parameters are set according to the baseline economic environment described in Section 4.4.

Figure 12: Implications of the discount rate on Δ^C and π^C/π^N.

and the average π^C/π^N decreases from 1.10% to 1.08%, respectively. Moreover, panels C and D of Figure 13 show that as ρ increases, both $\widehat{\chi}^C$ and relative price informativeness $\mathcal{I}^C/\mathcal{I}^N$ increase.

6.3 Hyperparameters

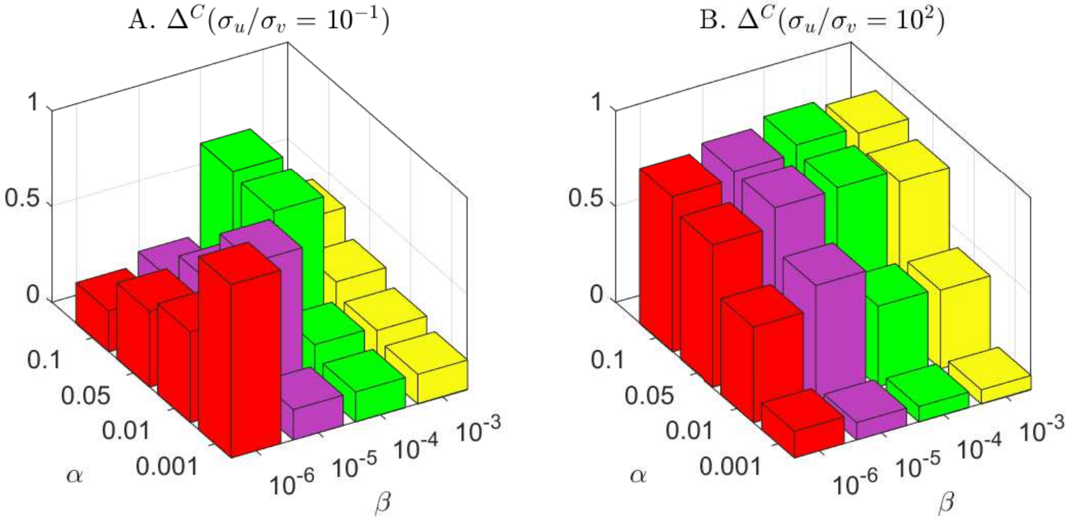
In this subsection, we study how the hyperparameters α and β affect AI traders' profits in equilibrium. Similar to the baseline economic environment, we consider AI traders



Note: This figure plots the average values of different metrics across $N = 1,000$ simulation sessions as the discount rate ρ varies. Panels A and B plot informed traders' order sensitivity to asset value ($\widehat{\chi}^C$) and the relative price informativeness ($\mathcal{I}^C/\mathcal{I}^N$) under small information asymmetry ($\sigma_u/\sigma_v = 10^{-1}$). Panels C and D plot these metrics under large information asymmetry ($\sigma_u/\sigma_v = 10^2$). The blue solid line represents the simulation experiments with AI traders; the red dash-dotted and black dotted lines represent the theoretical benchmarks in the noncollusive Nash equilibrium and perfect cartel equilibrium, respectively. The other parameters are set according to the baseline economic environment described in Section 4.4.

Figure 13: Implications of the discount rate on $\widehat{\chi}^C$ and $\mathcal{I}^C/\mathcal{I}^N$.

adopting the same value of α and β. In panel A of Figure 14, we plot the average Δ^C under small information asymmetry ($\sigma_u/\sigma_v = 10^{-1}$) for different values of α and β. As discussed in Subsection 5.1, AI traders need to conduct sufficient experimentations to learn punishment strategies, which is achieve through a long exploration process by setting a sufficiently low β. Indeed, when $\beta = 10^{-6}$, the red bars in panel A of Figure 14 show that AI traders can easily achieve a very high level of $\Delta^C = 0.90$ (corresponding to $\alpha = 0.001$) whereas when $\beta = 10^{-3}$, the yellow bars show that AI traders can only achieve a low level of $\Delta^C = 0.40$ (corresponding to $\alpha = 0.1$).

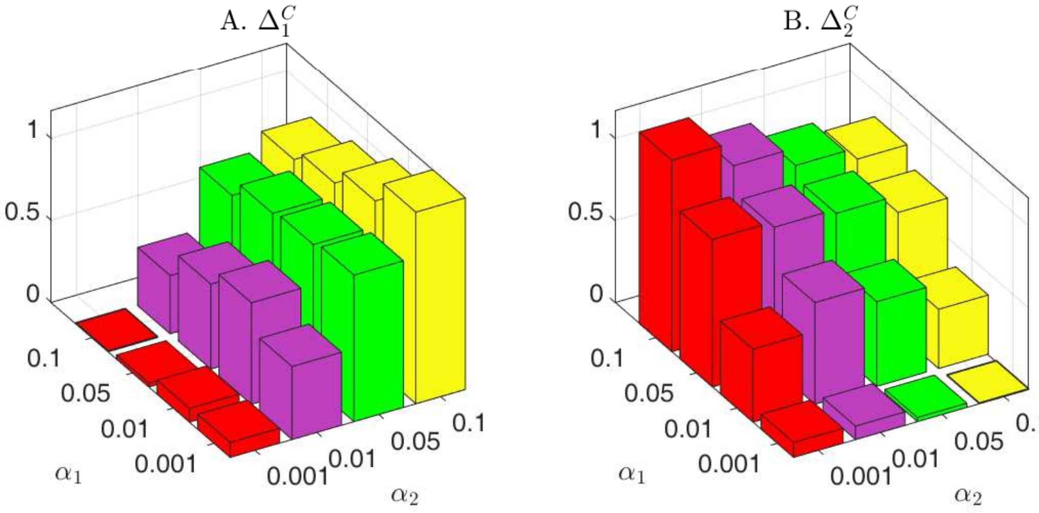


Note: Panel A plots Δ^C under small information asymmetry ($\sigma_u/\sigma_v = 10^{-1}$); panel B plots Δ^C under large information asymmetry ($\sigma_u/\sigma_v = 10^{2}$). The other parameters are set according to the baseline economic environment described in Section 4.4.

Figure 14: Implications of hyperparameters α and β on Δ^C.

Panel A of Figure 14 further shows that, to achieve the best collusive outcomes, the values of α and β have to be jointly determined. That is, the choice of a smaller β needs to be matched with a smaller α, and conversely, the choice of a larger β needs to be matched with a larger α. Intuitively, setting a small β ensures that AI traders will spend a long time in the exploration mode in which they randomly choose different actions, resulting in extensive experimentation. Then, setting a small α is necessary to record the value learned in the past whereas setting a large α will disrupt learning as the algorithm would forget what it has learned in the past too rapidly. By contrast, setting a large β means that AI traders only spend a short period of time in the exploration mode. Then, if we still set a small α, the Q-matrices of AI traders would not be updated significantly until the algorithms complete exploration. Thus, when β is large, setting a small α would backfire, making the initial exploration futile. Instead, setting a large α in this case would help AI traders to learn punishment strategies to achieve more collusive outcomes.

In panel B of Figure 14, we plot the average Δ^C under large information asymmetry ($\sigma_u/\sigma_v = 10^2$) for different values of α and β. Holding β unchanged at each value of $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$, panel B shows that the value of Δ^C declines monotonically as α decreases. This is because under large information asymmetry, the supra-competitive profits are attained because AI traders have biased learning. As discussed in Section 5.2,



Note: We allow the two AI traders to adopt different values of α, denoted by α_1 and α_2 for AI traders 1 and 2, respectively. We calculate the Delta metric for each trader separately, defined by $\Delta_i^C \equiv (\overline{\pi}_i - \overline{\pi}^N)/(\overline{\pi}^M - \overline{\pi}^N)$, where $\overline{\pi}_i \equiv \sum_{t=T_c}^{T_c+T} \pi_{i,t}(v_t, u_t)$, for $i = 1, 2$. Panels A and B plot Δ_1^C and Δ_2^C under large information asymmetry ($\sigma_u/\sigma_v = 10^2$). The other parameters are set according to the baseline economic environment described in Section 4.4.

Figure 15: Profit gain when AI traders adopt different values of α.

the biased learning due to the failure of the law of large numbers is mitigated when α becomes small (see equation (5.1)).

Taken together, a key feature that distinguishes collusion due to artificial intelligence (panel A of Figure 14) and collusion due to artificial stupidity (panel B of Figure 14) is whether improved learning through setting a sufficiently small α would significantly reduce the supra-competitive profits of AI traders.

Focusing on the economic environment with large information asymmetry, we now allow the two AI traders to adopt different values of α, but the same value of β. Intuitively, the AI trader adopting a smaller α would have less biased learning than the one adopting a larger α. As discussed in Subsection 5.2.4, biased learning induces AI traders to adopt more conservative trading strategies, i.e., placing orders with smaller amounts. Therefore, the AI trader with a larger α would adopt a more conservative trading strategy than the one with a smaller α. This essentially enables the AI trader with a smaller α would take advantage of the other AI trader and obtain more profits that what it would obtain when the other trader adopts the same α. Conversely, the AI trader with a larger α would obtain less profits than what it would obtain when the other trader adopts the same α.

The results of our simulation experiments are consistent with the above intuition. In

Figure 15, we allow each AI trader i to adopt different values of $\alpha_i = 0.001, 0.01, 0.05$ and 0.1 for $i = 1, 2$. Panels A and B plot the average Δ_1^C and Δ_2^C for AI traders 1 and 2, respectively. It is shown that for any combination of (α_1, α_2), the AI trader with a higher α_i attains a higher average Δ_i^C than the other AI trader. Moreover, holding α_1 unchanged at each value of $\{0.001, 0.01, 0.05, 0.1\}$, panel A shows that the average Δ_1^C for AI trader 1 increases as AI trader 2's α_2 increases. By contrast, holding α_2 unchanged at each value of $\{0.001, 0.01, 0.05, 0.1\}$, panel B shows that the average Δ_2^C for AI trader 2 increases as AI trader 1's α_1 increases.

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.

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.

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.

Harrington, Joseph E. 2019. "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.

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.

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

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

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 2.3

Given that $s_t = 0$, let $J^C(\chi_i)$ denote each informed trader 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 $x^C(v_t)$. 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

$$\begin{aligned} p^C(y_t) &= \overline{v} + \lambda^C y_t, \quad \text{with } \lambda^C = \frac{\theta\gamma^C + \xi}{\theta + \xi^2} \text{ and } \gamma^C = \frac{I\chi^C}{(I\chi^C)^2 + (\sigma_u/\sigma_v)^2} \\ y_t &= \chi_i(v_t - \overline{v}) + (I-1)x^C(v_t) + u_t. \end{aligned} \tag{B.2}$$

The probability of price trigger violation is

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

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

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

$$\begin{aligned} J^C(\chi^C) = &\left(1 - \lambda^C I \chi^C\right) \chi^C \sigma_v^2 \\ &+ \rho J^C(\chi^C)\Phi(\omega) \\ &+ \frac{\rho - \rho^T}{1-\rho} \left[1 - \Phi(\omega)\right] \mathbb{E}\left[\pi^N(v)\right] + \rho^T J^C(\chi^C)\left[1 - \Phi(\omega)\right]. \end{aligned} \tag{B.3}$$

Thus, we can obtain that

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

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

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

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.5), we obtain that

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

The first-order condition with respect to χ_i, characterized by $\nabla J^C(\chi^C) = 0$, leads to

$$\begin{aligned} 0 = &\left[1 - \lambda^C (I+1)\chi^C\right] \sigma_v^2 \\ &- \rho J^C(\chi^C)\frac{\sigma_v}{\sigma_u}\phi(\omega)\sqrt{\frac{2}{\pi}} \\ &+ \frac{\rho - \rho^T}{1-\rho}\mathbb{E}\left[\pi^N(v)\right]\frac{\sigma_v}{\sigma_u}\phi(\omega)\sqrt{\frac{2}{\pi}} \\ &+ \rho^T J^C(\chi^C)\frac{\sigma_v}{\sigma_u}\phi(\omega)\sqrt{\frac{2}{\pi}}, \end{aligned} \tag{B.7}$$

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

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, we only need 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 $\chi^M = \frac{\sqrt{I}}{I+1}\frac{\sigma_u}{\sigma_v}$. Denote $\chi^C = \hat{\chi}^C\frac{\sigma_u}{\sigma_v}$. Then, we show that there is no solution $\hat{\chi}^C \in [\frac{\sqrt{I}}{I+1}, \frac{1}{\sqrt{I}})$. In the Kyle case, $\mathbb{E}\left[\pi^N(v)\right] = \frac{\sigma_u\sigma_v}{(I+1)\sqrt{I}}$. Therefore, equations (B.4) and (B.7) can be rewritten, respectively, as follows:

$$J^C(\chi^C) = \frac{\left(1-\gamma^C I\chi^C\right)\chi^C\sigma_v^2 + \dfrac{\rho-\rho^T}{1-\rho}\left[1-\Phi(\omega)\right]\dfrac{\sigma_v\sigma_u}{(I+1)\sqrt{I}}}{1-\rho\Phi(\omega)-\rho^T\left[1-\Phi(\omega)\right]}. \tag{B.8}$$

and

$$0 = \left[1-\lambda^C(I+1)\chi^C\right]\sigma_v^2 - \left[\rho J^C(\chi^C) - \frac{\rho-\rho^T}{1-\rho}\frac{\sigma_v\sigma_u}{(I+1)\sqrt{I}} - \rho^T J^C(\chi^C)\right]\frac{\sigma_v}{\sigma_u}\phi(\omega)\sqrt{\frac{2}{\pi}} \tag{B.9}$$

Therefore, $\hat{\chi}^C$ is the root of the following quadratic equation that is different from $1/\sqrt{I}$:

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

Thus, we can obtain that

$$\hat{\chi}^C = -\frac{\left\{1-\rho+(\rho-\rho^T)[1-\Phi(\omega)]\right\}^{-1}\phi(\omega)\sqrt{\frac{2}{\pi}}}{\frac{I^2}{(I+1)\sqrt{I}}\left\{1-\rho+(\rho-\rho^T)[1-\Phi(\omega)]\right\}^{-1}\phi(\omega)\sqrt{\frac{2}{\pi}}+\frac{I}{\rho-\rho^T}} - \frac{1}{\sqrt{I}} < 0. \tag{B.11}$$

As a result, there is no root that lies in $[\frac{\sqrt{I}}{I+1}, \frac{1}{\sqrt{I}})$.

C Proof of Proposition 2.4

As $\theta \to 0$ or as $\xi \to \infty$, $\lambda^C \to 1/\xi$, that is, the market approaches to the environment where price is primarily determined by market clearing conditions. In this case, the demand of the preferred-habitat investor plays an important role. Because the system is continuous, we only need to show that there is a solution $\chi^C \in [\chi^M, \chi^N)$ in the environment of no price recovery, where $\chi^N = \frac{\xi}{I+1}$, $\chi^M = \frac{\xi}{2I}$, and $\mathbb{E}\left[\pi^N(v)\right] = \frac{\sigma_v^2}{(I+1)^2}$. We show that existence a solution $\chi^C \in [\frac{\xi}{I+1}, \frac{\xi}{2I})$. In this case, equations (B.4) and (B.7) can be rewritten, respectively, as follows:

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

and

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

Therefore, χ^C is the root of the following quadratic equation that is different from $\xi/(I+1)$:

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

where

$$K = 1-\rho\Phi(\omega)-\rho^T\left[1-\Phi(\omega)\right] \tag{C.3}$$

Thus, we can obtain that

$$\chi^C = \frac{\sigma_u}{\sigma_v}\left(1+\frac{1}{I}\right)\left[\frac{1-\rho^T}{\rho-\rho^T}-\Phi(\omega)\right]\phi(\omega)^{-1}\sqrt{\frac{2}{\pi}} + \frac{\xi}{I(I+1)} \tag{C.4}$$

To ensure that χ^C characterizes a collusion equilibrium, it requires that $\chi^C - \chi^N < 0$, that is,

$$\frac{\sigma_u}{\sigma_v}\left(1+\frac{1}{I}\right)\left[\frac{1-\rho^T}{\rho-\rho^T}-\Phi(\omega)\right]\phi(\omega)^{-1}\sqrt{\frac{2}{\pi}} - \frac{\xi(I-1)}{I(I+1)} < 0. \tag{C.5}$$

The above inequality is satisfied if information asymmetry σ_u/σ_v or I is not too large.

D Proof of Proposition 2.6

We only prove the proposition for the case of $\theta = 0$ here. More general cases with small θ or large ξ can be proved similarly with more involving derivations.

First, we compute $I\pi^C - I\pi^N$ as follows:

$$\begin{aligned} I\pi^C - I\pi^N &= \xi\left[1-\xi^{-1}\frac{\sigma_u}{\sigma_v}(I+1)A(\rho)-\frac{1}{I+1}\right]\left[\xi^{-1}\frac{\sigma_u}{\sigma_v}(I+1)A(\rho)+\frac{1}{I+1}\right] - \frac{\xi I}{(I+1)^2} \\ &= \xi\left[\xi^{-1}\frac{\sigma_u}{\sigma_v}(I+1)A(\rho)+\frac{1}{I+1}\right] - \xi\left[\left(\xi^{-1}\frac{\sigma_u}{\sigma_v}(I+1)A(\rho)\right)^2 + 2\xi^{-1}\frac{\sigma_u}{\sigma_v}A(\rho)+\frac{1}{(I+1)^2}\right] - \frac{\xi I}{(I+1)^2} \\ &= \xi\left[1-\xi^{-1}\frac{\sigma_u}{\sigma_v}(I+1)A(\rho)\right]\xi^{-1}\frac{\sigma_u}{\sigma_v}(I+1)A(\rho) - 2\frac{\sigma_u}{\sigma_v}A(\rho), \end{aligned}$$

where $A(\rho) \equiv \left[\frac{1-\rho^T}{\rho-\rho^T}-\Phi(\omega)\right]\phi(\omega)^{-1}\sqrt{\frac{2}{\pi}}$. We then compute $I\pi^M - I\pi^N$ as follows:

$$I\pi^M - I\pi^N = \xi\left[\frac{1}{4}-\frac{I}{(I+1)^2}\right] = \xi\frac{(I-1)^2}{4(I+1)^2} \tag{D.1}$$

Thus,

$$\Delta^C = \frac{4\left[1-\xi^{-1}\frac{\sigma_u}{\sigma_v}(I+1)A(\rho)\right]\xi^{-1}\frac{\sigma_u}{\sigma_v}(I+1)A(\rho) - 8\xi^{-1}\frac{\sigma_u}{\sigma_v}A(\rho)}{(I-1)^2/(I+1)^2} \tag{D.2}$$

The function $(1-x)x$ is strictly decreasing in x when $x > 1/2$. To ensure that $\chi^C \geq \chi^M$ for any I, we assume that $\xi^{-1}\frac{\sigma_u}{\sigma_v}(I+1)A(\rho) \geq 1/2$. Therefore, as I increases, $\xi^{-1}\frac{\sigma_u}{\sigma_v}(I+1)A(\rho) \geq 1/2$ increases, thereby making $\left[1-\xi^{-1}\frac{\sigma_u}{\sigma_v}(I+1)A(\rho)\right]\xi^{-1}\frac{\sigma_u}{\sigma_v}(I+1)A(\rho)$ decrease. In the meantime, as I increases, $(I-1)^2/(I+1)^2$ increases. Taken together, Δ^C decreases with I. Additionally, as σ_u/σ_v increases, $\xi^{-1}\frac{\sigma_u}{\sigma_v}(I+1)A(\rho) \geq 1/2$ increases, which decreases $\left[1-\xi^{-1}\frac{\sigma_u}{\sigma_v}(I+1)A(\rho)\right]\xi^{-1}\frac{\sigma_u}{\sigma_v}(I+1)A(\rho)$. In the meantime, as σ_u/σ_v increases, $\xi^{-1}\frac{\sigma_u}{\sigma_v}A(\rho)$ increases. Taken together, Δ^C decreases with σ_u/σ_v. Similarly, we can easily prove that Δ^C increases with ξ and ρ.

Price informativeness is

$$\begin{aligned}\mathcal{I}^C &= \log\left[\left(I\chi^C\right)^2(\sigma_v/\sigma_u)^2\right]\\ &= 2\log\left[(I+1)A(\rho)+\frac{\sigma_v}{\sigma_u}\frac{\xi}{I+1}\right]\end{aligned}$$

Because $\xi^{-1}\frac{\sigma_u}{\sigma_v}(I+1)A(\rho) \geq 1/2$, it follows that $(I+1)A(\rho)+\frac{\sigma_v}{\sigma_u}\frac{\xi}{I+1}$ increases with I. Consequently, price informativeness $\mathcal{I}^C$ is increasing in I. Obviously, price informativeness $\mathcal{I}^C$ is decreasing in ρ and σ_u/σ_v, and it is increasing in ξ.

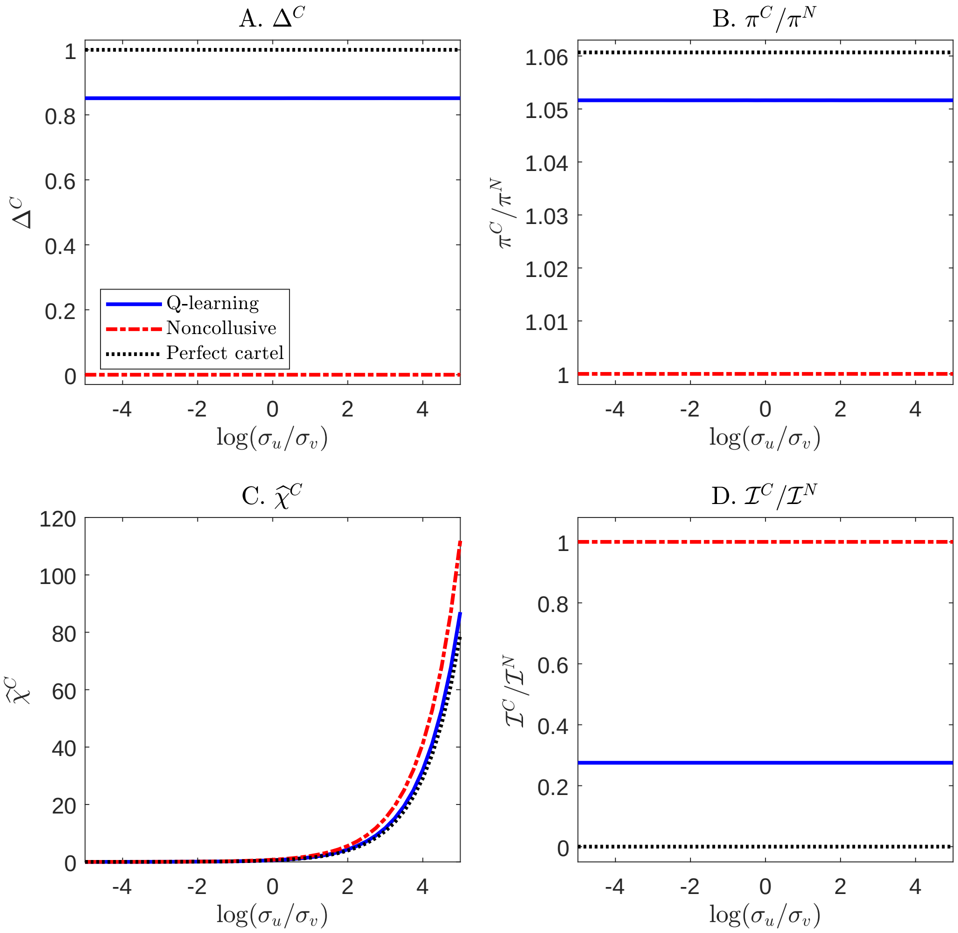
E Standard Kyle Setting with $\xi = 0$

In this appendix section, we study AI traders' behavior in the baseline economic environment except for setting $\xi = 0$, which essentially means that preferred-habitat investors do not exist. Thus, market makers set 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 traders. Our model in Section 2 shows that implicit collusion cannot be sustained by any price-trigger strategies when $\xi = 0$.

Figure A presents the simulation experiments with AI traders when $\xi = 0$. Although our model suggests that informed traders should not be able to achieve supra-competitive profits, our AI traders can attain an average Δ^C of 0.85 (panel A) and an average profit gain relative to noncollusion, $\pi^C/\pi^N = 1.05$ (panel B), due to biased learning. Moreover, AI traders' relative price informativeness (panel D) remain unchanged as $\log(\sigma_u/\sigma_v)$ varies along the x-axis, which is similar to the theoretical implication of the Kyle (1985) model. The AI traders' order sensitivity to asset value $\widehat{\chi}^C$ increases linearly with σ_u/σ_v (panel C).

F Market Makers with Q-Learning

In the baseline economic environment, market makers analyze historical data to estimate the pricing rule (ese Subsection 3.2). In this appendix section, we consider market makers adopting Q-learning algorithms to learn the pricing rule. All the results presented in



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, B, and C plot the average Δ^C, profit gain relative to noncollusion (π^C/π^N), and informed traders' order sensitivity to asset value ($\widehat{\chi}^C$). These metrics are defined in Section 4.6. Panel D plots the price informativeness relative to the theoretical benchmark of the noncollusive Nash equilibrium, i.e., $\mathcal{I}^C/\mathcal{I}^N$, where the price informativeness in the simulation experiment is calculated by $\mathcal{I}^C = \log\left[(I\widehat{\chi}^C)^2(\widehat{\sigma}_v/\sigma_u)^2\right]$. The blue solid line represents the simulation experiments with AI traders; the red dash-dotted and black dotted lines represent the theoretical benchmarks in the noncollusive Nash equilibrium and perfect cartel equilibrium, respectively. The parameters are set according to the baseline economic environment described in Section 4.4, except for $\xi = 0$.

Figure A: Implications of information asymmetry in the standard Kyle Setting with $\xi = 0$.

the main text are similar; they do not depend on whether market makers determine the pricing rule using statistical learning or Q-learning.

Below, we describe the Q-learning algorithms of market makers. We consider market makers adopting linear policies to price assets given total output y_t:

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

where v_t^{MM} and λ_t^{MM} are the decisions of market makers learned from their Q-learning programs. Specifically, market makers states are $s_t = \emptyset$ and actions are $a_t = \{v_t^{MM}, \lambda_t^{MM}\} \in \mathcal{V} \times \Lambda$. They update their Q-matrix according to the following learning equation:

$$\begin{aligned}\widehat{Q}_{t+1}^{MM}(v_t^{MM}, \lambda_t^{MM}) =&(1-\alpha^{MM})\widehat{Q}_t^{MM}(s_t, a_t) + \alpha \Bigg[(y_t - \xi(v_t^{MM} - \overline{v} + \lambda_t^{MM} y_t))^2 \\ &+\theta(v_t^{MM} + \lambda_t^{MM} y_t - v_t)^2 + \rho^{MM} \min_{v\in\mathcal{V},\lambda\in\Lambda} \widehat{Q}_t^{MM}(v,\lambda)\Bigg], \qquad \text{(F.2)}\end{aligned}$$

where the reward in period t is

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

The optimal decision v_t^{MM} and λ_t^{MM} are learned to minimize the Q-matrix. Similar to informed traders' Q-learning programs, market makers also do exploration with probability ε_t^{MM} and exploitation with $1 - \varepsilon_t^{MM}$. In the exploration mode, market makers randomly choose actions v and λ over the set $\mathcal{V} \times \Lambda$.

To implement the Q-learning programs for market makers, we construct discrete grid for v_t^{MM} and λ_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_λ 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 λ^{MM} correspond to the theoretical values in the noncollusive equilibrium. The parameter $\kappa > 0$ ensures that market makers can choose decisions different from these theoretical values.

For grid $(v_k^{MM}, \lambda_j^{MM}) \in \mathbb{V} \times \Lambda$, we initialize market makers' 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 (4.2). 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 parameters.