Internet Appendix for “AI-Powered Trading, Algorithmic Collusion, and Price Efficiency”

Winston Wei Dou   Itay Goldstein   Yan Ji

Abstract

This appendix provides supplemental materials for the paper titled “AI-Powered Trading, Algorithmic Collusion, and Price Efficiency” (Dou, Goldstein and Ji, 2024). Section 2 presents additional explanations for the model and sketches proofs for the propositions and lemmas. Section 3 provides detailed heuristic justifications and intuitions for AI equilibria. Section 4 provides supplemental material for the simulation results.

Keywords: Reinforcement learning, AI collusion, Homogenization, Experience-based and self-confirming equilibrium, Asymmetric information, Price informativeness, Market liquidity. (**JEL Classification:** D43, G10, G14, L13)

Contents

1 Further Literature Background

Our paper contributes to the growing literature on algorithmic collusion without explicit agreements, communication, or coded intent by introducing a novel mechanism, AI collusion through over-pruning bias in learning, and providing a comprehensive analysis of the theoretical benchmarks for different types of AI collusion and their underlying algorithmic mechanisms. Additionally, we examine which mechanism dominates under various market environments. Unlike our paper, which examines dynamically sophisticated Q-learning with endogenous state variables, earlier studies primarily focus on stateless Q-learning or simpler multi-armed bandit algorithms. Waltman and Kaymak (2008) show that even stateless Q-learning algorithms can sustain algorithmic collusion, provided the forgetting rate is high and the relative gains from collusion are substantial. More recently, Dolgopolov (2024) extended these results. However, their findings rely on the Boltzmann exploration scheme. When exploration follows a fixed-probability rule, such as the simple ε-greedy scheme used in this paper, their conclusions no longer hold. This distinction highlights that their algorithmic collusion mechanism is fundamentally different from ours, as it neither relies on price-trigger strategies nor stems from over-pruning bias in learning. Additionally, unlike this paper, some studies adopt a Markov perfect equilibrium framework. For example, Klein (2021) simulates a simplified sequential pricing duopoly where firms set prices alternately, focusing on Markov perfect equilibrium, similar to Brown and MacKay (2023), rather than simultaneous pricing, as examined in many other studies. While Klein (2021) also shows that Q-learning algorithms can coordinate on supra-competitive prices, its framework and underlying algorithmic mechanism are fundamentally different from ours by design.

In a contemporaneous study, Banchio and Mantegazza (2024) introduce a novel algorithmic collusion mechanism in stateless Q-learning, termed spontaneous coupling. This mechanism arises endogenously from statistical linkages between Q-learning algorithms' price estimates, leading to self-reinforcing cycles where firms sustain supra-competitive prices on average. Simulations confirm that spontaneous coupling can sustain price-fixing and market division. Importantly, learning rate asymmetries reinforce collusion, while greater noise disrupts it. This contrasts with the algorithmic collusion mechanisms highlighted in this paper, which operate through fundamentally different channels. Building on a similar stateless framework, Hansen, Misra and Pai (2021) show that duopolies using autonomous but dynamically unsophisticated reinforcement learning algorithms — specifically, upper confidence bound algorithms, a type of multi-armed bandit — can also sustain supra-competitive prices over time. In their setting, algorithms designed to learn demand through pricing experiments may misinterpret price sensitivity by failing to account for competitors' pricing behavior. This leads them to systematically overestimate their own price sensitivity. When firms engage in similar pricing experiments, their pricing decisions become correlated, further reinforcing supra-competitive pricing outcomes. A key difference between this mechanism and that in Banchio and Mantegazza (2024) is the source of correlation. In Hansen, Misra and Pai (2021), it arises through the timing of experimentation, while in Banchio and Mantegazza (2024), it stems from statistical coupling in price estimates generated under fully random experimentation. The mechanism in Hansen, Misra and Pai (2021) is fundamentally different from ours, because in stateless reinforcement learning, AI collusion cannot be sustained through price-trigger strategies, which require tracing past prices as state variables; additionally, greater noise strengthens AI collusion in our setting due to over-pruning bias in learning, whereas in Hansen, Misra and Pai (2021), greater noise weakens AI collusion. Lambin (2024) challenges Hansen, Misra and Pai (2021) by arguing that simultaneous exploration and slow learning, rather than price sensitivity overestimation, are the true drivers of correlated pricing and AI collusion. However, standard multi-agent reinforcement learning frameworks do not inherently require simultaneous moves in exploration, rendering Lambin (2024)' argument irrelevant for our proposed algorithmic mechanisms underlying AI collusion. Along similar lines, Abada and Lambin (2023) argue that AI collusion can arise from insufficient learning, particularly due to insufficient exploration. When reinforcement learning algorithms fail to sufficiently experiment with alternative pricing strategies, they may converge to stable outcomes that closely resemble collusive equilibria. However, Asker, Fershtman and Pakes (2024) show that while increasing exploration can reduce collusive tendencies, it does not eliminate algorithmic collusion. Expanding on this insight, a central contribution of this paper is to identify a novel mechanism underlying AI collusion: over-pruning bias in learning. This bias systematically and prematurely eliminates aggressive strategies from the set of potential optimal actions, thereby distorting the learning process. Its root cause lies in the asymmetric effect of exploitation, where adverse and beneficial shocks influence learning in fundamentally different ways. Exploitation reinforces strategies that appear successful based on past experience, while discouraging the reconsideration of those previously rated poorly due to past performance. As a result, reinforcement learning algorithms behave as if they are risk-averse to randomness in rewards. Since aggressive trading strategies are more vulnerable to adverse shocks from noise trading flows, they are disproportionately penalized and pruned prematurely from the learning path, reinforcing the over-pruning bias. This bias, in turn, embeds an implicit form of risk aversion into the algorithm's behavior, ultimately giving rise to AI collusion.

2 Model and Proofs

2.1 More Explanations on the Theoretical Benchmark

Further Explanations on the Model Setup. Analogous to Kyle and Xiong (2001), the demand curve (3.2) in the main text can be justified by a rational portfolio choice made by the information-insensitive investor under certain assumptions. These assumptions are summarized in Lemma 1 and the proof follows.

Lemma 1 (Demand Curve)**.** *If the information-insensitive investor possesses exponential utility with an* *absolute risk aversion coefficient of ϵ, then the demand curve has the functional form of (3.2) in the main text, where the slope ξ is given by $1/(\epsilon\sigma_v^2)$.*

Proof. The information-insensitive investor solves the following portfolio optimization problem for a given p_t:

$$\max_z \mathbb{E}\left[-e^{-\epsilon(v_t-p_t)z}/\epsilon\right]. \tag{IA.2.1}$$

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

$$0 = \left[(\overline{v} - p_t) - \epsilon z \sigma_v^2\right] e^{-\epsilon z(\overline{v}-p_t)+(\epsilon z)^2\sigma_v^2/2}. \tag{IA.2.2}$$

Thus, the optimal holding, z, is characterized as

$$z = -\frac{1}{\epsilon\sigma_v^2}(p_t - \overline{v}). \tag{IA.2.3}$$

□

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

The average asset holding of the information-insensitive investor, denoted as $\overline{z}$, is often substantial. This implies a small price elasticity of demand, given by $\varepsilon = \mathbb{E}[(\mathrm{d}z_t/\mathrm{d}p_t)(p_t/z_t)] = -\xi\mathbb{E}[p_t/z_t] \approx -\xi/\overline{z}$, as $\mathbb{E}[p_t]$ is close to $\overline{v} \equiv 1$. Studies indicate that information-insensitive 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).

Moreover, the concept of specifying exogenous demand curves within the framework of a noisy rational expectation equilibrium 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 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 information-insensitive investor to absorb the trading demand of informed speculators. The total supply $\widetilde{z}_t$ follows an exogenous noisy supply curve defined as:

$$\widetilde{z}_t = -u_t + \xi(p_t - \overline{v}), \tag{IA.2.4}$$

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

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

$$\min_{p_t} \mathbb{E}\left[(y_t + z_t)^2 + \theta(p_t - v_t)^2 \middle| y_t\right], \tag{IA.2.5}$$

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

The market maker's objective function (IA.2.5) captures both the inventory cost and asymmetric information faced by the market maker. Because the market maker takes the position $-(y_t + z_t)$ to clear the market, the term $(y_t + z_t)^2$ represents its inventory-holding costs. The quadratic form is adopted for tractability, consistent with the literature (e.g., Mildenstein and Schleef, 1983). The term $\theta(p_t - v_t)^2$ captures the market maker's efforts to reduce pricing errors arising from asymmetric information. The weight θ serves as a reduced-form way to capture the various benefits of reducing pricing errors, such as increased trading flows from a growing client base or enhanced competitive advantages over other trading platforms.[1] 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, $p_t = \mathbb{E}[v_t|y_t]$, as in the model of Kyle (1985).

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

$$x_{i,t} = \chi(v_t - \overline{v}), \quad \text{for all } i = 1, \cdots, I. \tag{IA.2.6}$$

[1] Similarly, in the context of e-commerce platforms, it is 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). The weight on consumer surplus in this context is a reduced-form way to capture various aspects of increasing consumer surplus. For example, increasing consumer surplus allows the platform to dynamically expand its consumer base over time and better compete with rival platforms.

The first-order condition of the minimization problem (IA.2.5) 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 market maker's pricing rule is

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

While the impact of the pricing error term may be minimal in practice, we choose to treat θ as a tiny, universally fixed positive constant in both our theoretical and simulation analyses. By fixing θ, we exclude it from the comparative-static analysis. This approach creates a more conceptually coherent theoretical framework or laboratory, featuring two meaningful extreme benchmarks. Specifically, as ξ approaches infinity, the price p_t converges to $\overline{v} + \xi^{-1}y_t$, set by the market clearing condition $y_t + z_t = 0$, as in Kyle and Xiong (2001). Conversely, as ξ decreases towards zero, p_t shifts to the efficient price $\mathbb{E}[v_t|y_t]$, as in Kyle (1985), if $\theta > 0$.

One of the main reasons for requiring $\theta > 0$ is to ensure a meaningful benchmark as $\xi \to 0$. If $\theta = 0$, the market price is determined solely by the market-clearing condition, leading to $p_t = \overline{v} + \xi^{-1}y_t$. As a result, in the extreme case where $\xi \to 0$, the benchmark becomes uninformative and does not provide a useful reference point for our analysis. Specifically, according to Propositions IA.1 and IA.2, it becomes evident that no informed speculator would act on their private information v_t, with $\chi^N \to 0$ in the non-collusive Nash equilibrium and $\chi^M \to 0$ in the perfect cartel benchmark, when $\xi \to 0$. This occurs while noise trading risk σ_u remain unchanged. As a result, the market price is $p_t^N = \overline{v} + \frac{I}{I+1}(v_t - \overline{v}) + \xi^{-1}u_t$ in the non-collusive Nash equilibrium, and it is $p_t^M = \overline{v} + \frac{1}{2}(v_t - \overline{v}) + \xi^{-1}u_t$ in the perfect cartel benchmark. Clearly, when $\xi \to 0$, the market price is infinitely volatile, solely driven by the noise trading flow u_t.

Suppose all informed speculators adopt the same trading strategy characterized by χ as in (IA.2.6). We define the trading profit function $\Pi(\chi, \lambda)$ as follows:

$$\Pi(\chi,\lambda) \equiv \mathbb{E}\left[\left(v_t - \overline{v} - \lambda\left(I\chi(v_t - \overline{v}) + u_t\right)\right)\chi(v_t - \overline{v})\right] = \left(1 - \lambda I \chi\right)\chi\sigma_v^2. \tag{IA.2.7}$$

Non-collusive Nash Equilibrium. We use the superscript N to denote the variables in the noncollusive Nash equilibrium. At the beginning of each period t, each informed speculator i

solves the following problem:

$$x^N(v_t) = \underset{x_i}{\operatorname{argmax}} \mathbb{E}\left[(v_t - p_t)\, x_i \middle| v_t\right], \tag{IA.2.8}$$

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 as $p_t = p^N(y_t)$. The equilibrium pricing function $p^N(\cdot)$ is given by:

$$p^N(y_t) = \overline{v} + \lambda^N y_t, \;\text{ with }\; \lambda^N = \frac{\theta\gamma^N + \xi}{\theta + \xi^2} \;\text{ and }\; \gamma^N = \frac{I\chi^N}{(I\chi^N)^2 + (\sigma_u/\sigma_v)^2}, \tag{IA.2.9}$$

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

$$y_t = x_i + (I-1)x^N(v_t) + u_t. \tag{IA.2.10}$$

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

Proposition IA.1. *The order flow of informed speculators and price in the non-collusive Nash equilibrium are*

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

where χ^N and λ^N satisfy

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

The expected profit of informed speculators is

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

Perfect Cartel Benchmark. 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

$$x^M(v_t) = \underset{x}{\operatorname{argmax}} \mathbb{E}\left[(v_t - p_t)\, x \middle| v_t\right], \tag{IA.2.11}$$

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 as $p_t = p^M(y_t)$. The equilibrium pricing function $p^M(\cdot)$ is given by:

$$p^M(y_t) = \overline{v} + \lambda^M y_t, \;\text{ with }\; \lambda^M = \frac{\theta\gamma^M + \xi}{\theta + \xi^2} \;\text{ and }\; \gamma^M = \frac{I\chi^M}{(I\chi^M)^2 + (\sigma_u/\sigma_v)^2}, \tag{IA.2.12}$$

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

$$y_t = Ix + u_t. \tag{IA.2.13}$$

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

Proposition IA.2. *The order flow of informed speculators and price in the perfect cartel benchmark 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

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

The expected profit of informed speculators is

$$\pi^M = \left(1 - \lambda^M I\chi^M\right)\chi^M\sigma_v^2 = \frac{\sigma_v^2}{4I\lambda^M}.$$

Collusive Nash Equilibrium. Information asymmetry in capital markets makes grim-trigger strategies ineffective for sustaining tacit collusion, as investors cannot accurately observe or monitor each other's actions.[2] However, tacit collusion can still be sustained under information asymmetry through price-trigger strategies with imperfect monitoring. If an informed speculator can reliably infer other informed speculators' total order flows from the market price, collusive incentives can be created.

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

We now describe the collusive Nash equilibrium sustained by price-trigger strategies under information asymmetry, as studied by Green and Porter (1984). Specifically, we focus on the symmetric collusive Nash equilibrium in which all I informed speculators choose the same

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

collusive order flow, denoted by $x^C(v_t)$. Such trading strategies are sustained by a price-trigger strategy: Informed speculators are initially in the "normal" state, $s_0^C = 0$, in period 0, and submit their respective order flows $x^C(v_t)$. Informed speculators 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 $v_t > \overline{v}$ in some period $t > 0$. Then, in period $t+1$, they will enter the "reversionary" state, $s_{t+1}^C = 1$, with probability η, or stay in the normal state, $s_{t+1}^C = 0$, with probability $1-\eta$. If they enter in the reversionary state in period $t+1$, they will trade noncollusively and continue to do so with the same probability η in each subsequent period up to period $t+T$. Thus, the reversionary episode can lasts for at most T periods.

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

$$x^C(v) \equiv \chi^C(v - \overline{v}), \tag{IA.2.14}$$

$$p^C(y) = \overline{v} + \lambda^C y, \ \text{ with } \ \lambda^C = \frac{\theta\gamma^C + \xi}{\theta + \xi^2} \ \text{ and } \ \gamma^C = \frac{I\chi^C}{(I\chi^C)^2 + \sigma_u^2/\sigma_v^2}. \tag{IA.2.15}$$

The price-trigger function $q(v)$ is specified based on the expected price when all informed speculators trade coordinately according to $x^C(v)$ conditional on v, namely, $\overline{p}^C(v) \equiv \mathbb{E}\left[p^C(y)|v\right]$. Specifically, plugging (IA.2.14) into (IA.2.15) and taking expectation over u, we obtain that $\overline{p}^C(v) \equiv \overline{v} + \lambda^C I \chi^C (v - \overline{v})$. The price-trigger function $q(v)$ is specified as follows:

$$q(v) \equiv \begin{cases} \overline{p}^C(v) + \lambda^C \sigma_u \omega, & \text{if } v > \overline{v} \\ \overline{p}^C(v) - \lambda^C \sigma_u \omega, & \text{if } v < \overline{v}, \end{cases} \tag{IA.2.16}$$

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

Equation (IA.2.16) warrants further discussions. First, when $v > \overline{v}$, informed investors have incentives to buy a large amount of the asset, which boosts up its price. As a result, when $v > \overline{v}$, a meaningful price-trigger strategy would punish the potential deviating counterparty by reverting to the noncollusive Nash equilibrium once the market price goes above a certain high-level threshold $q(v)$. In contrast, when $v < \overline{v}$, informed investors have incentives to sell a large amount of the asset, which suppresses down its price. As a result, when $v < \overline{v}$, a meaningful price-trigger strategy would punish the potential deviating counterparty by reverting to the noncollusive Nash equilibrium once the market price falls below a certain low-level threshold $q(v)$. Second, there is no price threshold when $v = \overline{v}$ because no informed investor would have incentives to trade in this case. Third, although there are infinitely many alternative ways to specify the functional form of the threshold $q(v)$, we focus on a specification that is not only statistically meaningful but also ensures a linear model solution as in Kyle (1985). If no one deviates from the coordinated trading, each informed speculator can infer that the noise order is $\widehat{u}_t = [p_t - \overline{p}^C(v)]/\lambda^C$ based on the observed price $p_t = p^C(y_t)$. If $\widehat{u}_t$ is excessively positive when $v_t > \overline{v}$, say $\widehat{u}_t > \omega\sigma_u$ for some constant $\omega > 0$, the informed speculator would suspect that some other informed speculators might have deviated from the implicit agreement. Analogously, if $\widehat{u}_t$ is excessively negative when $v_t < \overline{v}$, say $\widehat{u}_t < -\omega\sigma_u$ for some constant $\omega > 0$, the informed speculator would suspect that some other informed speculators might have deviated from the implicit agreement. Fourth, the multiplier 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.

2.2 Proof of Lemma 1

The information-insensitive investor solves the following portfolio optimization problem for a given p_t:

$$\max_z \mathbb{E}\left[-e^{-\epsilon(v_t-p_t)z}/\epsilon\right]. \tag{IA.2.17}$$

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

$$0 = \left[(\overline{v} - p_t) - \epsilon z \sigma_v^2\right] e^{-\epsilon z(\overline{v}-p_t)+(\epsilon z)^2\sigma_v^2/2}. \tag{IA.2.18}$$

Thus, the optimal holding, z, is characterized as

$$z = -\frac{1}{\epsilon\sigma_v^2}(p_t - \overline{v}). \tag{IA.2.19}$$

2.3 Proof of Proposition 3.1

Impossibility of Price-Trigger Collusive Nash Equilibrium. Given that $s_t^C = 0$ (i.e., informed speculators are in the collusive regime in period t), let $J^C(\chi_i)$ denote each informed speculator i's expected present value of future profits, when investor i chooses $x_{i,t} = \chi_i(v_t - \overline{v})$ and all other $I - 1$ informed investors choose $x^C(v_t) = \chi^C(v_t - \overline{v})$. The value of $J^C(\chi_i)$ satisfies the following recursive relation:

$$\begin{aligned} J^C(\chi_i) =& \mathbb{E}\left[\left(v_t - p^C(y_t)\right)\chi_i(v_t - \overline{v})\right] \\ &+ \rho J^C(\chi_i) \times \mathbb{P}\left\{\text{Price trigger is not violated in period } t \middle| \chi_i, \chi^C\right\} \\ &+ \rho\mathbb{E}\left[\sum_{\tau=1}^{T}(\rho\eta)^{\tau-1}\left[\eta\pi^N(v_{t+\tau}) + (1-\eta)J^C(\chi_i)\right] + (\rho\eta)^T J^C(\chi_i)\right] \\ &\qquad \times \left[1 - \mathbb{P}\left\{\text{Price trigger is not violated in period } t \middle| \chi_i, \chi^C\right\}\right], \end{aligned} \tag{IA.2.20}$$

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

$$p^C(y_t) = \overline{v} + \lambda^C y_t, \text{ with } \lambda^C = \frac{\theta\gamma^C + \xi}{\theta + \xi^2} \text{ and } \gamma^C = \frac{I\chi^C}{(I\chi^C)^2 + (\sigma_u/\sigma_v)^2}, \tag{IA.2.21}$$

$$y_t = \chi_i(v_t - \overline{v}) + (I-1)x^C(v_t) + u_t. \tag{IA.2.22}$$

The probability that the price trigger is not violated in period t is

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

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

Evaluating equality (IA.2.20) at $\chi_i = \chi^C$ leads to

$$\begin{aligned}
J^C(\chi^C) = {}& \left(1 - \lambda^C I\chi^C\right)\chi^C\sigma_v^2 + \rho J^C(\chi^C)\Phi(\omega) + \frac{\rho\eta\left[1-(\rho\eta)^T\right]}{1-\rho\eta}\mathbb{E}\left[\pi^N(v)\right]\left[1 - \Phi(\omega)\right] \\
&+ \frac{\rho(1-\eta) + (1-\rho)(\rho\eta)^{T+1}}{1-\rho\eta} J^C(\chi^C)\left[1 - \Phi(\omega)\right].
\end{aligned} \tag{IA.2.23}$$

Thus, we can obtain that

$$J^C(\chi^C) = \frac{\left(1 - \lambda^C I\chi^C\right)\chi^C\sigma_v^2 + \dfrac{\rho\eta\left[1-(\rho\eta)^T\right]}{1-\rho\eta}\mathbb{E}\left[\pi^N(v)\right]\left[1-\Phi(\omega)\right]}{1 - \rho\Phi(\omega) - \dfrac{\rho(1-\eta) + (1-\rho)(\rho\eta)^{T+1}}{1-\rho\eta}\left[1-\Phi(\omega)\right]}. \tag{IA.2.24}$$

The first-order derivative of the both sides of (IA.2.20) with respect to χ_i, evaluated at $\chi_i = \chi^C$, is

$$\begin{aligned}
\nabla J^C(\chi^C) = {}& \left[1 - \lambda^C(I+1)\chi^C\right]\sigma_v^2 + \frac{\rho\eta[1-(\rho\eta)^T]}{1-\rho\eta}\frac{1}{\sigma_u}\phi(\omega)\mathbb{E}\left[|v - \overline{v}|\right]\mathbb{E}\left[\pi^N(v)\right] \\
&+ \left[\frac{\rho(1-\eta) + (1-\rho)(\rho\eta)^{T+1}}{1-\rho\eta} - \rho\right] J^C(\chi^C)\frac{1}{\sigma_u}\phi(\omega)\mathbb{E}\left[|v - \overline{v}|\right] \\
&+ \rho\left[\nabla J^C(\chi^C)\right]\Phi(\omega) + \frac{\rho(1-\eta) + (1-\rho)(\rho\eta)^{T+1}}{1-\rho\eta}\left[\nabla J^C(\chi^C)\right]\left[1 - \Phi(\omega)\right],
\end{aligned} \tag{IA.2.25}$$

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

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

(IA.2.25), we obtain that

$$\begin{aligned}\nabla J^C(\chi^C) = {} & \left[1-\lambda^C(I+1)\chi^C\right]\sigma_v^2 + \frac{\rho\eta[1-(\rho\eta)^T]}{1-\rho\eta}\frac{\sigma_v}{\sigma_u}\phi(\omega)\sqrt{\frac{2}{\pi}}\mathbb{E}\left[\pi^N(v)\right] \\ & + \left[\frac{\rho(1-\eta)+(1-\rho)(\rho\eta)^{T+1}}{1-\rho\eta}-\rho\right]J^C(\chi^C)\frac{\sigma_v}{\sigma_u}\phi(\omega)\sqrt{\frac{2}{\pi}} \\ & +\rho\left[\nabla J^C(\chi^C)\right]\Phi(\omega)+\frac{\rho(1-\eta)+(1-\rho)(\rho\eta)^{T+1}}{1-\rho\eta}\left[\nabla J^C(\chi^C)\right][1-\Phi(\omega)],\end{aligned} \tag{IA.2.26}$$

The policy variable χ^C constitutes a collusive Nash equilibrium if speculator i has no incentive to deviate by setting $\chi_i \neq \chi^C$. The first-order condition with respect to χ_i, characterized by $\nabla J^C(\chi^C) = 0$, leads to

$$\begin{aligned}0 = {} & \left[1-\lambda^C(I+1)\chi^C\right]\sigma_v^2 + \frac{\rho\eta[1-(\rho\eta)^T]}{1-\rho\eta}\frac{\sigma_v}{\sigma_u}\phi(\omega)\sqrt{\frac{2}{\pi}}\mathbb{E}\left[\pi^N(v)\right] \\ & + \left[\frac{\rho(1-\eta)+(1-\rho)(\rho\eta)^{T+1}}{1-\rho\eta}-\rho\right]J^C(\chi^C)\frac{\sigma_v}{\sigma_u}\phi(\omega)\sqrt{\frac{2}{\pi}}.\end{aligned} \tag{IA.2.27}$$

We first consider the condition (i) that ξ is tiny relative to θ. In this case, $\xi/\theta \approx 0$, and thus, according to (IA.2.21), $\lambda^C \approx \gamma^C$. As a result, the market resembles the setting of Kyle (1985), where the demand from information-insensitive investors is negligible. Given that $\lambda^C = \gamma^C$, it follows that

$$\chi^N = \frac{1}{\sqrt{I}}\frac{\sigma_u}{\sigma_v} \text{ and } \chi^M = \frac{1}{I}\frac{\sigma_u}{\sigma_v}. \tag{IA.2.28}$$

Given the system's continuity, it suffices to show that no solution $\chi^C \in (0, \chi^N)$ exists such that $\Pi(\chi^C, \lambda^C) > \Pi(\chi^N, \lambda^N)$ in the Kyle (1985) environment, where $\Pi(\cdot,\cdot)$ is defined in (IA.2.7). Let $\chi^C = \widetilde{\chi}^C\frac{\sigma_u}{\sigma_v}$, and define $\widetilde{\chi}^M \equiv \frac{1}{I}$ and $\widetilde{\chi}^N = \frac{1}{\sqrt{I}}$. Then, it is equivalent to show that there is no solution $\widetilde{\chi}^C \in (0, \widetilde{\chi}^N)$ such that $\Pi(\widetilde{\chi}^C\frac{\sigma_u}{\sigma_v}, \lambda^C) > \Pi(\widetilde{\chi}^N\frac{\sigma_u}{\sigma_v}, \lambda^N)$.

In the Kyle case, $\mathbb{E}\left[\pi^N(v)\right] = \frac{\sigma_u\sigma_v}{(I+1)\sqrt{I}}$. Therefore, equations (IA.2.24) and (IA.2.27) can be rewritten, respectively, as follows:

$$J^C(\chi^C) = \frac{\left(1-\gamma^C I\chi^C\right)\chi^C\sigma_v^2 + \dfrac{\rho\eta\left[1-(\rho\eta)^T\right]}{1-\rho\eta}[1-\Phi(\omega)]\dfrac{\sigma_u\sigma_v}{(I+1)\sqrt{I}}}{1-\rho\Phi(\omega)-\dfrac{\rho(1-\eta)+(1-\rho)(\rho\eta)^{T+1}}{1-\rho\eta}[1-\Phi(\omega)]}. \tag{IA.2.29}$$

and

$$\begin{aligned}0 = {} & \left[1-\gamma^C(I+1)\chi^C\right]\sigma_v^2 + \frac{\rho\eta[1-(\rho\eta)^T]}{1-\rho\eta}\frac{\sigma_v}{\sigma_u}\phi(\omega)\sqrt{\frac{2}{\pi}}\frac{\sigma_u\sigma_v}{(I+1)\sqrt{I}} \\ & + \left[\frac{\rho(1-\eta)+(1-\rho)(\rho\eta)^{T+1}}{1-\rho\eta}-\rho\right]J^C(\chi^C)\frac{\sigma_v}{\sigma_u}\phi(\omega)\sqrt{\frac{2}{\pi}}.\end{aligned} \tag{IA.2.30}$$

After plugging (IA.2.29) into (IA.2.30) and rearranging terms, we obtain that $\widetilde{\chi}^C$ is the root of the following quadratic equation:

$$0 = 1 - I(\widetilde{\chi}^C)^2 - \vartheta \left\{ \widetilde{\chi}^C - \frac{1}{(I+1)\sqrt{I}} \left[1 + (I\widetilde{\chi}^C)^2\right] \frac{\epsilon(1-\rho)}{\rho - \kappa} \right\}, \tag{IA.2.31}$$

where the coefficients are

$$\vartheta = \frac{(\rho-\kappa)\phi(\omega)}{1 - \rho\Phi(\omega) - \kappa[1-\Phi(\omega)]}\sqrt{\frac{2}{\pi}}, \tag{IA.2.32}$$

$$\kappa = \frac{\rho(1-\eta) + (1-\rho)(\rho\eta)^{T+1}}{1-\rho\eta}, \tag{IA.2.33}$$

$$\epsilon = \frac{\rho\eta[1-(\rho\eta)^T]}{1-\rho\eta}. \tag{IA.2.34}$$

The algebra is cumbersome in the general case with $\eta \in (0,1)$. To simplify, we focus on $\eta = 1$ without loss of generality, ensuring that punishment occurs with probability one when the price trigger is violated. The case with $\eta \in (0,1)$ is isomorphic to the case with $\eta = 1$ after recalibrating ω and T to achieve the same intensity and expected profit loss of punishment. Therefore, restricting attention to $\eta = 1$ preserves generality in our proof.

When $\eta = 1$, we have $\kappa = \rho^{T+1}$ and $\epsilon = \frac{\rho(1-\rho^T)}{1-\rho}$. Thus, equations (IA.2.29) to (IA.2.32) are simplified as follows:

$$J^C(\chi^C) = \frac{\left(1 - \gamma^C I \chi^C\right)\chi^C\sigma_v^2 + \frac{\rho(1-\rho^T)}{1-\rho}\left[1-\Phi(\omega)\right]\frac{\sigma_u\sigma_v}{(I+1)\sqrt{I}}}{1 - \rho\Phi(\omega) - \rho^{T+1}\left[1-\Phi(\omega)\right]}. \tag{IA.2.35}$$

and

$$0 = \left[1 - \gamma^C(I+1)\chi^C\right]\sigma_v^2 + \left[\frac{\rho(1-\rho^T)}{1-\rho}\frac{\sigma_u\sigma_v}{(I+1)\sqrt{I}} + \left(\rho^{T+1} - \rho\right)J^C(\chi^C)\right]\frac{\sigma_v}{\sigma_u}\phi(\omega)\sqrt{\frac{2}{\pi}}. \tag{IA.2.36}$$

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

$$0 = 1 - I(\widetilde{\chi}^C)^2 - \vartheta \left\{ \widetilde{\chi}^C - \frac{1}{(I+1)\sqrt{I}} \left[1 + (I\widetilde{\chi}^C)^2\right] \right\}, \tag{IA.2.37}$$

where

$$\vartheta = \frac{\phi(\omega)}{\frac{1-\rho}{\rho-\rho^{T+1}} + 1 - \Phi(\omega)}\sqrt{\frac{2}{\pi}}. \tag{IA.2.38}$$

Solving the above problem, we obtain

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

There are three cases, which we examine below.

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

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

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

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

which is negative. Thus, there does not exist a solution $\widetilde{\chi}^C$ that lies in $(0, \frac{1}{\sqrt{I}})$, meaning that the collusive Nash equilibrium does not exist.

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

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

and the other root, which is larger, is greater than or equal to $\widetilde{\chi}^N$. Thus, there does not exist a solution $\widetilde{\chi}^C$ that lies in $(0, \frac{1}{\sqrt{I}})$, meaning that the collusive equilibrium does not exist.

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

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

Thus, the larger root is

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

and the smaller root is

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

For the smaller root to lie in $(0, \frac{1}{\sqrt{I}})$, we need $-I + \vartheta \frac{I\sqrt{I}}{I+1} > 0$ given that $\vartheta > 0$, which implies

$$\vartheta > \frac{I+1}{\sqrt{I}}.$$

Since $\frac{I+1}{\sqrt{I}} < 2\sqrt{I}\frac{I+1}{I-1}$, the existence of a collusive Nash equilibrium requires that $\vartheta > 2\sqrt{I}\frac{I+1}{I-1}$ in this case. To rule this out, we need to show the condition $\vartheta \leq 2\sqrt{I}\frac{I+1}{I-1}$, which ensures that Case 3 does not arise. Since $\frac{1-\rho}{\rho-\rho^{T+1}}$ is always positive for any $\rho \in (0,1)$, it follows that

$$\vartheta < \sqrt{\frac{2}{\pi}} H(\omega), \tag{IA.2.39}$$

where $H(\omega) \equiv \phi(\omega) / \left[1 - \Phi(\omega)\right]$ is the hazard rate function of the standard normal distribution. Because $H(\omega)$ is monotonically increasing when $\omega > 0$, it follows from (IA.2.39) that, when $\bar{\omega} = 8$,

$$\vartheta < \sqrt{\frac{2}{\pi}} H(\bar{\omega}) = 6.0515. \tag{IA.2.40}$$

On the other hand, it is straightforward to show that, for any integer $I \geq 2$,

$$2\sqrt{I}\frac{I+1}{I-1} \geq 6.666. \tag{IA.2.41}$$

Combining results from (IA.2.40) and (IA.2.41), we obtain

$$\vartheta < 2\sqrt{I}\frac{I+1}{I-1}. \tag{IA.2.42}$$

Therefore, Case 3 cannot occur.

We then consider the condition (ii) that σ_u is large. In this case, according to (IA.2.21), $\gamma^C \approx 0$ and thus $\lambda^C \approx \frac{\xi}{\theta+\xi^2}$. As a result, the market approximates a trading environment in which price recovery through market makers' learning about the fundamental value, based on the total trade flow of informed speculators and noise traders, is negligible. Given that $\sigma_v/\sigma_u \approx 0$, $\chi^N \approx \frac{1}{I+1}\frac{\theta+\xi^2}{\xi}$ and $\chi^M \approx \frac{1}{2I}\frac{\theta+\xi^2}{\xi}$. Due to the continuity, it suffices to show that no solution $\chi^C \in (0, \chi^N)$ exists such that $\Pi(\chi^C, \frac{\xi}{\theta+\xi^2}) > \Pi(\chi^N, \frac{\xi}{\theta+\xi^2})$ in the trading environment, where $\Pi(\cdot,\cdot)$ is defined in (IA.2.7). Let $\chi^C = \widetilde{\chi}^C \frac{\theta+\xi^2}{\xi}$, and define $\widetilde{\chi}^M \equiv \frac{1}{2I}$ and $\widetilde{\chi}^N \equiv \frac{1}{I+1}$. Then, it is equivalent to show that there is no solution $\widetilde{\chi}^C \in (0, \widetilde{\chi}^N)$ satisfies $\Pi(\widetilde{\chi}^C \frac{\theta+\xi^2}{\xi}, \frac{\xi}{\theta+\xi^2}) > \Pi(\widetilde{\chi}^N \frac{\theta+\xi^2}{\xi}, \frac{\xi}{\theta+\xi^2})$. In this case, $\mathbb{E}\left[\pi^N(v)\right] = \frac{\sigma_v^2}{(I+1)^2}\frac{\theta+\xi^2}{\xi}$. Given that $\sigma_v/\sigma_u \approx 0$, equation (IA.2.27) simplifies to:

$$\left(\chi^C - \frac{1}{I+1}\right)\left[\chi^C - \frac{1}{I(I+1)} - \frac{I+1}{I}\frac{\sigma_u}{\sigma_v}\frac{1}{\vartheta}\right] = 0. \tag{IA.2.43}$$

According to the definition of $\widetilde{\chi}^C$ here, which is $\widetilde{\chi}^C = \chi^C \frac{\xi}{\theta+\xi^2}$, the equation above can be further rewritten as:

$$\left(\widetilde{\chi}^C - \frac{1}{I+1}\right)\left[\widetilde{\chi}^C - \frac{1}{I(I+1)} - \frac{I+1}{I}\frac{\sigma_u}{\sigma_v}\frac{\xi}{\theta+\xi^2}\frac{1}{\vartheta}\right] = 0, \tag{IA.2.44}$$

Solving for $\widetilde{\chi}^C$ yields $\widetilde{\chi}^C = \frac{1}{I+1}$ or $\widetilde{\chi}^C = \frac{1}{I(I+1)} + \frac{I+1}{I}\frac{\sigma_u}{\sigma_v}\frac{\xi}{\theta+\xi^2}\frac{1}{\vartheta}$, with the second root becoming arbitrarily large as $\sigma_v/\sigma_u \to 0$. This implies that no solution exists within $(0, \widetilde{\chi}^N)$. Consequently, a collusive Nash equilibrium does not exist when $\sigma_v/\sigma_u \approx 0$.

Existence of Price-Trigger Collusive Nash Equilibrium. Given that $s_t^C = 0$ (i.e., informed speculators are in the collusive regime in period t), let $J^C(\chi_i)$ denote each informed speculator i's expected present value of future profits, when investor i chooses $x_{i,t} = \chi_i(v_t - \overline{v})$ and all other $I-1$ informed investors choose $x^C(v_t) = \chi^C(v_t - \overline{v})$. The value of $J^C(\chi_i)$ satisfies the following recursive relation in (IA.2.20). If $\theta/\xi \approx 0$, it must hold that $\lambda^C \approx 1/\xi$.

When $\lambda^C = 1/\xi$, it holds that $\chi^N = \frac{\xi}{I+1}$, $\chi^M = \frac{\xi}{2I}$, and $\mathbb{E}\left[\pi^N(v)\right] = \frac{\xi\sigma_v^2}{(I+1)^2}$. Because the system is continuous, it is sufficient to show that there exists a solution $\chi^C \in (0, \chi^N)$ such that $\Pi(\chi^C, 1/\xi) > \Pi(\chi^N, 1/\xi)$ in this environment. Without loss of generality, we focus on the case with $\eta = 1$, allowing us to rewrite Equations (IA.2.24) and (IA.2.27) as follows:

$$J^C(\chi^C) = \frac{\left(1 - \xi^{-1}I\chi^C\right)\chi^C\sigma_v^2 + \frac{\rho-\rho^{T+1}}{1-\rho}\left[1-\Phi(\omega)\right]\frac{\xi\sigma_v^2}{(I+1)^2}}{1 - \rho\Phi(\omega) - \rho^{T+1}\left[1-\Phi(\omega)\right]} \tag{IA.2.45}$$

and

$$0 = \left[1 - \xi^{-1}(I+1)\chi^C\right]\sigma_v^2 - \left[\rho J^C(\chi^C) - \frac{\rho-\rho^{T+1}}{1-\rho}\frac{\xi\sigma_v^2}{(I+1)^2} - \rho^{T+1}J^C(\chi^C)\right]\frac{\sigma_v}{\sigma_u}\phi(\omega)\sqrt{\frac{2}{\pi}}. \tag{IA.2.46}$$

Therefore, χ^C is the root of the following quadratic equation:

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

where

$$\vartheta_\sigma = \frac{\sigma_v}{\sigma_u}\vartheta = \frac{\sigma_v}{\sigma_u}\frac{\phi(\omega)}{\frac{1-\rho}{\rho-\rho^{T+1}} + 1 - \Phi(\omega)}\sqrt{\frac{2}{\pi}}. \tag{IA.2.47}$$

Solving the above problem, we obtain

$$\chi^C = \frac{\vartheta_\sigma + \frac{I+1}{\xi} \pm \left|\frac{\vartheta_\sigma(I-1)}{I+1} - \frac{I+1}{\xi}\right|}{\frac{2\vartheta_\sigma I}{\xi}}. \tag{IA.2.48}$$

There are two cases.

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

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

Thus, the larger root must be strictly greater than χ^N, ruling out the possibility of a collusive Nash equilibrium. To guarantee the existence of a collusive Nash equilibrium, we must show that this scenario never arises when θ/ξ and σ_u are both small. Specifically, to ensure that this case never occurs, it requires that $\vartheta_\sigma > \xi^{-1}\frac{(I+1)^2}{I-1}$, which, according to (IA.2.47), holds when σ_u is sufficiently small.

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

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

The smaller root is

$$\chi^C = \frac{\vartheta_\sigma + \frac{I+1}{\xi} - \left(\frac{\vartheta_\sigma(I-1)}{I+1} - \frac{I+1}{\xi}\right)}{\frac{2\vartheta_\sigma I}{\xi}} = \frac{\frac{\vartheta_\sigma \xi}{I+1} + I + 1}{\vartheta_\sigma I}. \tag{IA.2.49}$$

It is obvious that the smaller root lies between 0 and χ^N. To show that Equation (IA.2.49) characterizes a valid collusive Nash equilibrium, we need to verify that (IA.2.49) satisfies

$$\Pi(\chi^C, 1/\xi) > \Pi(\chi^N, 1/\xi), \tag{IA.2.50}$$

when θ/ξ and σ_u are both small. To prove the inequality in (IA.2.50), it suffices to show that

$$\left(1 - \xi^{-1} I \frac{\frac{\vartheta_\sigma \xi}{I+1} + I + 1}{\vartheta_\sigma I}\right) \frac{\frac{\vartheta_\sigma \xi}{I+1} + I + 1}{\vartheta_\sigma I} > \left(1 - \xi^{-1} I \frac{\xi}{I+1}\right) \frac{\xi}{I+1}, \tag{IA.2.51}$$

which can be further simplified into

$$\left(1 - \frac{\varepsilon_\sigma}{I}\right)(1 + \varepsilon_\sigma) > 1, \tag{IA.2.52}$$

with $\varepsilon_\sigma \equiv (I+1)^2/(\xi\vartheta_\sigma)$.

The inequality in (IA.2.52) is equivalent to

$$\varepsilon_\sigma < I - 1, \tag{IA.2.53}$$

which can be rewritten as

$$\xi\vartheta_\sigma > \frac{(I+1)^2}{I-1}. \tag{IA.2.54}$$

According to the definition of ϑ_σ in (IA.2.47), the inequality (IA.2.54) holds as long as σ_u is sufficiently close to zero. Therefore, with all other parameters held constant, a collusive Nash equilibrium through price-trigger strategies exists if both ξ is large relative to θ and σ_u is small.

2.4 Proof of Proposition 3.2

We first show the existence of a collusive experience-based equilibrium with trading strategy $\chi^C \in [\chi^M, \chi^N)$ for any $\xi > 0$ and $\sigma_u > 0$, where χ^N and χ^M characterize the optimal trading strategy of informed investors in the non-collusive Nash equilibrium and the perfect cartel benchmark, respectively. We show that, for any χ^C such that $\chi^M \leq \chi^C < \chi^N$, there exists an experience-based equilibrium, in which informed speculators uniformly undervalue aggressive trading strategies, perpetuating an incorrect system of outcome evaluation that remains uncorrected. Along the equilibrium path, the market price $p^C(y)$ satisfies

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

where y is the total order flow of informed speculators and noise traders.

If χ^C characterizes the trading strategy in an experience-based equilibrium, the "correct" outcome evaluation of any possible strategy χ is given by:

$$J(\chi) \equiv \mathbb{E}\left[(v_t - p^C(y_t))\chi(v_t - \overline{v})\right] + \rho J(\chi^C), \tag{IA.2.55}$$

where $y_t = \chi(v_t - \overline{v}) + (I-1)\chi^C(v_t - \overline{v}) + u_t$. It thus follows that

$$J(\chi) \equiv \left[1 - \lambda^C\chi - \lambda^C(I-1)\chi^C\right]\chi\sigma_v^2 + \rho J^C(\chi^C). \tag{IA.2.56}$$

As a result, the optimal trading strategy χ^C and its evaluation $J(\chi^C)$ in an experience-based equilibrium χ^C must satisfy the following condition:

$$J(\chi^C) = \frac{\sigma_v^2}{1-\rho}\left(1 - \lambda^C I\chi^C\right)\chi^C. \tag{IA.2.57}$$

Below, we establish existence by construction. Specifically, we construct a distortion function $D(\chi)$ that is strictly increasing in χ, i.e., $\dfrac{\partial D(\chi)}{\partial \chi} > 0$, and ensures that that χ^C can be sustained as an experience-based equilibrium outcome by an evaluation system

$$J^C(\chi) \equiv J(\chi) - D(\chi), \tag{IA.2.58}$$

If such a distortion function $D(\chi)$ exists, its strict increase in χ implies that the perceived value of more aggressive trading strategies is systematically lower, leading informed speculators to undervalue them.

Particularly, we construct a distortion function as follows:

$$D(\chi) \equiv \varsigma^C[1-(I+1)\lambda^C\chi^C]\sigma_v^2(\chi-\chi^C) + \frac{\varsigma}{2}\sigma_u^2\sigma_v^2\left[\chi^2-(\chi^C)^2\right], \quad \text{where } \varsigma^C \geq 0. \tag{IA.2.59}$$

It is obvious that $D(\chi)$ is strictly increasing in χ for $\chi > 0$. This follows from

$$\frac{\partial D(\chi)}{\partial \chi} = \varsigma^C\left[1-(I+1)\lambda^C\chi^C\right]\sigma_v^2 + \varsigma\sigma_u^2\sigma_v^2\chi. \tag{IA.2.60}$$

Because $\chi^C < \chi^N$, it holds that $\lambda^C\chi^C < \lambda^N\chi^N$ following the proof provided in Online Appendix 2.5. Thus, $\left[1-(I+1)\lambda^C\chi^C\right]\sigma_v^2 > \left[1-(I+1)\lambda^N\chi^N\right]\sigma_v^2 = 0$. Thus, the first term on the right-hand side of (IA.2.60) is positive. Since the second term on the right-hand side of (IA.2.60) is also clearly positive for $\chi > 0$, we conclude that

$$\frac{\partial D(\chi)}{\partial \chi} > 0, \quad \text{for } \chi > 0. \tag{IA.2.61}$$

Now, we verify the three conditions for the experience-based equilibrium. First, the trading game clearly follows a recurrent Markovian state process.

Second, the optimality condition for the experience-based equilibrium stipulates that the strategies must be optimal, given the evaluations system $J^C(\chi)$ in (IA.2.58). Consequently, $\left.\frac{\partial J^C(\chi)}{\partial \chi}\right|_{\chi=\chi^C} = 0$, leading to the condition that $\left.\frac{\partial J(\chi)}{\partial \chi}\right|_{\chi=\chi^C} = \left.\frac{\partial D(\chi)}{\partial \chi}\right|_{\chi=\chi^C}$. Thus, the first-order condition requires that

$$0 = (\varsigma^C - 1)\left[1-(I+1)\lambda^C\chi^C\right]\sigma_v^2 + \varsigma\sigma_u^2\sigma_v^2\chi^C. \tag{IA.2.62}$$

As a result, as long as the two free parameters ς^C and ς are chosen to satisfy the first-order condition (IA.2.62), the distortion function $D(\chi)$ is consistent with the optimal trading strategy χ^C.

The optimality condition also requires the Hessian condition at χ^C:

$$\left.\frac{\partial^2 J^C(\chi)}{\partial \chi^2}\right|_{\chi=\chi^C} < 0. \tag{IA.2.63}$$

The second-order derivative is $\left.\frac{\partial^2 J^C(\chi)}{\partial \chi^2}\right|_{\chi=\chi^C} = \left.\frac{\partial^2 J(\chi)}{\partial \chi^2}\right|_{\chi=\chi^C} - \left.\frac{\partial^2 D(\chi)}{\partial \chi^2}\right|_{\chi=\chi^C}$. Thus, it follows that

$$\left.\frac{\partial^2 J^C(\chi)}{\partial \chi^2}\right|_{\chi=\chi^C} = -2\lambda^C\sigma_v^2 - \varsigma\sigma_u^2\sigma_v^2 < -2\frac{\xi}{\theta+\xi^2}\sigma_v^2 - \varsigma\sigma_u^2\sigma_v^2 < 0. \tag{IA.2.64}$$

It is clear that the inequality in (IA.2.64) ensures that the condition in (IA.2.63) is satisfied. Given a chosen $\varsigma \geq 0$ that satisfies (IA.2.63), the remaining free parameter $\varsigma^C \in [0,1)$ can be set to ensure that the first-order condition in (IA.2.62) holds exactly.

The third condition for the experience-based equilibrium stipulates that equilibrium path behaviors yielding expected discounted net cash flows consistent with equilibrium path evaluations. Specifically, this condition requires that $J(\chi^C) = J^C(\chi^C)$, which holds true since $D(\chi^C) = 0$ as specified in (IA.2.59).

We next show the existence of a collusive experience-based equilibrium driven by over-perceived aversion against noise trading risk. If χ^C characterizes the trading strategy in an experience-based equilibrium, the "correct" outcome evaluation of the strategy χ^C is given by $J(\chi)$ in (IA.2.56). However, the "incorrect" outcome evaluation of the strategy χ^C due to the over-perceived noise trading risk aversion is given by

$$J^C(\chi) = \left[1 - \lambda^C\chi - \lambda^C(I-1)\chi^C\right]\chi\sigma_v^2 - D(\chi) + \rho J^C(\chi^C), \tag{IA.2.65}$$

where $D(\chi)$ is a distortion function specified as in (IA.2.59) with ς^C set to be zero.

In other words, the outcome evaluation is systematically biased purely due to the over-perceived noise trading risk aversion, captured by $D(\chi)$ in equation (IA.2.65). Our proof below focuses on demonstrating that there exists $\chi^C \in [\chi^M, \chi^N)$ and ς^C such that the trading strategy χ^C and the "incorrect" outcome evaluation system can be sustained in a collusive experience-based equilibrium driven by over-perceived aversion against noise trading risk. We consider $\varsigma^C = 0$. The evaluation of the equilibrium trading strategy χ^C is consistent with the valuation relation on the equilibrium path:

$$J^C(\chi^C) = \left(1 - \lambda^C I\chi^C\right)\chi^C\sigma_v^2 + \rho J^C(\chi^C) \quad \Rightarrow \quad J^C(\chi^C) = \frac{\sigma_v^2}{1-\rho}\left(1 - \lambda^C I\chi^C\right)\chi^C. \tag{IA.2.66}$$

This implies that $J^C(\chi^C) = J(\chi^C)$, that is, the equilibrium path behavior χ^C yielding expected discounted net cash flows consistent with equilibrium path evaluations.

Moreover, the equilibrium trading strategy χ is optimal under the "incorrect" outcome evaluation system of χ, which is characterized as follows:

$$\chi^C = \underset{\chi}{\operatorname{argmax}}\, J^C(\chi), \quad \text{which is defined in (IA.2.65).} \tag{IA.2.67}$$

The first-order condition of the above maximization problem is

$$\chi^C = \frac{1}{(I+1)\lambda^C + \varsigma\sigma_u^2}. \tag{IA.2.68}$$

As specified in (3.4) of the main text, $\theta \approx 0$ holds under the assumed specification. When $\theta = 0$, the first-order condition can be simplified into

$$\chi^C = \frac{1}{(I+1)\xi^{-1} + \varsigma\sigma_u^2} < \frac{1}{(I+1)\xi^{-1}} = \chi^N. \tag{IA.2.69}$$

Thus, as long as θ is sufficiently close to zero, the continuity implies that $\chi^C < \chi^N$.

At the same time, as long as ς is sufficiently close to zero so that $\varsigma\sigma_u^2 \leq (I-1)\xi^{-1}$, it holds that

$$\chi^C = \frac{1}{(I+1)\xi^{-1} + \varsigma\sigma_u^2} \geq \frac{1}{2I\xi^{-1}} = \chi^M. \tag{IA.2.70}$$

In addition, $\left.\frac{\partial^2 J^C(\chi)}{\partial\chi^2}\right|_{\chi=\chi^C} = -2\lambda^C\sigma_v^2 - \varsigma\sigma_u^2\sigma_v^2 < 0$. Thus, χ^C is the optimal trading strategy according to the evaluation system $J^C(\chi)$.

Finally, we verify that the "incorrect" outcome evaluation system is biased relative to the "correct" system, leading to a uniform underweighting of aggressive trading strategies. In fact, it is clear that

$$\frac{\partial D(\chi)}{\partial\chi} = \varsigma\chi\sigma_u^2\sigma_v^2 > 0, \text{ because } \chi > 0. \tag{IA.2.71}$$

2.5 Proof of Proposition 3.3

Informed speculators' order flows are given by $x(v_t,\chi) \equiv \chi(v_t - \overline{v})$, where χ is determined in equilibrium. We denote $\lambda(\chi) = \frac{\theta\gamma(\chi)+\xi}{\theta+\xi^2}$ with $\gamma(\chi) = \frac{I\chi}{(I\chi)^2+(\sigma_u/\sigma_v)^2}$. As defined in (IA.2.7), the profit function satisfies that

$$\begin{aligned}
\Pi(\chi,\lambda(\chi)) &= \mathbb{E}\left[(v_t - \overline{v} - \lambda(\chi)(Ix(v_t,\chi)+u_t))x(v_t,\chi)\right] \\
&= \mathbb{E}\left[(v_t - \overline{v} - \lambda(\chi)(I\chi(v_t-\overline{v})+u_t))\chi(v_t-\overline{v})\right] \\
&= \left[1 - \lambda(\chi)I\chi\right]\chi\sigma_v^2 \\
&= \left[1 - \frac{\theta}{\theta+\xi^2}\frac{(I\chi)^2}{(I\chi)^2+(\sigma_u/\sigma_v)^2} - \frac{\xi}{\theta+\xi^2}I\chi\right]\chi\sigma_v^2.
\end{aligned}$$

The first-order derivative of $\Pi(\chi,\lambda(\chi))$ is

$$\frac{\mathrm{d}\Pi(\chi,\lambda(\chi))}{\mathrm{d}\chi} = \sigma_v^2\left[1 - \frac{\theta I^2}{\theta+\xi^2}\frac{I^2\chi^2+3(\sigma_u/\sigma_v)^2}{[I^2+(\sigma_u/\sigma_v)^2/\chi^2]^2} - \frac{2\xi I}{\theta+\xi^2}\chi\right]. \tag{IA.2.72}$$

By definition of the perfect cartel benchmark, we have $\chi^M = \operatorname{argmax}\Pi(\chi,\lambda(\chi))$, which implies $\left.\frac{\Pi(\chi,\lambda(\chi))}{\mathrm{d}\chi}\right|_{\chi=\chi^M} = 0$. Equation (IA.2.72) shows that $\frac{\mathrm{d}\Pi(\chi,\lambda(\chi))}{\mathrm{d}\chi}$ is decreasing in χ. Thus, for $\chi > \chi^M$, we have $\frac{\mathrm{d}\Pi(\chi,\lambda(\chi))}{\mathrm{d}\chi} < 0$. Consequently, defining $\pi^i \equiv \Pi(\chi^i,\lambda(\chi^i))$ for $i \in \{N, M, C\}$, it follows that $\pi^C \in (\pi^N, \pi^M]$, which implies $\Delta^C \in (0,1]$, since $\chi^C \in [\chi^M, \chi^N)$.

2.6 Proof of Proposition 3.4

We prove the proposition in the environment where $\theta = 0$ to simplify the derivation. More general environments with sufficiently small $\theta > 0$, can be handled similarly with more complex derivations. By continuity, if the results hold for $\theta = 0$, they also hold for sufficiently small $\theta > 0$.

Proof for part (i). Given the assumption of this proposition, we can restrict the analysis to parameter choices that ensure the existence of a collusive Nash equilibrium.

Proof for the profit ratio Δ^C. Suppose π^N, π^C, and π^M are defined as $\pi^i \equiv \Pi(\chi^i, \lambda(\chi^i))$ for $i \in \{N, M, C\}$, where $\Pi(\cdot,\cdot)$ is defined in (IA.2.7). When $\theta = 0$, the expected profit associated with χ^i can be expressed as follows:

$$\pi^i = (1 - \xi^{-1} I \chi^i)\chi^i \sigma_v^2, \;\; \text{for } i \in \{N, M, C\}.$$

Thus, $\pi^C - \pi^N$ can be expressed as follows:

$$\pi^C - \pi^N = \left(\frac{I}{I+1} - \frac{I+1}{\vartheta_\sigma \xi}\right)\left[\frac{\xi}{I(I+1)} + \frac{I+1}{\vartheta_\sigma I}\right]\sigma_v^2 - \frac{\xi \sigma_v^2}{(I+1)^2},$$

where $\vartheta_\sigma = \dfrac{\phi(\omega)}{\dfrac{1-\rho}{\rho - \rho^{T+1}} + 1 - \Phi(\omega)} \dfrac{\sigma_v}{\sigma_u}\sqrt{\dfrac{2}{\pi}}$. At the same time, $\pi^M - \pi^N$ can be expressed as follows:

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

Thus, Δ^C is

$$\Delta^C = \frac{4}{(I-1)^2}\left[I - \frac{(I+1)^2}{\vartheta_\sigma \xi}\right]\left[1 + \frac{(I+1)^2}{\vartheta_\sigma \xi}\right] - \frac{4I}{(I-1)^2}. \tag{IA.2.73}$$

When I is sufficiently large, it follows that

$$\frac{(I+1)^2}{\vartheta_\sigma \xi} > \frac{I-1}{2}. \tag{IA.2.74}$$

Within this range of parameter space, it holds that Δ^C is increasing in ϑ_σ. Because ϑ_σ is decreasing in σ_u, Δ^C is decreasing in σ_u. The subjective discount rate parameter ρ affects Δ^C only through its impact on ϑ_σ. Clearly, Δ^C is increasing in ρ if and only if $\frac{1-\rho}{\rho-\rho^T}$ is decreasing in ρ. We know that

$$\frac{1-\rho}{\rho - \rho^T} = \frac{1}{\sum_{j=1}^{T-1} \rho^j}, \tag{IA.2.75}$$

Since $\sum_{j=1}^{T-1} \rho^j$ is increasing in ρ, it follows that $\frac{1-\rho}{\rho-\rho^T}$ is decreasing in ρ, establishing the result.

Next, we show that Δ^C is decreasing in I. The first-order derivative is

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

Thus, $\frac{\partial \Delta^C}{\partial I} < 0$, when $\vartheta_\sigma \xi < \frac{2(I+1)^2}{I-1}$ as in (IA.2.74). The condition is true when I is sufficiently large. Intuitively, when σ_u/σ_v is sufficiently small (as required to ensure the existence of the collusive Nash equilibrium sustained by price-trigger strategies), it follows that

$$\Delta^C \approx \frac{4(I+1)^2}{(I-1)\vartheta_\sigma \xi}. \tag{IA.2.76}$$

which implies that Δ^C is increasing in I when $I \geq 3$.

Proof for the price informativeness $\mathcal{I}^C$. By definition, the price informativeness $\mathcal{I}^C$ is

$$\mathcal{I}^C = \left(I\chi^C\right)^2 (\sigma_v/\sigma_u)^2 = \left(\frac{\xi}{I+1}+\frac{I+1}{\vartheta_\sigma}\right)^2\left(\frac{\sigma_v}{\sigma_u}\right)^2, \tag{IA.2.77}$$

where $\vartheta_\sigma = \dfrac{\phi(\omega)}{\dfrac{1-\rho}{\rho-\rho^{T+1}}+1-\Phi(\omega)}\dfrac{\sigma_v}{\sigma_u}\sqrt{\dfrac{2}{\pi}}$. The price informativeness $\mathcal{I}^M$ is

$$\mathcal{I}^M = \left(I\chi^M\right)^2 (\sigma_v/\sigma_u)^2 = \left(\frac{\xi}{2}\right)^2\left(\frac{\sigma_v}{\sigma_u}\right)^2. \tag{IA.2.78}$$

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

$$\frac{\mathcal{I}^C}{\mathcal{I}^M} = \left[\frac{2}{I+1}+\frac{2(I+1)}{\vartheta_\sigma \xi}\right]^2. \tag{IA.2.79}$$

Clearly, when I is sufficiently large so that $\vartheta_\sigma \xi < (I+1)^2$, it follows that $\mathcal{I}^C/\mathcal{I}^M$ is increasing in I. Moreover, since $\mathcal{I}^C/\mathcal{I}^M$ is decreasing in ϑ_σ, which, in turn, is increasing in ρ and decreasing in σ_u, it holds that $\mathcal{I}^C/\mathcal{I}^M$ is decreasing in ρ and increasing in σ_u.

Proof for the mispricing $\mathcal{E}^C$. When $\theta = 0$, by the definition of mispricing and substituting out χ^C using (IA.2.49), the mispricing $\mathcal{E}^C$ is

$$\mathcal{E}^C = \left(1-\lambda^C I\chi^C\right)|v_t-\overline{v}| = \left(1-\xi^{-1}I\frac{\frac{\vartheta_\sigma \xi}{I+1}+I+1}{\vartheta_\sigma I}\right)|v_t-\overline{v}| = \left(1-\frac{1}{I+1}-\frac{I+1}{\vartheta_\sigma \xi}\right)|v_t-\overline{v}|. \tag{IA.2.80}$$

The mispricing in the perfect cartel benchmark, denoted by $\mathcal{E}^M$, is

$$\mathcal{E}^M = \left(1 - \lambda^M I \chi^M\right) |v_t - \overline{v}| = \left(1 - \xi^{-1} I \frac{\xi}{2I}\right) |v_t - \overline{v}| = |v_t - \overline{v}| /2. \tag{IA.2.81}$$

Thus, $\mathcal{E}^C/\mathcal{E}^M$ is equal to

$$\mathcal{E}^C/\mathcal{E}^M = 2\left(1 - \frac{1}{I+1} - \frac{I+1}{\vartheta_\sigma \xi}\right). \tag{IA.2.82}$$

Obviously, $\mathcal{E}^C/\mathcal{E}^M$ is increasing in ϑ_σ. Since ϑ_σ is increasing in ρ and decreasing in σ_u, $\mathcal{E}^C/\mathcal{E}^M$ is increasing in ρ and decreasing in σ_u. In addition, when I is sufficiently large so that $\vartheta_\sigma \xi < (I+1)^2$, it follows that $\mathcal{E}^C/\mathcal{E}^M$ is decreasing in I.

Proof for the market liquidity $\mathcal{L}^C$. By definition, the market liquidity $\mathcal{L}^C$ is

$$\mathcal{L}^C = \frac{1}{\partial |z_t + y_t| / \partial u_t} = \frac{1}{1 - \xi \lambda^C}. \tag{IA.2.83}$$

In this environment with $\theta/\xi \approx 0$, the market liquidity $\mathcal{L}^C$ can be expressed as follows:

$$\mathcal{L}^C = \frac{1}{\left|1 - \xi \frac{\theta \gamma^C + \xi}{\theta + \xi^2}\right|} \approx \frac{1}{\left|1 - \xi \frac{\theta \gamma^C + \xi}{\xi^2}\right|} = \frac{\xi}{\theta \gamma^C} = \frac{\xi}{\theta}\left(I\chi^C + \frac{\sigma_u^2}{\sigma_v^2} \frac{1}{I\chi^C}\right). \tag{IA.2.84}$$

When σ_u/σ_v is sufficiently small (as required to ensure the existence of the collusive Nash equilibrium sustained by price-trigger strategies), it follows that

$$\mathcal{L}^C \approx \frac{\xi}{\theta} I \chi^C \tag{IA.2.85}$$

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

$$\mathcal{L}^M = \frac{\xi}{\theta}\left(I\chi^M + \frac{\sigma_u^2}{\sigma_v^2}\frac{1}{I\chi^M}\right) = \frac{\xi}{\theta}\left(I\frac{\xi}{2I} + \frac{\sigma_u^2}{\sigma_v^2}\frac{1}{I\frac{\xi}{2I}}\right) = \frac{1}{\theta}\left(\frac{\xi^2}{2} + \frac{2\sigma_u^2}{\sigma_v^2}\right). \tag{IA.2.86}$$

When σ_u/σ_v is sufficiently small (as required to ensure the existence of the collusive Nash equilibrium sustained by price-trigger strategies), it follows that

$$\mathcal{L}^M \approx \frac{\xi^2}{2\theta}. \tag{IA.2.87}$$

Thus, the relative market liquidity $\mathcal{L}^C/\mathcal{L}^M$ can be expressed as follows:

$$\frac{\mathcal{L}^C}{\mathcal{L}^M} = 2\xi^{-1} I \chi^C = 2\left(\frac{1}{I+1} + \frac{I+1}{\vartheta_\sigma \xi}\right). \tag{IA.2.88}$$

Obviously, $\mathcal{L}^C/\mathcal{L}^M$ is decreasing in ϑ_σ. Since ϑ_σ is increasing in ρ and decreasing in σ_u, $\mathcal{L}^C/\mathcal{L}^M$ is decreasing in ρ and decreasing in σ_u. In addition, when I is sufficiently large so that $\vartheta_\sigma \xi < (I+1)^2$, it follows that $\mathcal{L}^C/\mathcal{L}^M$ is increasing in I.

Proof for part (ii). Given the assumption of this proposition, we can restrict the analysis to parameter choices that ensure the existence of a collusive experience-based equilibrium sustained by an over-perceived aversion to noise trading risk, captured by $\varsigma > 0$. According to the proof for Proposition 3.2 in Online Appendix 2.4, we focus on the case with

$$\varsigma \sigma_u^2 \xi \leq I - 1. \tag{IA.2.89}$$

Proof for the profit ratio Δ^C. Suppose π^N, π^C, and π^M are defined as $\pi^i \equiv \Pi(\chi^i, \lambda(\chi^i))$ for $i \in \{N, M, C\}$, where $\Pi(\cdot, \cdot)$ is defined in (IA.2.7). When $\theta = 0$, the expected profit associated with χ^i can be expressed as follows:

$$\pi^i = (1 - \xi^{-1} I \chi^i)\chi^i \sigma_v^2, \ \text{ for } i \in \{N, M, C\}.$$

Thus, based on (IA.2.70), $\pi^C - \pi^N$ can be expressed as follows:

$$\begin{aligned} \pi^C - \pi^N &= \left[1 - \xi^{-1} I \frac{\xi}{(I+1) + \varsigma \sigma_u^2 \xi}\right] \frac{\xi}{(I+1) + \varsigma \sigma_u^2 \xi} \sigma_v^2 - \frac{\xi \sigma_v^2}{(I+1)^2} \\ &= \frac{\xi \sigma_v^2}{(I+1)^2} \frac{\varsigma \sigma_u^2 \xi \left[(I+1)(I-1) - \varsigma \sigma_u^2 \xi\right]}{\left[(I+1) + \varsigma \sigma_u^2 \xi\right]^2}. \end{aligned} \tag{IA.2.90}$$

At the same time, $\pi^M - \pi^N$ can be expressed as follows:

$$\pi^M - \pi^N = \frac{\xi \sigma_v^2}{4I} - \frac{\xi \sigma_v^2}{(I+1)^2} = \xi \frac{(I-1)^2}{4I(I+1)^2} \sigma_v^2. \tag{IA.2.91}$$

From (IA.2.90) and (IA.2.91), we know that

$$\Delta^C = \frac{4I}{(I-1)^2} \frac{(I+1)(I-1)z - 1}{\left[(I+1)z + 1\right]^2}, \ \text{ where } z \equiv \frac{1}{\varsigma \sigma_u^2 \xi}. \tag{IA.2.92}$$

Clearly, Δ^C is not affected by ρ at all. The first-order derivative of Δ^C with respect to z is

$$\frac{\partial \Delta^C}{\partial z} = \frac{4I(I+1)^2}{(I-1)^2} \frac{1 - (I-1)z}{[(I+1)z + 1]^3} \tag{IA.2.93}$$

When the inequality in (IA.2.89) holds, it follows that

$$\frac{\partial \Delta^C}{\partial z} < 0, \tag{IA.2.94}$$

implying that Δ^C is decreasing in z, and consequently, it is increasing in σ_u^2.

After arranging terms, the derivative of Δ^C with respect to I in its compact factorized form is:

$$\frac{\partial \Delta^C}{\partial I} = \frac{-4(I+1)\left[(I-1)z-1\right]\left(I^2 z+z+1\right)}{(I-1)^3\left[(I+1)z+1\right]^3}, \quad \text{where } z \equiv \frac{1}{\varsigma \sigma_u^2 \xi}. \tag{IA.2.95}$$

When the inequality in (IA.2.89) holds, it follows that

$$\frac{\partial \Delta^C}{\partial I} < 0, \tag{IA.2.96}$$

implying that Δ^C is decreasing in I.

Proof for the price informativeness $\mathcal{I}^C$. When $\theta = 0$, by definition, the price informativeness $\mathcal{I}^C$ is

$$\mathcal{I}^C = \left(I\chi^C\right)^2 (\sigma_v/\sigma_u)^2 = \left(\frac{I\xi}{(I+1)+\varsigma\sigma_u^2\xi}\right)^2 \left(\frac{\sigma_v}{\sigma_u}\right)^2. \tag{IA.2.97}$$

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

$$\mathcal{I}^M = \left(I\chi^M\right)^2 (\sigma_v/\sigma_u)^2 = \left(\frac{\xi}{2}\right)^2 \left(\frac{\sigma_v}{\sigma_u}\right)^2. \tag{IA.2.98}$$

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

$$\frac{\mathcal{I}^C}{\mathcal{I}^M} = \left[\frac{(I-1)-\varsigma\sigma_u^2\xi}{(I+1)+\varsigma\sigma_u^2\xi}+1\right]^2. \tag{IA.2.99}$$

Clearly, $\mathcal{I}^C/\mathcal{I}^M$ is not affected by ρ at all, and it is decreasing in σ_u^2. Since the inequality in (IA.2.89) holds, it follows that $\frac{(I-1)-\varsigma\sigma_u^2\xi}{(I+1)+\varsigma\sigma_u^2\xi}$ is between 0 and 1, and thus it is increasing in I. As a result, $\mathcal{I}^C/\mathcal{I}^M$ is increasing in I.

Proof for the mispricing $\mathcal{E}^C$. When $\theta = 0$, by definition, the mispricing $\mathcal{E}^C$ is

$$\begin{aligned}
\mathcal{E}^C &= \left(1-\lambda^C I\chi^C\right)|v_t-\overline{v}| \\
&= \left[1-\xi^{-1}I\frac{1}{(I+1)\xi^{-1}+\varsigma\sigma_u^2}\right]|v_t-\overline{v}| \\
&= \frac{1+\varsigma\sigma_u^2\xi}{(I+1)+\varsigma\sigma_u^2\xi}\,|v_t-\overline{v}|\,.
\end{aligned} \tag{IA.2.100}$$

The mispricing in the perfect cartel benchmark, denoted by $\mathcal{E}^M$, is

$$\mathcal{E}^M = \left(1-\lambda^M I\chi^M\right)|v_t-\overline{v}| = \left(1-\xi^{-1}I\frac{\xi}{2I}\right)|v_t-\overline{v}| = |v_t-\overline{v}|/2. \tag{IA.2.101}$$

Thus, $\mathcal{E}^C/\mathcal{E}^M$ is equal to

$$\mathcal{E}^C/\mathcal{E}^M = 2\left[\frac{1+\varsigma\sigma_u^2\xi}{(I+1)+\varsigma\sigma_u^2\xi}\right]. \tag{IA.2.102}$$

Clearly, $\mathcal{E}^C/\mathcal{E}^M$ is not affected by ρ at all, and it is increasing in σ_u^2, because $\frac{1+\varsigma\sigma_u^2\xi}{(I+1)+\varsigma\sigma_u^2\xi}$ is between 0 and 1. It is straightforward to see that $\mathcal{E}^C/\mathcal{E}^M$ is decreasing in I.

Proof for the market liquidity $\mathcal{L}^C$. By definition, the market liquidity $\mathcal{L}^C$ is

$$\mathcal{L}^C = \frac{1}{\partial|z_t+y_t|/\partial u_t} = \frac{1}{1-\xi\lambda^C}. \tag{IA.2.103}$$

Because $\theta/\xi \approx 0$, the market liquidity $\mathcal{L}^C$ can be expressed as follows:

$$\mathcal{L}^C = \frac{1}{\left|1-\xi\frac{\theta\gamma^C+\xi}{\theta+\xi^2}\right|} \approx \frac{1}{\left|1-\xi\frac{\theta\gamma^C+\xi}{\xi^2}\right|} = \frac{\xi}{\theta\gamma^C} = \frac{\xi}{\theta}\left(I\chi^C+\frac{\sigma_u^2}{\sigma_v^2}\frac{1}{I\chi^C}\right). \tag{IA.2.104}$$

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

$$\mathcal{L}^M = \frac{\xi}{\theta}\left(I\chi^M+\frac{\sigma_u^2}{\sigma_v^2}\frac{1}{I\chi^M}\right) = \frac{\xi}{\theta}\left(I\frac{\xi}{2I}+\frac{\sigma_u^2}{\sigma_v^2}\frac{1}{I\frac{\xi}{2I}}\right) = \frac{1}{\theta}\left(\frac{\xi^2}{2}+\frac{2\sigma_u^2}{\sigma_v^2}\right). \tag{IA.2.105}$$

Thus, the relative market liquidity $\mathcal{L}^C/\mathcal{L}^M$ can be expressed as follows:

$$\frac{\mathcal{L}^C}{\mathcal{L}^M} = \frac{I\chi^C+\frac{\sigma_u^2}{\sigma_v^2}\frac{1}{I\chi^C}}{\frac{\xi}{2}+\frac{\sigma_u^2}{\sigma_v^2}\frac{2}{\xi}}. \tag{IA.2.106}$$

Given that $\varsigma\sigma_u^2\xi < I-1$, it follows that when $\theta/\xi \approx 0$ and σ_u/σ_v is sufficiently small, $\mathcal{L}^C/\mathcal{L}^M$ can be further simplified as:

$$\frac{\mathcal{L}^C}{\mathcal{L}^M} \approx 2\xi^{-1}I\chi^C = \frac{2I}{(I+1)+\varsigma\sigma_u^2\xi}. \tag{IA.2.107}$$

Thus, it is straightforward to see that $\mathcal{L}^C/\mathcal{L}^M$ is increasing in I and decreasing in σ_u^2. Clearly, $\mathcal{L}^C/\mathcal{L}^M$ is independent of ρ.

3 Heuristic Justifications and Intuitions for AI Equilibria

In this subsection, we present intuitions and heuristic proofs on how interactions among Q-learning algorithms may or may not lead to a collusive trading equilibrium, sustained by one of the two distinct mechanisms examined in Section 5, in various trading environments. A heuristic proof refers to an argument or reasoning that offers an intuitive explanation for a result, often relying on approximations or informal reasoning rather than strict mathematical rigor.

To capture the key ideas, we assume each informed speculator selects between two strategies: the non-collusive Nash equilibrium strategy, χ^N, and a more conservative alternative, χ^C. Specifically, our discussion on intuitions and heuristic proofs is based on a simple model with $I = 2$ and $\mathbb{X} = \{\chi^N, \chi^C\}$. For simplicity, we also set $v_t \equiv \overline{v} + \sigma_v$, with σ_v normalized to 1 throughout this paper. These assumptions make Q-learning speculators' behavior semi-tractable analytically, allowing us to heuristically characterize the properties of the steady state. In this case, the steady state arises from interactions and learning among multiple Q-learning speculators, manifesting as either (i) an AI collusive equilibrium sustained by price-trigger strategies, (ii) an AI collusive equilibrium sustained by over-pruning bias in learning, or (iii) the non-collusive Nash equilibrium. We emphasize that, in general, deriving theoretical convergence results for multiple interacting Q-learning algorithms is extremely difficult, which makes it even more challenging to gain analytical insights into the properties of their steady states.

We first consider the case where ξ is not close to zero, relative to θ, with two sub-scenarios: one where σ_u is low and one where σ_u is high. In the first sub-scenario, with low σ_u, informed AI speculators using Q-learning algorithms can learn price-trigger strategies to collude. However, achieving an AI collusive equilibrium sustained by over-pruning bias in learning is impossible in this case. In the second sub-scenario, with high σ_u, informed AI speculators can reach and sustain an AI collusive equilibrium due to over-pruning bias in learning, but they cannot learn to achieve an AI collusive equilibrium sustained by price-trigger strategies.

Next, we examine the case where ξ is close to zero relative to θ, regardless of the level of σ_u. In this scenario, informed AI speculators using Q-learning algorithms can reach and sustain an AI collusive equilibrium, where the conservative trading strategy χ^C is chosen due to over-pruning bias in learning, regardless of the level of noise trading risk σ_u. However, they cannot sustain a collusive equilibrium with χ^C chosen due to price-trigger strategies, irrespective of σ_u.

To elaborate further, the so-called "AI collusive equilibrium" is fundamentally connected to the theoretical concept of collusive equilibrium formalized in Definition 3.1 in Section 3.2 and Definitions 3.2 and 3.3 in Section 3.3 of the main text. However, these AI equilibria are reached and sustained through algorithmic mechanisms rather than theoretical assumptions or economic reasoning.

3.1 Significant Presence of Information-Insensitive Investors (i.e., Large ξ)

We consider the case where ξ is not close to zero, relative to θ, indicating a significant presence of information-insensitive investors, and thus a negligible effect of pricing error minimization in the market maker's objective, as specified in (IA.2.5). For transparency, we assume without loss of generality that $\theta = 0$. Thus,

$$\chi^C < \chi^N \equiv \xi/3 \quad \text{and} \quad \chi^C > \chi^M \equiv \xi/4. \tag{IA.3.1}$$

Let the focal speculator be i, and the other speculator is denoted by $-i$. At t, the market price, p_t, is

$$p_t = \begin{cases} p_t^N \equiv \overline{v} + 2\xi^{-1}\chi^N + \xi^{-1}u_t, & \text{if } (\chi_i, \chi_{-i}) = (\chi^N, \chi^N), \\ p_t^C \equiv \overline{v} + 2\xi^{-1}\chi^C + \xi^{-1}u_t, & \text{if } (\chi_i, \chi_{-i}) = (\chi^C, \chi^C), \\ p_t^D \equiv \overline{v} + \xi^{-1}(\chi^N + \chi^C) + \xi^{-1}u_t, & \text{otherwise,} \end{cases} \tag{IA.3.2}$$

and the trading profit for speculator i is $\pi_{i,t} = \left[1 - \xi^{-1}(\chi_{i,t} + \chi_{-i,t})\right] \chi_{i,t} - \xi^{-1}\chi_{i,t}u_t$, where

$$\pi_{i,t} = \begin{cases} \pi^{NN} - \xi^{-1}\chi^N u_t, \text{with } \pi^{NN} = \left(1 - 2\xi^{-1}\chi^N\right)\chi^N, & \text{if } (\chi_i, \chi_{-i}) = (\chi^N, \chi^N), \\ \pi^{CC} - \xi^{-1}\chi^C u_t, \text{with } \pi^{CC} = \left(1 - 2\xi^{-1}\chi^C\right)\chi^C, & \text{if } (\chi_i, \chi_{-i}) = (\chi^C, \chi^C), \\ \pi^{CN} - \xi^{-1}\chi^C u_t, \text{with } \pi^{CN} = \left[1 - \xi^{-1}(\chi^C + \chi^N)\right]\chi^C, & \text{if } (\chi_i, \chi_{-i}) = (\chi^C, \chi^N), \\ \pi^{NC} - \xi^{-1}\chi^N u_t, \text{with } \pi^{NC} = \left[1 - \xi^{-1}(\chi^C + \chi^N)\right]\chi^N, & \text{if } (\chi_i, \chi_{-i}) = (\chi^N, \chi^C). \end{cases} \tag{IA.3.3}$$

It is straightforward to show that $p_t^C < p_t^D < p_t^N$. The following inequalities hold

$$\begin{aligned} &\pi^{CN} < \pi^{NN} < \pi^{CC} < \pi^{NC}, \text{ and} \\ &\pi^{CN} + \pi^{CC} < \pi^{NN} + \pi^{NC}. \end{aligned} \tag{IA.3.4}$$

The inequality (IA.3.4) can be shown as follows:

$$\pi^{NN} + \pi^{NC} - (\pi^{CN} + \pi^{CC}) = (\chi^N - \chi^C)\left[2 - 3\xi^{-1}(\chi^N + \chi^C)\right] \tag{IA.3.5}$$

$$> (\chi^N - \chi^C)\left[2 - 3\xi^{-1}(\chi^N + \chi^N)\right] = 0. \tag{IA.3.6}$$

The simplified model, with a significant presence of information-insensitive investors, essentially resembles a prisoner's dilemma game in product market competition, where a "nature" player submits demand with an amount randomly drawn from a zero-mean normal distribution.

In the presence of a significant force of information-insensitive investors, that is, when ξ is not close to zero, relative to θ, we first demonstrate in Section 3.1.1 how an AI collusive equilibrium driven by price-trigger strategies can robustly arise in a low-noise trading environment. Conversely, we explain why an AI collusive equilibrium sustained by over-pruning bias in learning cannot arise in the same trading environment. In Section 3.1.2, we then show how an AI collusive equilibrium sustained by over-pruning bias in learning can arise in a high-noise trading environment, while no collusive equilibrium driven by price-trigger strategies can occur in the same trading environment. These findings confirm the results of our simulations, particularly those illustrated in Figure 2 of the main text.

3.1.1 Low Noise Trading Risk σ_u with Large ξ

We first consider the case where σ_u is low. For illustrative purposes, we assume $\sigma_u = 0$, meaning monitoring is perfect. Although rigorous mathematical proofs for the convergence of multiple

Q-learning algorithms in this simplified model have been demonstrated (e.g., Cartea et al., 2022; Possnig, 2024), they offer little in terms of intuitive explanation or insight, relying mainly on the application of stochastic approximation results. Specifically, these proofs for the simplified model depend on directly applying existing results in stochastic approximation, which primarily involves verifying high-level regularity conditions and emphasizes technical details. As a result, they lack intuitive insights. Our heuristic proofs add significant value by providing clear intuition behind the convergence of multiple Q-learning algorithms, which can be extended to more general settings, as outlined in our main model, even though we use this simplified model for illustration.

In this scenario, the market price in (IA.3.2) becomes

$$p_t = \begin{cases} p^N \equiv \overline{v} + 2\xi^{-1}\chi^N, & \text{if } (\chi_i, \chi_{-i}) = (\chi^N, \chi^N), \\ p^C \equiv \overline{v} + 2\xi^{-1}\chi^C, & \text{if } (\chi_i, \chi_{-i}) = (\chi^C, \chi^C), \\ p_t^D \equiv \overline{v} + \xi^{-1}(\chi^N + \chi^C), & \text{otherwise,} \end{cases} \tag{IA.3.7}$$

and the trading profit in (IA.3.3) becomes

$$\pi_{i,t} = \begin{cases} \pi^{NN} = \left(1 - 2\xi^{-1}\chi^N\right)\chi^N, & \text{if } (\chi_i, \chi_{-i}) = (\chi^N, \chi^N), \\ \pi^{CC} = \left(1 - 2\xi^{-1}\chi^C\right)\chi^C, & \text{if } (\chi_i, \chi_{-i}) = (\chi^C, \chi^C), \\ \pi^{CN} = \left[1 - \xi^{-1}(\chi^C + \chi^N)\right]\chi^C, & \text{if } (\chi_i, \chi_{-i}) = (\chi^C, \chi^N), \\ \pi^{NC} = \left[1 - \xi^{-1}(\chi^C + \chi^N)\right]\chi^N, & \text{if } (\chi_i, \chi_{-i}) = (\chi^N, \chi^C). \end{cases} \tag{IA.3.8}$$

And, in this scenario, the state variable s_t can take one of three values: $\{p^C, p^N, p^D\}$. To elaborate, based on the state variable s_t specification in Section 4.1, we define $s_t = p^C$ if and only if $\chi_{1,t-1} = \chi_{2,t-1} = \chi^C$, $s_t = p^N$ if and only if $\chi_{1,t-1} = \chi_{2,t-1} = \chi^N$, and $s_t = p^D$ in all other cases.

According to the theoretical recursive relation of the Q-function in (2.3), the theoretical Q-function associated with the collusive equilibrium sustained by the price-trigger strategy under perfect monitoring with perpetual punishment, can be summarized as follows:

If the state is $s = p^C$,

$$Q_i(s, \chi^C) = \frac{\pi^{CC}}{1-\rho}, \text{ and } Q_i(s, \chi^N) = \pi^{NC} + \frac{\rho\pi^{NN}}{1-\rho}, \quad \text{for } i = 1, 2; \tag{IA.3.9}$$

if the state is $s \neq p^C$,

$$Q_i(s, \chi^C) = \pi^{CN} + \frac{\rho\pi^{NN}}{1-\rho}, \text{ and } Q_i(s, \chi^N) = \frac{\pi^{NN}}{1-\rho}, \quad \text{for } i = 1, 2. \tag{IA.3.10}$$

Result 1: *In a market environment where $\sigma_u = 0$ and ξ is large relative to θ, no AI collusive equilibrium sustained by over-pruning bias in learning can be achieved by multiple informed AI speculators using Q-learning algorithms.*

Below, we provide a heuristic justification for **Result 1**. Intuitively, in a low-noise trading environment, the exploration-exploitation tradeoff works effectively, avoiding severe over-pruning bias in learning. As a result, AI collusive equilibrium sustained by over-pruning bias in learning cannot be achieved or sustained by multiple informed AI speculators.

In theory, this collusive experience-based equilibrium represents a steady state where informed speculators have learning bias that leads to an overvaluation of conservative trading strategies. In this equilibrium, $Q_i(s, \chi^N) < Q_i(s, \chi^C)$ for all $i = 1, 2$ and all s, indicating that the conservative strategy χ^C is consistently preferred over the aggressive strategy χ^N, and $Q_i(s, \chi^C)$ represents an on-path evaluation that satisfies the consistency condition of an experience-based equilibrium. As a result, the following holds:

$$Q_i(s, \chi^C) = \frac{\pi^{CC}}{1-\rho}, \text{ for } i = 1, 2 \text{ and all } s, \tag{IA.3.11}$$

and $Q_i(s, \chi^N)$ can take arbitrary values as long as:

$$Q_i(s, \chi^N) < \frac{\pi^{CC}}{1-\rho} \text{ for } i = 1, 2 \text{ and all } s. \tag{IA.3.12}$$

Specifically, we prove **Result 1** by contradiction. Assume that the system of multiple Q-learning algorithms converges to the collusive equilibrium sustained by learning bias, described above in (IA.3.11) and (IA.3.12). In this case, an AI collusive equilibrium sustained by over-pruning bias in learning would be achievable.

Thus, after many iterations in the exploitation-intensive stage, the estimated Q-function $\widehat{Q}_{i,t}$ approaches its theoretical counterpart Q_i, so that $\widehat{Q}_{i,t}(s, \chi^N) < \widehat{Q}_i(s, \chi^C)$ for $i = 1, 2$ and all s, for sufficiently large t.

What would $\widehat{Q}_{i,t}(s, \chi^N)$ and $\widehat{Q}_i(s, \chi^C)$ converge to in this scenario? During the exploitation-intensive stage, the following recursive relationship holds:

$$\widehat{Q}_{i,t_{\tau+1}}(p^C, \chi^C) = (1-\alpha)\widehat{Q}_{i,t_\tau}(p^C, \chi^C) + \alpha \left[\pi^{CC} + \rho \max_{\chi \in \{\chi^C, \chi^N\}} \widehat{Q}_{i,t_\tau}(p^C, \chi) \right] \tag{IA.3.13}$$

where t_τ is the index of the iteration in which i visits the state-and-action pair (p^C, χ^C) for the τ-th time.

For sufficiently large τ, given that χ^C is preferred over χ^N, the maximum value simplifies to $\widehat{Q}_{i,t_\tau}(p^C, \chi^C)$, and we have:

$$\widehat{Q}_{i,t_{\tau+1}}(p^C, \chi^C) = (1-\alpha+\alpha\rho)\widehat{Q}_{i,t_\tau}(p^C, \chi^C) + \alpha\pi^{CC}. \tag{IA.3.14}$$

Thus, as τ grows large, $\widehat{Q}_{i,t_\tau}(p^C, \chi^C)$ approaches $\dfrac{\pi^{CC}}{1-\rho}$ for $i = 1, 2$, described in (IA.3.11). Thus,

$\widehat{Q}_{i,t}(p^C, \chi^C)$ approaches $\frac{\pi^{CC}}{1-\rho}$, as t approaches infinity. Similarly, we can show that $\widehat{Q}_{i,t}(s, \chi^C)$ approaches $\frac{\pi^{CC}}{1-\rho}$, as t approaches infinity, for $i = 1, 2$ and all other s. During the exploitation-intensive stage, the following recursive relationship holds:

$$\widehat{Q}_{i,t_{\tau+1}}(p^C, \chi^N) = (1-\alpha)\widehat{Q}_{i,t_\tau}(p^C, \chi^N) + \alpha \left[\pi^{NC} + \rho \max_{\chi \in \{\chi^C, \chi^N\}} \widehat{Q}_{i,t_\tau}(p^D, \chi) \right] \tag{IA.3.15}$$

where t_τ is the index of the iteration in which i visits the state-and-action pair (p^C, χ^N) for the τ-th time.

For sufficiently large τ, given that χ^C is preferred over χ^N, the maximum value simplifies to $\widehat{Q}_{i,t_\tau}(p^D, \chi^C) \approx \frac{\rho \pi^{CC}}{1-\rho}$, and we have:

$$\widehat{Q}_{i,t_{\tau+1}}(p^C, \chi^N) = (1-\alpha)\widehat{Q}_{i,t_\tau}(p^C, \chi^N) + \alpha \left[\pi^{NC} + \rho \widehat{Q}_{i,t_\tau}(p^D, \chi^C) \right] \tag{IA.3.16}$$

$$\approx (1-\alpha)\widehat{Q}_{i,t_\tau}(p^C, \chi^N) + \alpha \left(\pi^{NC} + \frac{\rho \pi^{CC}}{1-\rho} \right). \tag{IA.3.17}$$

Thus, as τ grows large, $\widehat{Q}_{i,t_\tau}(p^C, \chi^N)$ approaches $\pi^{NC} + \frac{\rho \pi^{CC}}{1-\rho}$ for $i = 1, 2$. As a result, $\widehat{Q}_{i,t}(p^C, \chi^N)$ converges to the same value, $\pi^{NC} + \frac{\rho \pi^{CC}}{1-\rho}$, for $i = 1, 2$, as t approaches infinity.

Since $\pi^{NC} > \pi^{CC}$, as stated in (IA.3.4), it follows that $\widehat{Q}_{i,t}(p^C, \chi^N) > \widehat{Q}_i(p^C, \chi^C)$ for both $i = 1, 2$. This contradicts the assumption that the system of multiple Q-learning algorithms converges to the AI collusive equilibrium sustained by over-pruning bias in learning. Therefore, by contradiction, we conclude that the collusive equilibrium sustained by over-pruning bias in learning cannot be achieved by multiple Q-learning algorithms when the noise trading risk σ_u is low.

Result 2: *In a market environment where $\sigma_u = 0$ and ξ is large relative to θ, an AI collusive equilibrium sustained by price-trigger strategies can be achieved by multiple informed AI speculators using Q-learning algorithms with a sufficiently small forgetting rate α.*

Below, we provide a heuristic proof and key intuitions for **Result 2**. By the nature of reinforcement learning algorithms, both in the simple ε-greedy method in (4.3) of the main text and in more general exploitation-exploration trade-off specifications, random exploration dominates the early iterations, referred to as the "exploration-intensive stage" (**Phase 1**). As the exploration rate ε_t approaches zero, exploitation becomes dominant in the later iterations, marking the "exploitation-intensive stage" (**Phase 2**). After sufficient iterations in these first two stages, the system of estimated Q-functions $\widehat{Q}_{i,t}(s, \chi)$ for $i = 1, 2$ stabilizes within a small neighborhood of the steady state. This is referred to as the "absorbing stage" (**Phase 3**).

In line with the three phases described above, we illustrate the intuition behind the convergence process through the following three steps, consistent with the nature of reinforcement learning algorithms:

Phase 1 ("The Exploration-Intensive Stage"): We provide a heuristic proof that, after a sufficient number of iterations in the exploration-intensive stage, it holds that $\widehat{Q}_{i,t}(s,\chi^C) < \widehat{Q}_{i,t}(s,\chi^N)$ for all s, as shown below.

When exploration dominates and exploitation is minimal, that is, $\varepsilon_t \approx 1$, the two actions χ^C and χ^N are randomly chosen with equal probabilities in each iteration, regardless of the state variable s. As a result, the Q-learning algorithms behave as if they lack dynamic sophistication, not tracking any state variables. Consequently, the Q-functions do not depend on the state variable s after many iterations in the exploration-intensive stage. As a result, in this stage, when t is sufficiently large, for any state s, it holds that

$$\widehat{Q}_{i,t}(s,\chi^C) < \widehat{Q}_{i,t}(s,\chi^N), \quad \text{for any state } s, \tag{IA.3.18}$$

and specifically, it holds that

$$\widehat{Q}_{i,t}(s,\chi^C) \approx \frac{\pi^{CC}+\pi^{CN}}{2} + \frac{\rho(\pi^{NC}+\pi^{NN})}{2(1-\rho)}, \text{ and} \tag{IA.3.19}$$

$$\widehat{Q}_{i,t}(s,\chi^N) \approx \frac{\pi^{NC}+\pi^{NN}}{2(1-\rho)}. \tag{IA.3.20}$$

Below, we prove (IA.3.18) by contradiction. Suppose $\widehat{Q}_{i,t}(s,\chi^C) \geq \widehat{Q}_{i,t}(s,\chi^N)$, for any state s. As a result, the following cumulative updates hold:

$$\widehat{Q}_{i,t_\tau}(s,\chi^C) = \sum_{h=0}^{\tau-1} \alpha(1-\alpha)^h \left[D_h\pi^{CC} + (1-D_h)\pi^{CN} + \rho \max_{\chi\in\{\chi^C,\chi^N\}} \widehat{Q}_{i,t_{\tau-h}}(s,\chi)\right], \tag{IA.3.21}$$

$$= \sum_{h=0}^{\tau-1} \alpha(1-\alpha)^h \left[D_h\pi^{CC} + (1-D_h)\pi^{CN} + \rho\widehat{Q}_{i,t_{\tau-h}}(s,\chi^C)\right], \tag{IA.3.22}$$

where t_τ is the index of the iteration in which i visits the state-and-action pair (s,χ^N) for the τ-th time, and D_h are i.i.d. Bernoulli variables with mean $1/2$, capturing random exploration. When τ is sufficiently large, there exists sufficiently large τ_0 such that the following conditions are satisfied: (i) $\tau_0 \ll \tau$, (ii) $(1-\alpha)^{\tau-\tau_0}$ is sufficiently tiny, and (iii) $\widehat{Q}_{i,t_{\tau'}}(s,\chi^C) \approx \widehat{Q}_{i,t_{\tau_0}}(s,\chi^C)$ for all $\tau' \geq \tau_0$ in the exploration-intensive stage, including iteration index τ. Therefore, the following approximation works:

$$\begin{aligned}\widehat{Q}_{i,t_\tau}(s,\chi^C) &= \sum_{h=0}^{\tau-\tau_0} \alpha(1-\alpha)^h \left[D_h\pi^{CC} + (1-D_h)\pi^{CN} + \rho\widehat{Q}_{i,t_{\tau-h}}(s,\chi^C)\right] \\ &\quad + \sum_{h=\tau-\tau_0+1}^{\tau-1} \alpha(1-\alpha)^h \left[D_h\pi^{CC} + (1-D_h)\pi^{CN} + \rho\widehat{Q}_{i,t_{\tau-h}}(s,\chi^C)\right] \qquad \text{(IA.3.23)} \\ &\approx \sum_{h=0}^{\tau-\tau_0} \alpha(1-\alpha)^h \left[D_h\pi^{CC} + (1-D_h)\pi^{CN} + \rho\widehat{Q}_{i,t_\tau}(s,\chi^C)\right] \qquad \text{(IA.3.24)} \\ &\approx \sum_{h=0}^{\infty} \alpha(1-\alpha)^h \left[D_h\pi^{CC} + (1-D_h)\pi^{CN}\right] + \rho\widehat{Q}_{i,t_\tau}(s,\chi^C). \qquad \text{(IA.3.25)}\end{aligned}$$

Thus, it follows that

$$\widehat{Q}_{i,t_\tau}(s,\chi^C) \approx \frac{1}{1-\rho}\sum_{h=0}^{\infty} \alpha(1-\alpha)^h \left[D_h\pi^{CC} + (1-D_h)\pi^{CN}\right]. \qquad \text{(IA.3.26)}$$

To understand how the Law of Large Numbers (LLN) approximation applies to the right-hand side of equation (IA.3.26), we examine the following two terms:

$$\begin{aligned}\mathbb{E}\left[\sum_{h=0}^{\infty} \alpha(1-\alpha)^h \left[D_h\pi^{CC} + (1-D_h)\pi^{CN}\right]\right] &= \frac{\pi^{CC}+\pi^{CN}}{2}, \qquad \text{(IA.3.27)} \\ \operatorname{var}\left[\sum_{h=0}^{\infty} \alpha(1-\alpha)^h \left[D_h\pi^{CC} + (1-D_h)\pi^{CN}\right]\right] &= \frac{\alpha}{4(2-\alpha)}\left(\pi^{CC}-\pi^{CN}\right)^2. \qquad \text{(IA.3.28)}\end{aligned}$$

Applying Markov's inequality, it follows from equations (IA.3.27) and (IA.3.28) that as α approaches zero, the following LLN approximation becomes increasingly accurate:

$$\sum_{h=0}^{\infty} \alpha(1-\alpha)^h \left[D_h\pi^{CC} + (1-D_h)\pi^{CN}\right] \approx \frac{\pi^{CC}+\pi^{CN}}{2}. \qquad \text{(IA.3.29)}$$

Since α is small, the Law of Large Numbers implies that, as t becomes sufficiently large,

$$\widehat{Q}_{i,t}(s,\chi^C) \approx \frac{1}{1-\rho}\frac{\pi^{CC}+\pi^{CN}}{2}. \qquad \text{(IA.3.30)}$$

In the meantime, we would have

$$\widehat{Q}_{i,t_\tau}(s,\chi^N) = \sum_{h=0}^{\tau-1} \alpha(1-\alpha)^h \left[D_h \pi^{NC} + (1-D_h)\pi^{NN} + \rho \max_{\chi \in \{\chi^C, \chi^N\}} \widehat{Q}_{i,t_{\tau-h}}(s,\chi) \right] \tag{IA.3.31}$$

$$= \sum_{h=0}^{\tau-1} \alpha(1-\alpha)^h \left[D_h \pi^{NC} + (1-D_h)\pi^{NN} + \rho \widehat{Q}_{i,t_{\tau-h}}(s,\chi^C) \right] \tag{IA.3.32}$$

$$= \sum_{h=0}^{\tau-\tau_0} \alpha(1-\alpha)^h \left[D_h \pi^{NC} + (1-D_h)\pi^{NN} + \rho \widehat{Q}_{i,t_{\tau-h}}(s,\chi^C) \right] + \sum_{h=\tau-\tau_0+1}^{\tau-1} \alpha(1-\alpha)^h \left[D_h \pi^{NC} + (1-D_h)\pi^{NN} + \rho \widehat{Q}_{i,t_{\tau-h}}(s,\chi^C) \right] \tag{IA.3.33}$$

$$\approx \sum_{h=0}^{\tau-\tau_0} \alpha(1-\alpha)^h \left[D_h \pi^{NC} + (1-D_h)\pi^{NN} + \rho \widehat{Q}_{i,t_\tau}(s,\chi^C) \right] \tag{IA.3.34}$$

$$\approx \sum_{h=0}^{\infty} \alpha(1-\alpha)^h \left[D_h \pi^{NC} + (1-D_h)\pi^{NN} \right] + \rho \widehat{Q}_{i,t_\tau}(s,\chi^C) \tag{IA.3.35}$$

$$\approx \sum_{h=0}^{\infty} \alpha(1-\alpha)^h \left[D_h \pi^{NC} + (1-D_h)\pi^{NN} \right] + \frac{\rho}{1-\rho} \frac{\pi^{CC} + \pi^{CN}}{2}. \tag{IA.3.36}$$

where t_τ is the index of the iteration in which i visits the state-and-action pair (s, χ^N) for the τ-th time, and D_h are i.i.d. Bernoulli variables with mean $1/2$, capturing random exploration.

As α is small, the Law of Large Numbers implies that, as t is sufficiently large,

$$\widehat{Q}_{i,t}(s,\chi^N) = \frac{\pi^{NC} + \pi^{NN}}{2} + \frac{\rho}{1-\rho} \frac{\pi^{CC} + \pi^{CN}}{2}. \tag{IA.3.37}$$

Comparing (IA.3.30) with (IA.3.37), the inequality (IA.3.4) results in a contraction. Therefore, it follows that the inequalities in (IA.3.18) must hold after numerous iterations, by the end of the exploration-intensive stage.

Given the inequalities in (IA.3.18), we now show that the results in (IA.3.19) and (IA.3.20) hold. The following relationship exists between the estimated Q-function and its cumulative updates from the past:

$$\widehat{Q}_{i,t_\tau}(s,\chi^N) = \sum_{h=0}^{\tau-1} \alpha(1-\alpha)^h \left[D_h \pi^{NC} + (1-D_h)\pi^{NN} + \rho \max_{\chi \in \{\chi^C, \chi^N\}} \widehat{Q}_{i,t_{\tau-h}}(s,\chi) \right], \tag{IA.3.38}$$

$$= \sum_{h=0}^{\tau-1} \alpha(1-\alpha)^h \left[D_h \pi^{NC} + (1-D_h)\pi^{NN} + \rho \widehat{Q}_{i,t_{\tau-h}}(s,\chi^N) \right], \tag{IA.3.39}$$

where t_τ is the index of the iteration in which i visits the state-and-action pair (s, χ^N) for the τ-th time, and D_h are i.i.d. Bernoulli variables with mean $1/2$, capturing random exploration. When τ is sufficiently large, there exists sufficiently large τ_0 such that the following conditions are satisfied: (i) $\tau_0 \ll \tau$, (ii) $(1-\alpha)^{\tau-\tau_0}$ is sufficiently tiny, and (iii) $\widehat{Q}_{i,t_{\tau'}}(s,\chi^N) \approx \widehat{Q}_{i,t_{\tau_0}}(s,\chi^N)$ for all $\tau' \geq \tau_0$, including iteration index τ. Thus, the following approximation works:

$$\begin{aligned}\widehat{Q}_{i,t_\tau}(s,\chi^N) = & \sum_{h=0}^{\tau-\tau_0} \alpha(1-\alpha)^h \left[D_h \pi^{NC} + (1-D_h)\pi^{NN} + \rho \widehat{Q}_{i,t_{\tau-h}}(s,\chi^N)\right] \\ & + \sum_{h=\tau-\tau_0+1}^{\tau} \alpha(1-\alpha)^h \left[D_h \pi^{NC} + (1-D_h)\pi^{NN} + \rho \widehat{Q}_{i,t_{\tau-h}}(s,\chi^N)\right]\end{aligned} \tag{IA.3.40}$$

$$\approx \sum_{h=0}^{\tau-\tau_0} \alpha(1-\alpha)^h \left[D_h \pi^{NC} + (1-D_h)\pi^{NN} + \rho \widehat{Q}_{i,t_\tau}(s,\chi^N)\right] \tag{IA.3.41}$$

$$\approx \sum_{h=0}^{\infty} \alpha(1-\alpha)^h \left[D_h \pi^{NC} + (1-D_h)\pi^{NN}\right] + \rho \widehat{Q}_{i,t_\tau}(s,\chi^N). \tag{IA.3.42}$$

Thus, it follows that

$$\widehat{Q}_{i,t_\tau}(s,\chi^N) \approx \frac{1}{1-\rho} \sum_{h=0}^{\infty} \alpha(1-\alpha)^h \left[D_h \pi^{NC} + (1-D_h)\pi^{NN}\right]. \tag{IA.3.43}$$

Since α is small, the Law of Large Numbers implies that, as t becomes sufficiently large,

$$\widehat{Q}_{i,t}(s,\chi^N) \approx \frac{1}{1-\rho} \frac{\pi^{NC} + \pi^{NN}}{2}. \tag{IA.3.44}$$

With the result in (IA.3.20) established, we now demonstrate that the result in (IA.3.19) also holds. For sufficiently large τ, the Q-function for trading strategy χ^C is

$$\widehat{Q}_{i,t_\tau}(s,\chi^C) = \sum_{h=0}^{\tau-1} \alpha(1-\alpha)^h \left[D_h \pi^{CC} + (1-D_h)\pi^{CN} + \rho \max_{\chi \in \{\chi^C,\chi^N\}} \widehat{Q}_{i,t_{\tau-h}}(s,\chi)\right] \tag{IA.3.45}$$

$$= \sum_{h=0}^{\tau-1} \alpha(1-\alpha)^h \left[D_h \pi^{CC} + (1-D_h)\pi^{CN} + \rho \widehat{Q}_{i,t_{\tau-h}}(s,\chi^N)\right] \tag{IA.3.46}$$

$$\approx \sum_{h=0}^{\infty} \alpha(1-\alpha)^h \left[D_h \pi^{CC} + (1-D_h)\pi^{CN}\right] + \rho \widehat{Q}_{i,t_\tau}(s,\chi^N) \tag{IA.3.47}$$

$$\approx \sum_{h=0}^{\infty} \alpha(1-\alpha)^h \left[D_h \pi^{CC} + (1-D_h)\pi^{CN}\right] + \frac{\rho}{1-\rho} \frac{\pi^{NC} + \pi^{NN}}{2}. \tag{IA.3.48}$$

where t_τ is the index of the iteration in which i visits the state-and-action pair (s,χ^C) for the τ-th time, and D_h are i.i.d. Bernoulli variables with mean $1/2$, capturing random exploration.

Since α is small, the Law of Large Numbers implies that, as t becomes sufficiently large,

$$\widehat{Q}_{i,t}(s,\chi^C) \approx \frac{\pi^{CC}+\pi^{CN}}{2} + \frac{\rho}{1-\rho}\frac{\pi^{NC}+\pi^{NN}}{2}. \tag{IA.3.49}$$

Phase 2 ("The Exploitation-Intensive Stage"): After many iterations in the exploration-intensive stage, **Phase 1** achieves convergence with the estimated Q-functions stabilizing around the limits in (IA.3.19) and (IA.3.20). We now show that as the algorithms transition to **Phase 2**, the estimated Q-function $\widehat{Q}_{i,t}$ converges toward the true Q-function Q_i for $i = 1, 2$, after many iterations of exploitation during the exploitation-intensive stage, where the exploration rate is near zero (i.e., $\varepsilon_t \approx 0$). Notably, this convergence is driven by the reinforcing nature of the algorithms.

We show that, in the exploitation-intensive stage with $\varepsilon_t \approx 0$, the estimated Q-functions converge to their true counterparts, specified in equations (IA.3.9) and (IA.3.10), after numerous iterations of exploitation. Below, we first consider the four cases with $s_t \neq p^C$:

- When the state is $s_t = p^N$ and $\chi_{i,t} = \chi^N$, the other speculator, $-i$, will also choose strategy χ^N based on exploitation, as it was favored over χ^C after numerous explorations in **Phase 1**. As a result, the state in the next iteration remains $s_{t+1} = p^N$, leading to the following update:

$$\widehat{Q}_{i,t_{\tau+1}}(p^N,\chi^N) = (1-\alpha)\widehat{Q}_{i,t_\tau}(p^N,\chi^N) + \alpha\left[\pi^{NN} + \rho \max_{\chi\in\{\chi^C,\chi^N\}} \widehat{Q}_{i,t_\tau}(p^N,\chi)\right] \tag{IA.3.50}$$

$$= (1-\alpha+\alpha\rho)\widehat{Q}_{i,t_\tau}(p^N,\chi^N) + \alpha\pi^{NN}, \tag{IA.3.51}$$

  where t_τ is the index of the iteration in which i visits the state-and-action pair (p^N,χ^N). As τ increases, $\widehat{Q}_{i,t_\tau}(p^N,\chi^N)$ converges to $Q_i(p^N,\chi^N) \equiv \frac{\pi^{NN}}{1-\rho}$ for $i = 1, 2$, as described in (IA.3.10). Therefore, $\widehat{Q}_{i,t}(p^N,\chi^N)$ approaches $\frac{\pi^{NN}}{1-\rho}$ for both $i = 1, 2$ as t approaches infinity.

- When the state is $s_t = p^D$ and $\chi_{i,t} = \chi^N$, the other speculator, $-i$, following the rule of exploitation, will also choose strategy χ^N, as it was favored over χ^C after numerous explorations in **Phase 1**. Consequently, the state in the next iteration becomes $s_{t+1} = p^N$, leading to the following update:

$$\widehat{Q}_{i,t_{\tau+1}}(p^D,\chi^N) = (1-\alpha)\widehat{Q}_{i,t_\tau}(p^D,\chi^N) + \alpha\left[\pi^{NN} + \rho \max_{\chi\in\{\chi^C,\chi^N\}} \widehat{Q}_{i,t_\tau}(p^N,\chi)\right] \quad \text{(IA.3.52)}$$

$$= (1-\alpha)\widehat{Q}_{i,t_\tau}(p^D,\chi^N) + \alpha\left[\pi^{NN} + \rho\widehat{Q}_{i,t_\tau}(p^N,\chi^N)\right] \quad \text{(IA.3.53)}$$

$$\approx (1-\alpha)\widehat{Q}_{i,t_\tau}(p^D,\chi^N) + \alpha\left(\pi^{NN} + \rho\frac{\pi^{NN}}{1-\rho}\right), \quad \text{(IA.3.54)}$$

where t_τ is the index of the iteration in which i visits the state-and-action pair (p^D,χ^N). As τ grows large, $\widehat{Q}_{i,t_\tau}(p^D,\chi^N)$ approaches $Q_i(p^D,\chi^N) \equiv \frac{\pi^{NN}}{1-\rho}$ for $i=1,2$, as described in (IA.3.10). Thus, $\widehat{Q}_{i,t}(p^D,\chi^N)$ approaches $\frac{\pi^{NN}}{1-\rho}$ for $i=1,2$, as t approaches infinity.

- When the state is $s_t = p^N$ and $\chi_{i,t} = \chi^C$, the other speculator, $-i$, following the rule of exploitation, will continue to choose strategy χ^N, as it was favored over χ^C after numerous explorations in **Phase 1**. Consequently, the state in the next iteration becomes $s_{t+1} = p^D$, leading to the following update:

$$\widehat{Q}_{i,t_{\tau+1}}(p^N,\chi^C) = (1-\alpha)\widehat{Q}_{i,t_\tau}(p^N,\chi^C) + \alpha\left[\pi^{CN} + \rho \max_{\chi\in\{\chi^C,\chi^N\}} \widehat{Q}_{i,t_\tau}(p^D,\chi)\right] \quad \text{(IA.3.55)}$$

$$= (1-\alpha)\widehat{Q}_{i,t_\tau}(p^N,\chi^C) + \alpha\left[\pi^{CN} + \rho\widehat{Q}_{i,t_\tau}(p^D,\chi^N)\right] \quad \text{(IA.3.56)}$$

$$\approx (1-\alpha)\widehat{Q}_{i,t_\tau}(p^N,\chi^C) + \alpha\left(\pi^{CN} + \rho\frac{\pi^{NN}}{1-\rho}\right), \quad \text{(IA.3.57)}$$

where t_τ is the index of the iteration in which i visits the state-and-action pair (p^N,χ^C). As τ grows large, $\widehat{Q}_{i,t_\tau}(p^N,\chi^C)$ approaches $Q_i(p^N,\chi^C) \equiv \pi^{CN} + \frac{\rho\pi^{NN}}{1-\rho}$ for $i=1,2$, as described in (IA.3.10). Thus, $\widehat{Q}_{i,t}(p^N,\chi^C)$ approaches $\pi^{CN} + \frac{\rho\pi^{NN}}{1-\rho}$ for $i=1,2$, as t approaches infinity.

- When the state is $s_t = p^D$ and $\chi_{i,t} = \chi^C$, the other speculator $-i$, following the rule of exploitation, will choose strategy χ^N, as it was favored over χ^C after numerous explorations in **Phase 1**. Consequently, the state in the next iteration becomes $s_{t+1} = p^D$, leading to the following update:

$$\widehat{Q}_{i,t_{\tau+1}}(p^D,\chi^C) = (1-\alpha)\widehat{Q}_{i,t_\tau}(p^D,\chi^C) + \alpha\left[\pi^{CN} + \rho \max_{\chi\in\{\chi^C,\chi^N\}} \widehat{Q}_{i,t_\tau}(p^D,\chi)\right] \quad \text{(IA.3.58)}$$

$$= (1-\alpha)\widehat{Q}_{i,t_\tau}(p^D,\chi^C) + \alpha\left[\pi^{CN} + \rho\widehat{Q}_{i,t_\tau}(p^D,\chi^N)\right] \quad \text{(IA.3.59)}$$

$$\approx (1-\alpha)\widehat{Q}_{i,t_\tau}(p^D,\chi^C) + \alpha\left(\pi^{CN} + \rho\frac{\pi^{NN}}{1-\rho}\right), \quad \text{(IA.3.60)}$$

where t_τ is the index of the iteration in which i visits the state-and-action pair (p^D, χ^C). As τ grows large, $\widehat{Q}_{i,t_\tau}(p^D, \chi^C)$ approaches $Q_i(p^D, \chi^C) \equiv \pi^{CN} + \frac{\rho\pi^{NN}}{1-\rho}$ for $i = 1, 2$, as described in (IA.3.10). Thus, $\widehat{Q}_{i,t}(p^D, \chi^C)$ approaches $\pi^{CN} + \frac{\rho\pi^{NN}}{1-\rho}$ for $i = 1, 2$, as t approaches infinity.

Now, we consider the case where the state is $s_t = p^C$. We first show that there must exist some t^C such that $\widehat{Q}_{i,t^C}(p^C, \chi^C) > \widehat{Q}_{i,t^C}(p^C, \chi^N)$ for both $i = 1, 2$. We then show that, if this condition holds, for any $t > t^C$, the state p^C and the associated optimal strategy χ^C will continue to be reinforced during the exploitation-intensive stage (**Phase 2**). This reinforcement enables the estimated Q-functions at state p^C for both actions χ^C and χ^N, denoted by $\widehat{Q}_{i,t}(p^C, \chi^C)$ and $\widehat{Q}_{i,t}(p^C, \chi^N)$, to converge toward the theoretical values $Q_i(p^C, \chi^C)$ and $Q_i(p^C, \chi^N)$, respectively, defined in (IA.3.9), as t approaches infinity. Combining the convergence results for Q-functions evaluated at state p^C (derived below) with those at p^D and p^N (derived above in (IA.3.50) to (IA.3.60)), we find that informed AI speculators using Q-learning algorithms reach and maintain an AI collusive equilibrium, sustained by price-trigger strategies. This occurs because χ^C is the optimal strategy at state p^C, which keeps the system in p^C. If one informed AI speculator deviates from χ^C to χ^N, the state shifts from p^C to p^D, where χ^N becomes the optimal strategy, eventually leading the system to state p^N in subsequent periods.

– We first show that there must exist some t^C such that $\widehat{Q}_{i,t^C}(p^C, \chi^C) > \widehat{Q}_{i,t^C}(p^C, \chi^N)$ for both $i = 1, 2$. As the system of Q-learning algorithms transitions from the exploration-intensive stage (**Phase 1**) to **Phase 2**, both speculators have $\widehat{Q}_{i,t}(p^C, \chi^C) < \widehat{Q}_{i,t}(p^C, \chi^N)$ at every t in the beginning of the exploitation-intensive stage (**Phase 2**). Thus, for each $i = 1, 2$, we have:

$$\widehat{Q}_{i,t_{\tau+1}}(p^C, \chi^N) = (1-\alpha)\widehat{Q}_{i,t_\tau}(p^C, \chi^N) + \alpha \left[\pi^{NN} + \rho \max_{\chi \in \{\chi^C, \chi^N\}} \widehat{Q}_{i,t_\tau}(p^N, \chi)\right] \tag{IA.3.61}$$

$$= (1-\alpha)\widehat{Q}_{i,t_\tau}(p^C, \chi^N) + \alpha \left[\pi^{NN} + \rho \widehat{Q}_{i,t_\tau}(p^N, \chi^N)\right] \tag{IA.3.62}$$

$$\approx (1-\alpha)\widehat{Q}_{i,t_\tau}(p^C, \chi^N) + \alpha \left(\pi^{NN} + \rho\frac{\pi^{NN}}{1-\rho}\right), \tag{IA.3.63}$$

where t_τ is the index of the iteration in which speculator i visits the state-and-action pair (p^C, χ^N) for the τ-th time. In this case, (p^C, χ^C) would be insufficiently, even rarely, updated from the beginning of the exploitation-intensive stage (**Phase 2**) for both $i = 1, 2$. Consequently, $\widehat{Q}_{i,t}(p^C, \chi^C)$ would be remain close to its initial value at the beginning of the exploitation-intensive stage (**Phase 2**), as given by equation (IA.3.49), which implies that

$$\widehat{Q}_{i,t}(p^C, \chi^C) \approx \frac{\pi^{CC} + \pi^{CN}}{2} + \frac{\rho(\pi^{NC} + \pi^{NN})}{2(1-\rho)}, \tag{IA.3.64}$$

which is strictly greater than $\widehat{Q}_{i,t}(p^C, \chi^N) \approx \frac{\pi^{NN}}{1-\rho}$, derived from (IA.3.63), for sufficiently large t. This is true because $\pi^{NC} > \pi^{NN}$ and

$$\begin{aligned}
\pi^{CC} + \pi^{CN} - 2\pi^{NN} &= \left(1 - 2\xi^{-1}\chi^C\right)\chi^C + \left[1 - \xi^{-1}(\chi^C + \chi^N)\right]\chi^C - 2\left(1 - 2\xi^{-1}\chi^N\right)\chi^N \\
&= (\chi^N - \chi^C)\left(4\xi^{-1}\chi^N + 3\xi^{-1}\chi^C - 2\right) && \text{(IA.3.65)} \\
&\geq (\chi^N - \chi^C)\left(4\xi^{-1}\chi^N + 3\xi^{-1}\chi^M - 2\right) \quad \text{(due to (IA.3.1))} && \text{(IA.3.66)} \\
&= (\chi^N - \chi^C)\left(\frac{4}{3} + \frac{3}{4} - 2\right) > 0. && \text{(IA.3.67)}
\end{aligned}$$

Therefore, after many iterations from the starting point of the exploitation-intensive stage (**Phase 2**), the estimated Q-value $\widehat{Q}_{i,t}(p^C, \chi^N)$ will become lower than $\widehat{Q}_{i,t}(p^C, \chi^C)$. This shift in Q-values will lead both speculators to increasingly prefer χ^C over χ^N at the state p^C. The exploitation iterations will further reinforce this preference, as illustrated by the following process.

- We next show that, if there exists t^C such that $\widehat{Q}_{i,t^C}(p^C, \chi^C) > \widehat{Q}_{i,t^C}(p^C, \chi^N)$ for both $i = 1, 2$, the state p^C and the associated optimal strategy χ^C will continue to be reinforced during the exploitation-intensive stage (**Phase 2**) for any $t > t^C$. Consequently, the system of Q-learning algorithms converges to a collusive equilibrium sustained by price-trigger strategies, as t approaches infinity.

Because Q-functions and profits are bounded, once the system reaches t^C, where

$$\widehat{Q}_{i,t^C}(p^C, \chi^C) > \widehat{Q}_{i,t^C}(p^C, \chi^N), \quad \text{for both } i = 1, 2, \tag{IA.3.68}$$

a substantial number of updates at state p^C would be required to reverse any of these inequalities, if reversal is possible at all, particularly with a sufficiently small forgetting rate α.[3]

To elaborate further, given that Q-functions and profits are bounded, they are unlikely to

[3] From this point, we observe that a small α plays two critical roles in the convergence of multiple Q-learning algorithms. First, it ensures that the Law of Large Numbers holds as a reasonable approximation during the exploration-intensive stage (**Phase 1**), allowing the algorithms to accurately learn the ranking of all strategies without considering collusive equilibria based on past behavior, as previously emphasized. Second, it facilitates convergence to a collusive equilibrium sustained by price-trigger strategies in the exploitation-intensive stage (**Phase 2**), where the focus shifts to exploiting the strategies that have been learned and correcting over-estimated Q-values of the aggressive trading strategy χ^N in the state p^C.

change much in the short run. Once χ^C becomes the better choice in state p^C, that preference tends to persist and gets reinforced over time. Reversing the inequality (IA.3.68), i.e., reaching a point t where $\widehat{Q}_{i,t}(p^C, \chi^N)$ might become greater than $\widehat{Q}_{i,t}(p^C, \chi^C)$ again, would require an exceptionally large number of updates to the Q-functions at state p^C. This is especially true when the forgetting rate, α, is sufficiently small. A small α restricts the rate at which recent experiences influence the Q-values, thereby making Q-values slow moving and insensitive to individual updates. Consequently, once the inequality in favor of χ^C is established, it is likely to remain for a prolonged period, or potentially indefinitely, as the system continues to reinforce the choice of χ^C at state p^C.

In fact, before any possible reversal of the inequalities in (IA.3.68) could happen, the system would already be effectively "locked" in the state p^C, with both speculators choosing strategy χ^C as the optimal strategy due to the following Q-learning recursive relationship:

$$\begin{aligned}
\widehat{Q}_{i,t_{\tau+1}}(p^C, \chi^C) &= (1-\alpha)\widehat{Q}_{i,t_\tau}(p^C, \chi^C) + \alpha\left[\pi^{CC} + \rho \max_{\chi \in \{\chi^C, \chi^N\}} \widehat{Q}_{i,t_\tau}(p^C, \chi)\right] && \text{(IA.3.69)}\\
&= (1-\alpha)\widehat{Q}_{i,t_\tau}(p^C, \chi^C) + \alpha\left[\pi^{CC} + \rho\widehat{Q}_{i,t_\tau}(p^C, \chi^C)\right] && \text{(IA.3.70)}\\
&= (1-\alpha+\alpha\rho)\widehat{Q}_{i,t_\tau}(p^C, \chi^C) + \alpha\pi^{CC}, && \text{(IA.3.71)}
\end{aligned}$$

where t_τ is the index of the iteration in which i visits the state-and-action pair (p^C, χ^C) for the τ-th time. As τ grows large, $\widehat{Q}_{i,t_\tau}(p^C, \chi^C)$ approaches $Q_i(p^C, \chi^C) \equiv \frac{\pi^{CC}}{1-\rho}$ for $i = 1, 2$, as described in (IA.3.9). Thus, $\widehat{Q}_{i,t}(p^C, \chi^C)$ approaches $\frac{\pi^{CC}}{1-\rho}$ for $i = 1, 2$, as t approaches infinity.

As a consequence of the convergence result for $\widehat{Q}_{i,t}(p^C, \chi^C)$ discussed above, the following convergence result for $\widehat{Q}_{i,t}(p^C, \chi^N)$ also holds after sufficiently many iterations in the exploitation-intensive stage (**Phase 2**):

$$\begin{aligned}
\widehat{Q}_{i,t_{\tau+1}}(p^C, \chi^N) &= (1-\alpha)\widehat{Q}_{i,t_\tau}(p^C, \chi^N) + \alpha\left[\pi^{NC} + \rho \max_{\chi \in \{\chi^C, \chi^N\}} \widehat{Q}_{i,t_\tau}(p^D, \chi)\right] && \text{(IA.3.72)}\\
&= (1-\alpha)\widehat{Q}_{i,t_\tau}(p^C, \chi^N) + \alpha\left[\pi^{NC} + \rho\widehat{Q}_{i,t_\tau}(p^D, \chi^N)\right] && \text{(IA.3.73)}\\
&\approx (1-\alpha)\widehat{Q}_{i,t_\tau}(p^C, \chi^N) + \alpha\left(\pi^{NC} + \frac{\rho\pi^{NN}}{1-\rho}\right), && \text{(IA.3.74)}
\end{aligned}$$

where t_τ is the index of the iteration in which i visits the state-and-action pair (p^C, χ^N) for the τ-th time. As τ grows large, $\widehat{Q}_{i,t_\tau}(p^C, \chi^N)$ approaches $Q_i(p^C, \chi^N) \equiv \pi^{NC} + \frac{\rho\pi^{NN}}{1-\rho}$ for $i = 1, 2$, as described in (IA.3.9). Thus, $\widehat{Q}_{i,t}(p^C, \chi^N)$ approaches $\pi^{NC} + \frac{\rho\pi^{NN}}{1-\rho}$ for $i = 1, 2$, as t approaches infinity.

In summary, we have shown that $\widehat{Q}_{i,t}(p^C, \chi^C) \approx \frac{\pi^{CC}}{1-\rho}$ as t approaches infinity, and $\widehat{Q}_{i,t}(p^C, \chi^N) \approx \pi^{NC} + \frac{\rho \pi^{NN}}{1-\rho}$ as t approaches infinity. Since $\frac{\pi^{CC}}{1-\rho}$ is significantly greater than $\pi^{NC} + \frac{\rho \pi^{NN}}{1-\rho}$ as long as the discount factor ρ is sufficiently close to 1, the system of multiple Q-learning algorithms will converge to the long-run steady state where χ^C is preferred over χ^N at state p^C (i.e., $\widehat{Q}_{i,t}(p^C, \chi^C) > \widehat{Q}_{i,t}(p^C, \chi^N)$ for any sufficiently large t). This inequality will hold consistently, preventing any reversion of the inequalities in (IA.3.68).

Phase 3 ("The Absorbing Stage"): Finally, we show that the collusive equilibrium, sustained by the price-trigger strategy under perfect monitoring, constitutes the long-run steady state of the multi-agent Q-learning system and, importantly, is stable. Specifically, in what follows, we show that, once $\widehat{Q}_{i,t}$ reaches a small neighborhood of the theoretical Q-function Q_i for all $i = 1, 2$, described in (IA.3.9) and (IA.3.10), the system remains within this neighborhood permanently.[4]

To be more specific, we show that, once $\widehat{Q}_{i,t}(s, \chi) = Q_i(s, \chi) + U_{i,t}(s, \chi)$ with $|U_{i,t}(s, \chi)| < \delta$ for all s, χ, and i, for some arbitrarily small $\delta > 0$, the asynchronous update in Q-learning algorithms ensures that $\widehat{Q}_{i,t+1}(s, \chi) = Q_i(s, \chi) + U_{i,t+1}(s, \chi)$ with $|U_{i,t+1}(s, \chi)| < \delta$ for all s, χ, and i. Below, we illustrate the case for (p^C, χ^C), with similar arguments applying to all other cases. If (p^C, χ^C) is not updated for speculator i in the iteration t, the claim is obviously true. If (p^C, χ^C) is updated for speculator i in the iteration t, according to the Q-learning recursive relationship, we have

$$
\begin{aligned}
\widehat{Q}_{i,t+1}(p^C, \chi^C) &= (1-\alpha)\widehat{Q}_{i,t}(p^C, \chi^C) + \alpha \left[\pi^{CC} + \rho \max_{\chi \in \{\chi^C, \chi^N\}} \widehat{Q}_{i,t}(p^C, \chi) \right] && \text{(IA.3.75)} \\
&= (1-\alpha)\widehat{Q}_{i,t}(p^C, \chi^C) + \alpha \left[\pi^{CC} + \rho \widehat{Q}_{i,t}(p^C, \chi^C) \right] && \text{(IA.3.76)} \\
&= (1-\alpha+\alpha\rho) \left[Q_i(p^C, \chi^C) + U_{i,t}(p^C, \chi^C) \right] + \alpha \pi^{CC} && \text{(IA.3.77)} \\
&= Q_i(p^C, \chi^C) + U_{i,t+1}(p^C, \chi^C), && \text{(IA.3.78)}
\end{aligned}
$$

where $U_{i,t+1}(p^C, \chi^C) \equiv (1-\alpha+\alpha\rho) U_{i,t}(p^C, \chi^C)$. Therefore, it holds that

$$|U_{i,t+1}(p^C, \chi^C)| = (1-\alpha+\alpha\rho)|U_{i,t}(p^C, \chi^C)| < \delta. \tag{IA.3.79}$$

[4] This demonstrates that the collusive equilibrium, sustained by the price-trigger strategy under perfect monitoring with perpetual punishment, can be achieved by multiple Q-learning algorithms through effective learning, rather than insufficient learning, despite the theoretical stochastic instability of the limiting equilibrium. The convergence remains approximate when the forgetting rate α is a small constant, rather than decaying to zero over the iterations. The key is that, with any finite number of iterations, small but positive approximation errors and exploration rates lead to an AI equilibrium with limited punishment for deviations, ensuring stochastic stability throughout the learning process, even if not for the limiting equilibrium.

3.1.2 High Noise Trading Risk σ_u with Large ξ

Result 3: *In a market environment characterized by a high σ_u and a large ξ relative to θ, no AI collusive equilibrium sustained by price-trigger strategies can be achieved by multiple informed AI speculators using Q-learning algorithms.*

When σ_u is high, the state variable p_t becomes very noisy, providing little useful information for the Q-learning algorithms to track. Consequently, the algorithms learn to make optimal decisions with minimal reliance on the state variables, effectively behaving as if no state variable is being used. In this scenario, the optimization problem becomes static, and the Q-learning algorithms operate more like bandit algorithms, lacking dynamic sophistication. When price is not an informative state variable, the mechanism behind price-trigger strategies becomes ineffective, as the state variable p_t is now primarily driven by noise trading flows u_t rather than by the trading behavior of informed AI speculators. As a result, no AI collusive equilibrium sustained by price-trigger strategies can be achieved by multiple informed AI speculators using Q-learning algorithms when σ_u is high, even if ξ is large.

Result 4: *In a market environment characterized by a high σ_u and a large ξ relative to θ, an AI collusive equilibrium sustained by over-pruning bias in learning can be achieved by multiple informed AI speculators using Q-learning algorithms.*

While AI collusion via price-trigger strategies is infeasible in this scenario, an alternative mechanism for AI-driven collusion emerges. High noise trading risk disrupts the exploration-exploitation tradeoff, a critical factor for effective learning in reinforcement learning algorithms like Q-learning. In scenarios with high noise trading risk, where Q-learning algorithms become effectively static, we assume without loss of generality that $\rho = 0$. Thus, $\widehat{Q}_{i,t}(\chi)$ follows the recursive relation:

$$\widehat{Q}_{i,t+1}(\chi_{i,t}) = (1-\alpha)\widehat{Q}_{i,t}(\chi_{i,t}) + \ \alpha\pi_{i,t}, \tag{IA.3.80}$$

where $\pi_{i,t}$ is defined in (IA.3.3) and

$$\chi_{i,t} = \begin{cases} \text{argmax}_{\chi\in\{\chi^C,\chi^N\}} \widehat{Q}_{i,t}(\chi), & \text{with probability } 1-\varepsilon_t \\ \text{randomly drawn from } \{\chi^C,\chi^N\}, & \text{with probability } \varepsilon_t. \end{cases} \tag{IA.3.81}$$

In line with the three phases described in Online Appendix 3.1.1, we illustrate the intuition behind the convergence process through the following three steps, consistent with the nature of reinforcement learning algorithms:

Phase 1 ("The Exploration-Intensive Stage"): We provide a heuristic proof that, after many iterations during the exploration-intensive stage with $\varepsilon_t \approx 1$, the Q-function values $\widehat{Q}_{i,t}(\chi^C)$

and $\widehat{Q}_{i,t}(\chi^N)$ are

$$\widehat{Q}_{i,t_\tau}(\chi^C) = \sum_{h=0}^{\tau-1} \alpha(1-\alpha)^h \left[D_h \pi^{CC} + (1-D_h)\pi^{CN} - \xi^{-1}\chi^C u_{t^C_{\tau-h}}\right], \text{ and} \tag{IA.3.82}$$

$$\widehat{Q}_{i,t_\tau}(\chi^N) = \sum_{h=0}^{\tau-1} \alpha(1-\alpha)^h \left[D_h \pi^{NN} + (1-D_h)\pi^{NC} - \xi^{-1}\chi^N u_{t^N_{\tau-h}}\right], \tag{IA.3.83}$$

where t^C_τ and t^N_τ denote the indices of the iterations when the actions χ^C and χ^N are visited for the τ-th time, and D_h are i.i.d. Bernoulli variables with mean $1/2$, capturing random exploration. Thus, given a sufficiently small α, the Law of Large Numbers provides the following approximations:

$$\widehat{Q}_{i,t}(\chi^C) \approx \frac{\pi^{CC} + \pi^{CN}}{2} - \xi^{-1}\chi^C \left(\frac{\alpha\sigma_u^2}{2-\alpha}\right)^{1/2} z_C, \tag{IA.3.84}$$

$$\widehat{Q}_{i,t}(\chi^N) \approx \frac{\pi^{NC} + \pi^{NN}}{2} - \xi^{-1}\chi^N \left(\frac{\alpha\sigma_u^2}{2-\alpha}\right)^{1/2} z_N, \tag{IA.3.85}$$

where z_C and z_N are independent standard normal random variables, defined as follows:

$$z_C \equiv \left(\frac{\alpha\sigma_u^2}{2-\alpha}\right)^{-1/2} \sum_{h=0}^{\infty} u_{t^C_{\tau-h}}, \tag{IA.3.86}$$

$$z_N \equiv \left(\frac{\alpha\sigma_u^2}{2-\alpha}\right)^{-1/2} \sum_{h=0}^{\infty} u_{t^N_{\tau-h}}. \tag{IA.3.87}$$

To further elaborate on the derivation of the approximations in (IA.3.84) – (IA.3.87) above, let us focus on $\widehat{Q}_{i,t}(\chi^C)$ as an illustrative example. Similar to the proof for (IA.3.29), it follows that

$$\sum_{h=0}^{\tau-1} \alpha(1-\alpha)^h \left[D_h \pi^{CC} + (1-D_h)\pi^{CN}\right] \approx \frac{\pi^{CC} + \pi^{CN}}{2}. \tag{IA.3.88}$$

In addition, it holds that

$$\sum_{h=0}^{\tau-1} \alpha(1-\alpha)^h u_{t^C_{\tau-h}} \text{ is a normal random variable with zero mean and variance } \frac{\alpha\sigma_u^2}{2-\alpha}.$$

Importantly, when σ_u is sufficiently large with α kept constant, the noise terms in (IA.3.84) and (IA.3.85) dominate the constant terms. Consequently, unlike in a low-noise trading environment, intensive exploration does not guide Q-learning algorithms toward adopting or learning aggressive trading strategies. However, we emphasize that, if α is sufficiently close to zero, even with a large σ_u, the constant terms in (IA.3.84) and (IA.3.85) may still dominate the noise terms. As a result, even in a high-noise trading environment, as long as α is sufficiently small, intensive exploration does guide Q-learning algorithms toward aggressive trading strategies, which is consistent with our simulation experimental results discussed in Figure IA.9 of the main text.

Phase 2 ("The Exploitation-Intensive Stage"): From (IA.3.82) – (IA.3.85), it is clear that aggressive trading behavior, χ^N, is particularly vulnerable to poor trading outcomes driven by large noise trading flows moving in the same direction (i.e., when u_t is highly positive). In these cases, the algorithm "labels" χ^N as "disastrous action," meaning that the estimated Q-value, $\widehat{Q}_{i,t+1}(\chi^N)$, is updated to a level significantly lower than $\widehat{Q}_{i,t+1}(\chi^C)$.

As a result, during the exploitation-intensive stage, the algorithm is strongly discouraged from revisiting χ^N, effectively pruning χ^N from the learning process and causing it to remain persistently undervalued. In other words, when noise trading risk σ_u is high, the Q-learning system becomes excessively noisy, amplifying the potential learning bias caused by over-pruning during exploitation. At the same time, during the exploitation-intensive phase, exploration not only becomes infrequent but also ineffective when it occurs due to high noise. Taken together, the potential learning bias from exploitation cannot be fully corrected by exploration, disrupting the balance of the exploration-exploitation tradeoff. This imbalance leads to persistent over-pruning bias, where the agent's policy prematurely converges to a suboptimal solution, ignoring other potentially beneficial strategies.

To be more specific, suppose χ^N is initially preferred over χ^C for both informed AI speculators using Q-learning algorithms at the start of the exploitation-intensive stage (**Phase 2**). After many exploitation iterations during the beginning period of the exploitation-intensive stage (**Phase 2**), the estimated Q-values $\widehat{Q}_{i,t}(\chi^N)$ for $i = 1, 2$ are

$$\widehat{Q}_{i,t}(\chi^N) \approx \pi^{NN} - \xi^{-1}\chi^N \left(\frac{\alpha\sigma_u^2}{2-\alpha}\right)^{1/2} z_N, \tag{IA.3.89}$$

where z_N is a standard normal variable. During the iterations of the exploitation-intensive stage (**Stage 2**), it is almost certain that a 3-standard-deviation positive noise trading flow ($u_t = 3$) will occur when z_N takes a high value ($z_N \approx 3$), corresponding to three standard deviations above the mean of a standard normal distribution. This event drives the Q-value of χ^N down to an excessively low level: for $i = 1, 2$, it holds that

$$\begin{aligned}
&\widehat{Q}_{i,t+1}(\chi^N) = (1-\alpha)\widehat{Q}_{i,t}(\chi^N) + \alpha\pi_{i,t} \\
&\approx (1-\alpha)\left[\pi^{NN} - 3\xi^{-1}\chi^N\left(\frac{\alpha\sigma_u^2}{2-\alpha}\right)^{1/2}\right] + \alpha\pi_{i,t} \quad \text{according to (IA.3.89)} \\
&\approx (1-\alpha)\left[\pi^{NN} - 3\xi^{-1}\chi^N\left(\frac{\alpha\sigma_u^2}{2-\alpha}\right)^{1/2}\right] + \alpha\left[\pi^{NN} - 3\xi^{-1}\chi^N\right] \quad \text{according to (IA.3.3)} \\
&= \pi^{NN} - 3(1-\alpha)\xi^{-1}\chi^N\left(\frac{\alpha\sigma_u^2}{2-\alpha}\right)^{1/2} - 3\alpha\xi^{-1}\chi^N \qquad \text{(IA.3.90)}
\end{aligned}$$

As a result, χ^C is likely to be preferred over χ^N following this large positive noise trading flow, leading the exploitation iterations to focus on updating the Q-value of χ^C:

$$\widehat{Q}_{i,t}(\chi^C) \approx \pi^{CC} - \xi^{-1}\chi^C\left(\frac{\alpha\sigma_u^2}{2-\alpha}\right)^{1/2} z_C, \quad \text{for } i = 1, 2, \qquad \text{(IA.3.91)}$$

which can rarely go below $\widehat{Q}_{i,t}(\chi^N)$ in (IA.3.90).

On the other hand, when noise trading flows move in the opposite direction (i.e., with a highly negative u_t), χ^N may temporarily yield exceptional profits. This will cause the algorithm to "label" χ^N as "favorable action," meaning that the estimated Q-value, $\widehat{Q}_{i,t+1}(\chi^N)$, is updated to a level significantly greater than $\widehat{Q}_{i,t+1}(\chi^C)$. This inflation in its Q-value, relative to χ^C, leads to repeated reinforcement of χ^N. For instance, a 3-standard-deviation negative noise flow occurs ($u_t = -3$) when z_N takes a large negative value, say $z_N \approx -3$, pushing the Q-value of χ^N to an excessively high level (following the same reasoning as the approximation in (IA.3.89)):

$$\widehat{Q}_{i,t}(\chi^N) \approx \pi^{NN} + 3(1-\alpha)\xi^{-1}\chi^N\left(\frac{\alpha\sigma_u^2}{2-\alpha}\right)^{1/2} + 3\alpha\xi^{-1}\chi^N, \quad \text{for } i = 1, 2. \qquad \text{(IA.3.92)}$$

In this way, the Q-value of χ^N can temporarily exceed that of χ^C, leading the algorithm to "label" χ^N as "favorable action" and reinforce its selection. However, over many iterations, this temporary inflation effect, represented by the last two terms on the right-hand side of (IA.3.92), will average out, resulting in the long-run Q-value for χ^N, specified in (IA.3.90).

Taken together, the exploration-exploitation tradeoff in reinforcement learning parallels the bias-variance tradeoff in supervised learning and high-dimensional statistics. Both tradeoffs seek a balance between pruning the choice space[5] and reducing the variability of the outcome. In reinforcement learning, exploring (i.e., trying new actions) is necessary to minimize bias in

[5] The choice space varies across contexts: in reinforcement learning, it is the action space; in supervised learning and high-dimensional statistics, it is the parameter space (e.g., Hastie, Tibshirani and Friedman, 2009; Hall and Horowitz, 2007; Dou, Pollard and Zhou, 2012).

estimating the optimal action, while exploiting (i.e., choosing believed optimal actions) helps reduce action and reward variability. Much like shrinkage techniques in supervised learning and dimensionality reduction in high-dimensional statistics, exploitation in reinforcement learning serves to prune the choice space, reducing variability and introducing a degree of bias to enhance learning efficiency overall.

Exploitation asymmetrically influences the learning process when responding to adverse versus beneficial large noise trading shocks. After an adverse noise shock, the strategy chosen is often "marked" as "disastrous action," with a very low estimated Q-value assigned to that strategy. Exploitation discourages the algorithm from revisiting this strategy, preventing the correction of biased evaluations for such off-equilibrium actions. Conversely, after a beneficial noise shock, the strategy is labeled "favorable action" and a very high estimated Q-value is assigned. Exploitation ensures the algorithm continues to play and update this chosen strategy, allowing any bias for this on-equilibrium-path action to be largely corrected.

Because more aggressive strategies are significantly more susceptible to large adverse shocks, this asymmetry causes their persistent undervaluation. These aggressive strategies are prematurely pruned from the set of potential optimal strategies. In the case with large noise trading shocks, informed AI speculators using Q-learning algorithms eventually sustain conservative strategies after the learning process, which aligns with collusive trading behavior as defined in Definition 3.1.

3.2 Little Presence of Information-Insensitive Investors (i.e., Small ξ)

We now consider the case with high price efficiency, that is, the case when ξ is close to zero, relative to θ. Throughout Section 3, we consider $I = 2$. For illustration purpose, we focus on the benchmark where $\xi = 0$. According to Propositions IA.1 and IA.2, it holds that

$$\chi^N = \widehat{\chi}^N \sigma_u, \text{ with } \widehat{\chi}^N = 1/\sqrt{I}, \tag{IA.3.93}$$

$$\chi^M = \widehat{\chi}^M \sigma_u, \text{ with } \widehat{\chi}^M = 1/I. \tag{IA.3.94}$$

Thus, the choice set can be further reduced to $\mathbf{X} = \{\widehat{\chi}^N, \widehat{\chi}^C\}$ with $\widehat{\chi}^M \leq \widehat{\chi}^C < \widehat{\chi}^N$.

A key characteristic of financial markets is that some investors infer the value of an asset by observing the trading behavior of informed speculators. As a result, the asset's demand curve is endogenously shaped and varies across different equilibria, rather than being fixed. In the equilibrium where informed AI speculators mostly opt trading strategy $\widehat{\chi}^C$, the demand curve for the asset set by the market maker is

$$p^C(y_t) = \overline{v} + \lambda^C y_t, \text{ with } \lambda^C = \frac{1}{\sigma_u} \frac{I\widehat{\chi}^C}{(I\widehat{\chi}^C)^2 + 1}. \tag{IA.3.95}$$

Thus, at t, the market price, p_t, is

$$p_t = \begin{cases} p_t^{C,N} \equiv \overline{v} + \dfrac{(I\widehat{\chi}^C)(I\widehat{\chi}^N)}{(I\widehat{\chi}^C)^2+1} + \dfrac{I\widehat{\chi}^C}{(I\widehat{\chi}^C)^2+1}\widehat{u}_t, & \text{if } (\widehat{\chi}_i, \widehat{\chi}_{-i}) = (\widehat{\chi}^N, \widehat{\chi}^N), \\ p_t^{C,C} \equiv \overline{v} + \dfrac{(I\widehat{\chi}^C)^2}{(I\widehat{\chi}^C)^2+1} + \dfrac{I\widehat{\chi}^C}{(I\widehat{\chi}^C)^2+1}\widehat{u}_t, & \text{if } (\widehat{\chi}_i, \widehat{\chi}_{-i}) = (\widehat{\chi}^C, \widehat{\chi}^C), \\ p_t^{C,D} \equiv \overline{v} + \dfrac{(I\widehat{\chi}^C)[(I-1)\widehat{\chi}^C + \widehat{\chi}^N]}{(I\widehat{\chi}^C)^2+1} + \dfrac{I\widehat{\chi}^C}{(I\widehat{\chi}^C)^2+1}\widehat{u}_t, & \text{otherwise,} \end{cases} \tag{IA.3.96}$$

where $\widehat{u}_t \equiv u_t/\sigma_u$ is a standard normal variable, and the trading profit for speculator i is

$$\pi_{i,t} = \left[1 - \frac{I\widehat{\chi}^C}{(I\widehat{\chi}^C)^2+1}(\widehat{\chi}_{i,t} + \widehat{\chi}_{-i,t})\right]\widehat{\chi}_{i,t}\sigma_u - \frac{I\widehat{\chi}^C}{(I\widehat{\chi}^C)^2+1}\widehat{\chi}_{i,t}u_t, \tag{IA.3.97}$$

where

$$\pi_{i,t} = \begin{cases} \pi^{C,NN} - \lambda^C\chi^N u_t, & \text{with } \pi^{C,NN} = \left(1 - I\lambda^C\chi^N\right)\chi^N, & \text{if } (\chi_i, \chi_{-i}) = (\chi^N, \chi^N), \\ \pi^{C,CC} - \lambda^C\chi^C u_t, & \text{with } \pi^{C,CC} = \left(1 - I\lambda^C\chi^C\right)\chi^C, & \text{if } (\chi_i, \chi_{-i}) = (\chi^C, \chi^C), \\ \pi^{C,CN} - \lambda^C\chi^C u_t, & \text{with } \pi^{C,CN} = \left[1 - \lambda^C(\chi^C + (I-1)\chi^N)\right]\chi^C, & \text{if } (\chi_i, \chi_{-i}) = (\chi^C, \chi^N), \\ \pi^{C,NC} - \lambda^C\chi^N u_t, & \text{with } \pi^{C,NC} = \left[1 - \lambda^C((I-1)\chi^C + \chi^N)\right]\chi^N, & \text{if } (\chi_i, \chi_{-i}) = (\chi^N, \chi^C). \end{cases} \tag{IA.3.98}$$

It is straightforward to show that $p_t^{C,C} < p_t^{C,D} < p_t^{C,N}$, and $\pi^{C,CN} < \pi^{C,NN} < \pi^{C,CC} < \pi^{C,NC}$, and $\pi^{C,CN} + \pi^{C,CC} < \pi^{C,NN} + \pi^{C,NC}$.

On the other hand, in the equilibrium where informed AI speculators mostly opt trading strategy $\widehat{\chi}^N$, the demand curve for the asset set by the market maker is

$$p^N(y_t) = \overline{v} + \lambda^N y_t, \text{ with } \lambda^N = \frac{1}{\sigma_u}\frac{I\widehat{\chi}^N}{(I\widehat{\chi}^N)^2+1}. \tag{IA.3.99}$$

Thus, at t, the market price, p_t, is

$$p_t = \begin{cases} p_t^{N,N} \equiv \overline{v} + \dfrac{(I\widehat{\chi}^N)^2}{(I\widehat{\chi}^N)^2+1} + \dfrac{I\widehat{\chi}^N}{(I\widehat{\chi}^N)^2+1}\widehat{u}_t, & \text{if } (\widehat{\chi}_i, \widehat{\chi}_{-i}) = (\widehat{\chi}^N, \widehat{\chi}^N), \\ p_t^{N,C} \equiv \overline{v} + \dfrac{(I\widehat{\chi}^N)(I\widehat{\chi}^C)}{(I\widehat{\chi}^N)^2+1} + \dfrac{I\widehat{\chi}^N}{(I\widehat{\chi}^N)^2+1}\widehat{u}_t, & \text{if } (\widehat{\chi}_i, \widehat{\chi}_{-i}) = (\widehat{\chi}^C, \widehat{\chi}^C), \\ p_t^{N,D} \equiv \overline{v} + \dfrac{(I\widehat{\chi}^N)(\widehat{\chi}^C + (I-1)\widehat{\chi}^N)}{(I\widehat{\chi}^N)^2+1} + \dfrac{I\widehat{\chi}^N}{(I\widehat{\chi}^N)^2+1}\widehat{u}_t, & \text{otherwise,} \end{cases} \tag{IA.3.100}$$

where $\widehat{u}_t \equiv u_t/\sigma_u$ is a standard normal variable, and the trading profit for speculator i is

$$\pi_{i,t} = \left[1 - \frac{I\widehat{\chi}^N}{(I\widehat{\chi}^N)^2+1}(\widehat{\chi}_{i,t} + \widehat{\chi}_{-i,t})\right]\widehat{\chi}_{i,t}\sigma_u - \frac{I\widehat{\chi}^N}{(I\widehat{\chi}^N)^2+1}\widehat{\chi}_{i,t}u_t, \tag{IA.3.101}$$

where

$$\pi_{i,t}=\begin{cases}\pi^{N,NN}-\lambda^N\chi^N u_t, & \text{with } \pi^{N,NN}=\left(1-I\lambda^N\chi^N\right)\chi^N, & \text{if } (\chi_i,\chi_{-i})=(\chi^N,\chi^N),\\ \pi^{N,CC}-\lambda^N\chi^C u_t, & \text{with } \pi^{N,CC}=\left(1-I\lambda^N\chi^C\right)\chi^C, & \text{if } (\chi_i,\chi_{-i})=(\chi^C,\chi^C),\\ \pi^{N,CN}-\lambda^N\chi^C u_t, & \text{with } \pi^{N,CN}=\left[1-\lambda^N(\chi^C+(I-1)\chi^N)\right]\chi^C, & \text{if } (\chi_i,\chi_{-i})=(\chi^C,\chi^N),\\ \pi^{N,NC}-\lambda^N\chi^N u_t, & \text{with } \pi^{N,NC}=\left[1-\lambda^N((I-1)\chi^C+\chi^N)\right]\chi^N, & \text{if } (\chi_i,\chi_{-i})=(\chi^N,\chi^C).\end{cases}\tag{IA.3.102}$$

It is straightforward to show that $p_t^{N,C}<p_t^{N,D}<p_t^{N,N}$, and $\pi^{N,CN}<\pi^{N,NN}<\pi^{N,CC}<\pi^{N,NC}$, and $\pi^{N,CN}+\pi^{N,CC}<\pi^{N,NN}+\pi^{N,NC}$.

Result 5: *In a market environment characterized by a small ξ relative to θ, no AI collusive equilibrium sustained by price-trigger strategies can be achieved by multiple informed AI speculators using Q-learning algorithms.*

Regardless of the market's equilibrium or the noise trading level, σ_u, the maximum standardized price deviation is given by

$$\frac{\Delta_D(p_t)}{\sigma(p_t)}=\widehat{\chi}^N-\widehat{\chi}^C\le\frac{1}{\sqrt{I}}-\frac{1}{I},\tag{IA.3.103}$$

where $\Delta_D(p_t)=|p_t^{C,C}-p_t^{C,D}|$ and $\sigma(p_t)$ denote the maximum price deviation caused by the deviation of an informed AI speculator, and the price volatility driven by randomness in the model, respectively, in two different equilibria. Specifically:

- $\Delta_D(p_t)=|p_t^{C,C}-p_t^{C,D}|$ in the equilibrium where informed AI speculators predominantly opt trading strategy $\widehat{\chi}^C$, and $\Delta_D(p_t)=|p_t^{N,N}-p_t^{N,D}|$ in the equilibrium where informed AI speculators predominantly opt trading strategy $\widehat{\chi}^N$.
- $\sigma(p_t)$ denotes the standard deviation of $p_t^{C,C}$ (or $p_t^{C,N}$ or $p_t^{C,D}$) in the equilibrium where informed AI speculators predominantly opt trading strategy $\widehat{\chi}^C$, and the same notation $\sigma(p_t)$ denotes the standard deviation of $p_t^{N,C}$ (or $p_t^{N,N}$ or $p_t^{N,D}$) in the equilibrium where informed AI speculators predominantly opt trading strategy $\widehat{\chi}^N$.

This implies that, regardless of σ_u, the equilibrium market price remains too noisy to serve as an informative state variable in Q-learning algorithms, preventing the sustainability of price-triggered collusion based on private information. This is key to the theoretical result: collusion through price-trigger strategies becomes unsustainable when ξ is sufficiently small, all else being equal. Intuitively, when $\xi\approx 0$, informed speculators must trade conservatively relative to the noise trading risk σ_u to earn information rents. Since information rents are the sole source of profits and market makers learn about the fundamental value v_t through trading flows, informed speculators are compelled to trade conservatively. As a result, the equilibrium market price is significantly driven by noise trading u_t, rather than reflecting the fundamental value v_t, regardless of the level of σ_u.

Result 6: *In a market environment characterized by a small ξ relative to θ, an AI collusive equilibrium sustained by over-pruning bias in learning can be achieved by multiple informed AI speculators using Q-learning algorithms.*

Although AI collusion through price-trigger strategies cannot be learned once ξ becomes close to zero, relative to θ, a distinct mechanism for AI collusion begins to emerge. We now illustrate the intuition behind collusion sustained by over-pruning bias in learning. As discussed in the heuristic proof for **Result 5** above, when ξ is close to zero relative to θ, the state variable p_t becomes too noisy to provide useful information for the Q-learning algorithms to track. In this case, the algorithms learn to make optimal decisions with minimal reliance on the state variables, effectively behaving as if no state variable is being used.

Similar to the case with excessively large σ_u, the endogenous trading profit is primarily influenced by the noise trading flow u_t. As a result, the observed trading profit lacks enough information about the strategy choice to guide the selection of optimal trading strategies, rendering the algorithms' exploration ineffective. Specifically, following the same reasoning in the heuristic proof for **Result 4** in Section 3.1.2, analogous to the case with excessively large σ_u, intensive exploration during the exploration-intensive stage fails to enable Q-learning algorithms to adopt or learn more aggressive trading strategies. Subsequently, during the exploitation-intensive stage, the aggressive trading strategy χ^N is under-learned due to over-pruning, as asymmetric exploitation prematurely dismisses the aggressive strategy χ^N.

4 Supplementary Simulation Results

4.1 Measures in Simulation Experiments

In each of the $N_{sim} = 1,000$ simulation experiments, we compute the following measures over $T = 100,000$ periods after Q-learning algorithms reach convergence at T_c.

Collusion Capacity Δ^C**.** We compute Δ^C as follows:

$$\Delta^C = \frac{1}{I}\sum_{i=1}^{I}\Delta_i^C, \text{ with } \Delta_i^C = \frac{\overline{\pi}_i - \overline{\pi}_i^N}{\overline{\pi}_i^M - \overline{\pi}_i^N}, \tag{IA.4.1}$$

where $\overline{\pi}_i \equiv \sum_{t=T_c}^{T_c+T}\pi_{i,t}$ is the average profits of informed AI speculator i.[6] The values of $\overline{\pi}_i^N = \sum_{t=T_c}^{T_c+T}\pi_{i,t}^N$ and $\overline{\pi}_i^M = \sum_{t=T_c}^{T_c+T}\pi_{i,t}^M$ are the average profit that informed speculator i would obtain in the theoretical benchmarks for the noncollusive Nash equilibrium and perfect cartel equlibrium, respectively. Because informed speculators are symmetric, we have $\pi_{i,t}^N \equiv \pi_t^N$ and $\pi_{i,t}^M \equiv \pi_t^M$ for

[6] By taking the average over a large number of periods, we smooth out the randomness in the underlying economic environment, caused by the randomness in the noise trader's order flow u_t and the asset's value v_t.

all $i = 1, ..., I$. Specifically, conditional on the realized values of v_t and u_t in period t, informed speculator i's profit in the theoretical benchmark for the noncollusive Nash equilibrium is

$$\pi_t^N = \left[v_t - p^N(Ix^N(v_t) + u_t)\right] x^N(v_t), \quad \text{for } i = 1, ..., I, \tag{IA.4.2}$$

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

$$\pi_t^M = \left[v_t - p^M(Ix^M(v_t) + u_t)\right] x^M(v_t), \quad \text{for } i = 1, ..., I, \tag{IA.4.3}$$

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

In principle, the value of Δ^C should range from 0 to 1. A larger Δ^C implies that informed AI speculators attain higher profits. The value of Δ^C can never be larger than 1 because $\overline{\pi}_i^M$ is the highest theoretically possible average profit. In fact, because informed AI speculators can only choose actions over discrete grids, by design, it is not possible to obtain $\Delta^C = 1$ in our simulation experiments. However, it is possible to achieve a Δ^C below 0 under the limit strategies of informed AI speculators. This outcome implies that informed AI speculators failed to learn a good approximation of the theoretical Q-matrix, and as a result, they achieve average profits lower than those in the theoretical benchmark for the noncollusive Nash equilibrium.

Profit Gain Relative to Noncollusion. The Δ^C measure is informative about collusive behavior. However, it does not tell us the relative magnitude of supra-competitive profits. We thus also calculate the extra profit gain relative to the profits that informed AI speculators would obtain in the theoretical benchmark for the noncollusive Nash equilibrium. Specifically, the relative profit gain is $\sum_{i=1}^{I} \overline{\pi}_i / \sum_{i=1}^{I} \overline{\pi}_i^N$, where $\overline{\pi}_i$ and $\overline{\pi}_i^N$ are calculated similarly as those in equation (IA.4.1).

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

In our simulation experiments, informed AI speculators directly learn $x_{i,t}$ without imposing the linearity restriction between $x_{i,t}$ and v_t. Despite this, we find that informed AI speculators learn roughly linear strategies (see Figure IA.5 in Online Appendix 4.7). We estimate the trading policy $\widehat{\chi}^C$ based on the recorded asset's values and order flows $\{v_t, x_{i,t}\}_{t=T_c}^{T_c+T}$ for each informed AI speculator $i = 1, ..., I$, by running the following linear regression:

$$x_{i,t} = \chi_{i,0}^C + \chi_{i,1}^C v_t + \epsilon_t. \tag{IA.4.4}$$

Consistent with our model, the estimates based on the simulated data satisfy $\widehat{\chi}_{i,0}^C \approx -\overline{v}\widehat{\chi}_{i,1}^C$ in the unrestricted regression (IA.4.4). The estimate $\widehat{\chi}_{i,1}^{C}$ captures the sensitivity of $x_{i,t}$ to v_t corresponding to the limit trading strategies of informed AI speculators after their Q-learning algorithms converge. We further compute the average trading policy of informed AI speculators as $\widehat{\chi}^{C} = \frac{1}{I}\sum_{i=1}^{I}\widehat{\chi}_{i,1}^{C}$.

Price Informativeness. Consistent with Definition 3.4 in the main text, the degree of price informativeness in our simulation experiments is measured by the signal-noise ratio as follows:

$$\mathcal{I}^{C} = \frac{\mathrm{var}(x_t^{C})}{\mathrm{var}(u_t)} = \frac{\mathrm{var}(\sum_{i=1}^{I} x_{i,t}^{C})}{\mathrm{var}(u_t)} = (I\widehat{\chi}^{C})^2(\widehat{\sigma}_v/\sigma_u)^2, \tag{IA.4.5}$$

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

Market Liquidity. Consistent with Definition 3.4 in the main text, the market liquidity in period t is measured by the inverse sensitivity of the absolute value of the market maker's inventory $m_t = -(z_t + y_t)$ to noise order flows u_t

$$\mathcal{L}_t^{C} = \left[\frac{\partial |m_t|}{\partial u_t}\right]^{-1} = \frac{1}{|1-\xi\widehat{\lambda}_t|}, \tag{IA.4.6}$$

where $z_t = -\xi(p_t - \overline{v}) = -\xi\widehat{\lambda}_t y_t$ and $\widehat{\lambda}_t$ is given by equation (4.2) in the main text. The average market liquidity is computed as $\mathcal{L}^{C} = \sum_{t=T_c}^{T_c+T}\mathcal{L}_t^{C}$.

Mispricing. Consistent with Definition 3.4 in the main text, the magnitude of mispricing in period t is measured by the deviation of the conditional value of the asset's price p_t from the asset's value v_t

$$\mathcal{E}_t^{C} = |\mathbb{E}[p_t|v_t] - v_t| = \left(1 - \widehat{\lambda}_t I\widehat{\chi}^{C}\right)|v_t - \overline{v}|, \tag{IA.4.7}$$

where $\widehat{\lambda}_t$ is given by equation (4.2) in the main text. The average mispricing is computed as $\mathcal{E}^{C} = \sum_{t=T_c}^{T_c+T}\mathcal{E}_t^{C}$.

4.2 Testing If Outcomes Form An Experience-Based Equilibrium

In this appendix section, we formally test whether the simulation outcomes of informed AI speculators constitute an experience-based equilibrium or a restricted experience-based equilibrium.

Test for Experience-Based Equilibrium. Let $\widetilde{Q}_i(s, x_i)$ and $\widetilde{x}_i(s)$ denote the estimated Q-matrix and limit trading strategies for each informed AI speculator $i \in \mathcal{I}$, respectively, after all informed AI speculators' Q-learning algorithms converge. That is, $\widetilde{Q}_i(s, x_i) = \widehat{Q}_{i,T_c}(s, x_i)$ if all algorithms converge at T_c, and $\widetilde{x}_i(s) = \mathrm{argmax}_{x_i}\,\widetilde{Q}_i(s, x_i)$. We adapt the test for the experience-based equilibrium proposed by Fershtman and Pakes (2012) in our setting with Q-learning algorithms. Specifically, Fershtman and Pakes (2012, Section IV.A) show that to test whether the output of an algorithm constitutes an experience-based equilibrium, one needs to check whether the three conditions that define an experience-based equilibrium are simultaneously satisfied.

Condition 1 requires the Markov process generated by any initial condition $s_0 = \{p_{-1}, v_{-1}, v_0\} \in \mathcal{R}$ and the transition kernel generated by the limit trading strategies $\{\widetilde{x}_i(s)\}_{i=1}^{I}$ to have $\mathcal{R}$ as a recurrent class. Thus, with probability 1, any subgame starting from an arbitrary $s_t \in \mathcal{R}$ in period t will result in sample paths that are within $\mathcal{R}$ forever. As noted by Fershtman and Pakes (2012), condition 1 essentially requires one to find the recurrent class of states. Thus, we start with an arbitrary state s_0 and use the limit trading strategies $\{\widetilde{x}_i(s)\}_{i=1}^{I}$ to simulate a sample path $\{s_t\}_{t=1}^{T_0+T}$, with sufficiently large T_0 and T. The first T_0 periods are dropped as burn-in. The set of states that are visited between $T_0 + 1$ and $T_0 + T$ constitute a recurrent class $\mathcal{R}$ generated by $\{\widetilde{Q}_i(s, x_i)\}_{i=1}^{I}$. In our test, we set $T_0 = 100,000$ and $T = 1,000,000$. We further verify that setting larger values for T_0 or T will not alter $\mathcal{R}$.

Condition 2 requires that for each state $s \in \mathcal{R}$, trading policies are optimal for each informed AI speculator $i \in \mathcal{I}$ given its Q-matrix $\widetilde{Q}_i(s, x_i)$.[7] That is, $\widetilde{x}_i(s)$ solves

$$\max_{x_i \in \mathcal{X}} \widetilde{Q}_i(s, x_i) \text{ for all } i \in \mathcal{I}. \tag{IA.4.8}$$

Condition 2 is satisfied by construction of Q-learning algorithms because a necessary criterion for convergence is that all Q-learning algorithms are purely in the exploitation mode. In the exploitation mode, all informed AI speculators choose order flows to maximize their Q values at each state s.

Condition 3 requires consistency of values on $\mathcal{R}$. Specifically, for each state $s \in \mathcal{R}$ and each informed AI speculator $i \in \mathcal{I}$,

$$\widetilde{Q}_i(s, x_i(s)) = \pi_i(\widetilde{x}_i(s), \widetilde{x}_{-i}(s)) + \rho \sum_{s'} \widetilde{Q}_i(s', x_i(s')) prob(s'|s), \tag{IA.4.9}$$

where $\pi_i(\widetilde{x}_i(s), \widetilde{x}_{-i}(s))$ is the profit that informed AI speculator i receives by choosing the limit trading strategy $\widetilde{x}_i(s)$, conditional on other informed AI speculators $j \in \mathcal{I}$ and $j \neq i$ choosing limit trading strategies. In principle, π_i also depends on the realized noise order flow u and the market maker's pricing rule. We omit this dependence to simplify the notations. The variable $prob(s'|s)$ captures the transition probability from the current state s to the state s' in the next period. This transition probability also depends on informed AI speculators' order flows, the noise order flow u, and the market maker's pricing rule.

To verify condition 3, we check the consistency of $\widetilde{Q}_i(s, x_i)$ with outcomes generated by the limit policies $\widetilde{x}_i(s)$ in each state in $\mathcal{R}$. We follow the method of Fershtman and Pakes (2012) and

[7] Note that in our simulation experiments, the state variables $s_t = \{p_{t-1}, v_{t-1}, v_t\}$ only include public information that is available to all informed AI speculators. Thus, each informed AI speculator's own information set $\mathcal{J}_{i,t}$ is identical to s_t. Fershtman and Pakes (2012) consider the general setting by allowing $\mathcal{J}_{i,t}$ to include private information that is only available to agent i. In this case, $s_t = \bigcup_{i=1}^{I} \mathcal{J}_{i,t}$.

directly use the simulated sample path from T_0+1 to T_0+T to conduct the test. Specifically, for each period t along the simulated sample path, we compute the perceived value as follows:

$$\widetilde{V}_{i,t}(s_t) = (v_t - p_t)\widetilde{x}_i(s_t) + \rho \max_{x' \in \mathcal{X}} \widetilde{Q}_i(s_{t+1}, x'). \tag{IA.4.10}$$

Then, for each state $s \in \mathcal{R}$, we find all the periods from T_0+1 to T_0+T in which the state s is visited. Let the total number of visits be $N(s)$ and the period for the τth visit be $t(\tau)$, with $\tau = 1, 2, ..., N(s)$. The α-weighted average of the simulated perceived values across visits to the same state s is computed as follows:

$$\widehat{\mu}_i(s) = \sum_{\tau=1}^{N(s)} \alpha(1-\alpha)^{\tau-1} \widetilde{V}_{i,t(N(s)-\tau+1)}(s). \tag{IA.4.11}$$

If the simulation results constitute an experience-based equilibrium, the value of $\widehat{\mu}_i(s)$ should be close to $\widetilde{Q}_i(s, \widetilde{x}_i(s))$ for all $s \in \mathcal{R}$ and $i \in \mathcal{I}$, when $N(s)$ is sufficiently large.[8]

Formally, the deviation (i.e., bias) of $\widehat{\mu}_i(s)$ from $\widetilde{Q}_i(s, \widetilde{x}_i(s))$ can be estimated as follows. For each of the $N_{sim} = 1,000$ simulation sessions, we use the limit trading strategies $\{\widetilde{x}_i(s)\}_{i=1}^{I}$ to simulate $K = 500$ sample paths. When simulating each sample path, we start from an arbitrary s_0 and simulate for T_0+T periods, with the first T_0 periods dropped as burn-in. Next, we compute $\widehat{\mu}_i(s)$ for each sample path (from T_0+1 to T) according to equation (IA.4.11). Let $\mathbb{E}[\cdot]$ denote the average value over $K = 500$ sample paths. The percentage bias squared of $\widehat{\mu}_i(s)$ as an estimate of $\widetilde{Q}_i(s, \widetilde{x}_i(s))$ is computed as follows:

$$\begin{aligned} Bias_i^2(s) = & \left(\frac{\mathbb{E}[\widehat{\mu}_i(s)] - \widetilde{Q}_i(s, \widetilde{x}_i(s))}{\widetilde{Q}_i(s, \widetilde{x}_i(s))} \right)^2 \\ = & \mathbb{E}\left[\left(\frac{\widehat{\mu}_i(s) - \widetilde{Q}_i(s, \widetilde{x}_i(s))}{\widetilde{Q}_i(s, \widetilde{x}_i(s))} \right)^2 \right] - \mathbb{E}\left[\left(\frac{\widehat{\mu}_i(s) - \mathbb{E}[\widehat{\mu}_i(s)]}{\widetilde{Q}_i(s, \widetilde{x}_i(s))} \right)^2 \right] \\ = & \widehat{MSE}_i(s) - \widehat{\sigma}_i^2(s). \end{aligned} \tag{IA.4.12}$$

The average percentage bias squared across all states $s \in \mathcal{R}$ and all informed AI speculators $i \in \mathcal{I}$

[8] Because our Q-learning algorithm is different from the reinforcement learning algorithm used by Fershtman and Pakes (2012), rather than taking a simple average, we compute the α-weighted average in equation (IA.4.11). This modification for the computation of $\widehat{\mu}_i(s)$ in equation (IA.4.11) follows the idea of Fershtman and Pakes (2012). Specifically, were we to substitute the true unbiased equilibrium $Q_i(s,x)$ and $x_i(s)$ for the $\widetilde{Q}_i(s,x)$ and $\widetilde{x}_i(s)$ in equation (IA.4.10) to compute $\widetilde{V}_{i,t}(s_t)$, then $\widehat{\mu}_i(s)$ computed in equation (IA.4.11) will also be the true unbiased equilibrium $Q_i(s, x_i(s))$ as $N(s) \to \infty$.

is computed as follows:

$$\begin{aligned} Avg\_Bias^2 &= \sum_{s\in\mathcal{R}} w(s) Bias_i^2(s) \\ &= \sum_{s\in\mathcal{R}} w(s)\left(\widehat{MSE}_i(s) - \widehat{\sigma}_i^2(s)\right), \end{aligned} \tag{IA.4.13}$$

where the weight, $w(s) = \frac{\mathbb{E}[N(s)]}{\sum_{s\in\mathcal{R}}\mathbb{E}[N(s)]}$, is determined by the average number of visits to state s across $K = 500$ sample paths.

Test for Restricted Experience-Based Equilibrium. The test for restricted experience-based equilibrium is similar to the test for experience-based equilibrium except that for each period t along the simulated sample path, we compute the perceived value for each $x \in \mathcal{X}$. That is, equation (IA.4.10) is modified as follows:

$$\widetilde{V}_{i,t}(s_t, x) = (v_t - p_t)x + \rho \max_{x'\in\mathcal{X}} \widetilde{Q}_i(s_{t+1}, x') \quad \text{for all} \quad x \in \mathcal{X}. \tag{IA.4.14}$$

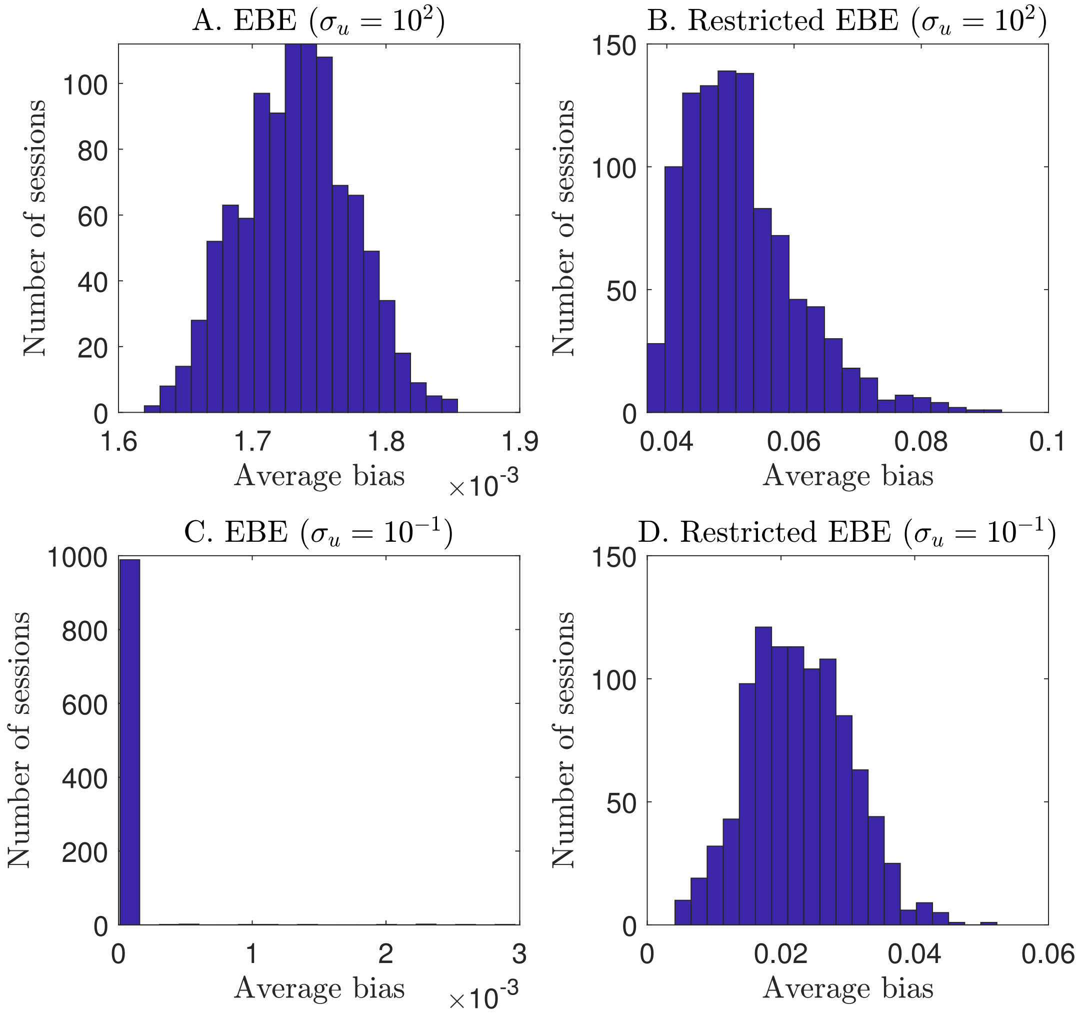
The evolution of s_t is still determined by the limit trading strategies $\{\widetilde{x}_i(s)\}_{i=1}^I$. Moreover, when computing $\widetilde{V}_{i,t}(s_t, x)$ for informed AI speculator i using trading strategy x, all other informed AI speculators (i.e., $j \in \mathcal{I}$ and $j \neq i$) still following their limit trading strategies.

Next, we modify equation (IA.4.11) to compute the α-weighted average of the simulated perceived values for each $x \in \mathcal{X}$ as follows:

$$\widehat{\mu}_i(s, x) = \sum_{\tau=1}^{N(s)} \alpha(1-\alpha)^{\tau-1} \widetilde{V}_{i,t(N(s)-\tau+1)}(s, x). \tag{IA.4.15}$$

We then use the analog of equation (IA.4.12) to compute the estimates of $\widehat{MSE}_i(s, x)$ and $\widehat{\sigma}_i^2(s, x)$ for each $x \in \mathcal{X}$ and average over $x \in \mathcal{X}$ to obtain the estimates of $\widehat{MSE}_i(s)$ and $\widehat{\sigma}_i^2(s)$. The average percentage bias squared is computed using equation (IA.4.13).

Test Results. In Figure IA.1, we examine the test results in the environment with a significant presence of information-insensitive investors (i.e., $\xi = 500$). Panels A and B present the test results in the environment with high noise trading risk (i.e., $\sigma_u = 10^2$). Specifically, panel A presents the average bias based on the test for experience-based equilibrium. It is shown that the average bias is fairly small, between 0.16% and 0.19% across the $N_{sim} = 1,000$ simulation sessions. This indicates that the simulation outcomes of informed AI speculators constitute an experience-based equilibrium. Intuitively, informed AI speculators' trading strategies are optimally derived from their estimated Q matrix, which records the rewards associated with each state-action pair (s, x) based on their past experience. By contrast, panel B plots the distribution of the average bias based on the test for restricted experience-based equilibrium. The bias is large, ranging from 4.0%



Note: For each simulation session, we compute the average bias according to equation (IA.4.13). We then plot the distribution of average bias based on the tests for the experience-based equilibrium (EBE) and restricted EBE, respectively, across all $N_{sim} = 1,000$ simulation sessions. Panels A and B represent the environment with high noise trading risk (i.e., $\sigma_u = 10^2$) whereas panels C and D represent the environment with low noise trading risk (i.e., $\sigma_u = 10^{-1}$) The other parameters are set according to the baseline economic environment described in Section 4.2 in the main text.

Figure IA.1: Test results for the experience-based equilibrium and the restricted experience-based equilibrium when $\xi = 500$.

to 9.2% across the $N_{sim} = 1,000$ simulation sessions. This implies that the simulation outcomes of informed AI speculators do not constitute a restricted experience-based equilibrium. The reason is that non-optimal actions (i.e., $x_i \neq \widetilde{x}_i(s)$) for each state s are not frequently visited once informed AI speculators enter the pure exploitation mode, a necessary condition for algorithms to converge. Thus, the outcomes generated by non-optimal actions, $x_i \neq \widetilde{x}_i(s)$ for each state s, are not consistent with the estimated Q-matrix $\widetilde{Q}_i(s, x_i)$ corresponding to these non-optimal actions, after Q-learning algorithms converge. Panels C and D present the test results in the environment with low noise trading risk (i.e., $\sigma_u = 10^{-1}$). Similar to panels A and B, the average bias is fairly small in the test for the experience-based equilibrium, indicating that the simulation outcomes of informed AI speculators constitute an experience-based equilibrium. The average bias is relatively large in the test for the restricted experience-based equilibrium. However, comparing panels B and D, it is clear that the average bias is much smaller in the environment with low noise trading risk.

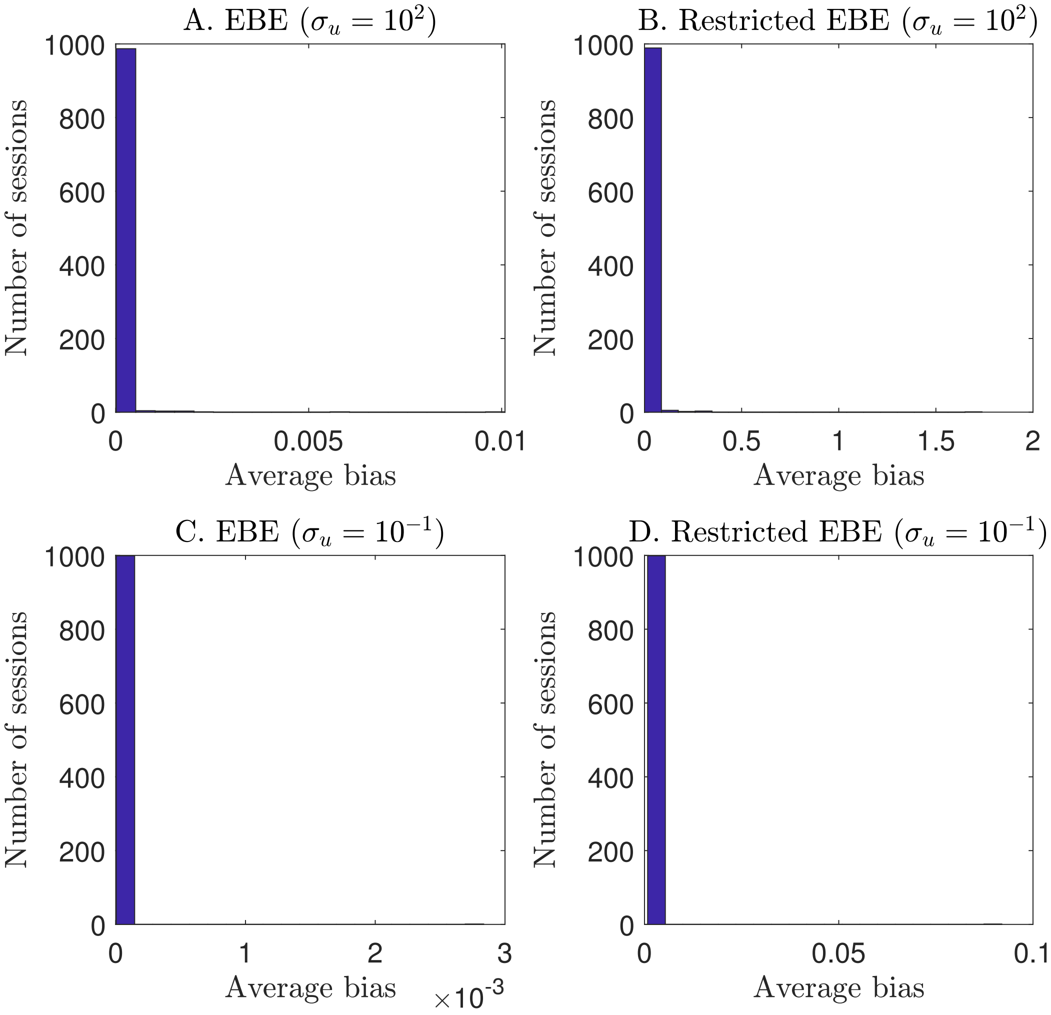
In Figure IA.2, we examine the test results in the environment with an insignificant presence of information-insensitive investors (i.e., $\xi = 5$). The results are similar to Figure IA.1. Regardless of whether the environment has low or high noise trading risk, the simulation outcomes of informed AI speculators constitute an experience-based equilibrium, but not a restricted experience-based equilibrium.

4.3 Discussions of Price-Trigger Strategies of Informed AI Speculators

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

In particular, our economic environment differs from that of Calvano et al. (2020) in two main aspects. First, we consider a stochastic environment where the asset's value v_t in each period is drawn from an i.i.d. distribution. In this stochastic setting, it becomes more difficult for the two informed AI speculators to learn punishment strategies to sustain collusion than in the deterministic setting with a constant v_t.[9] Second, the noise trader's random order flows

[9] 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 algorithms 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.



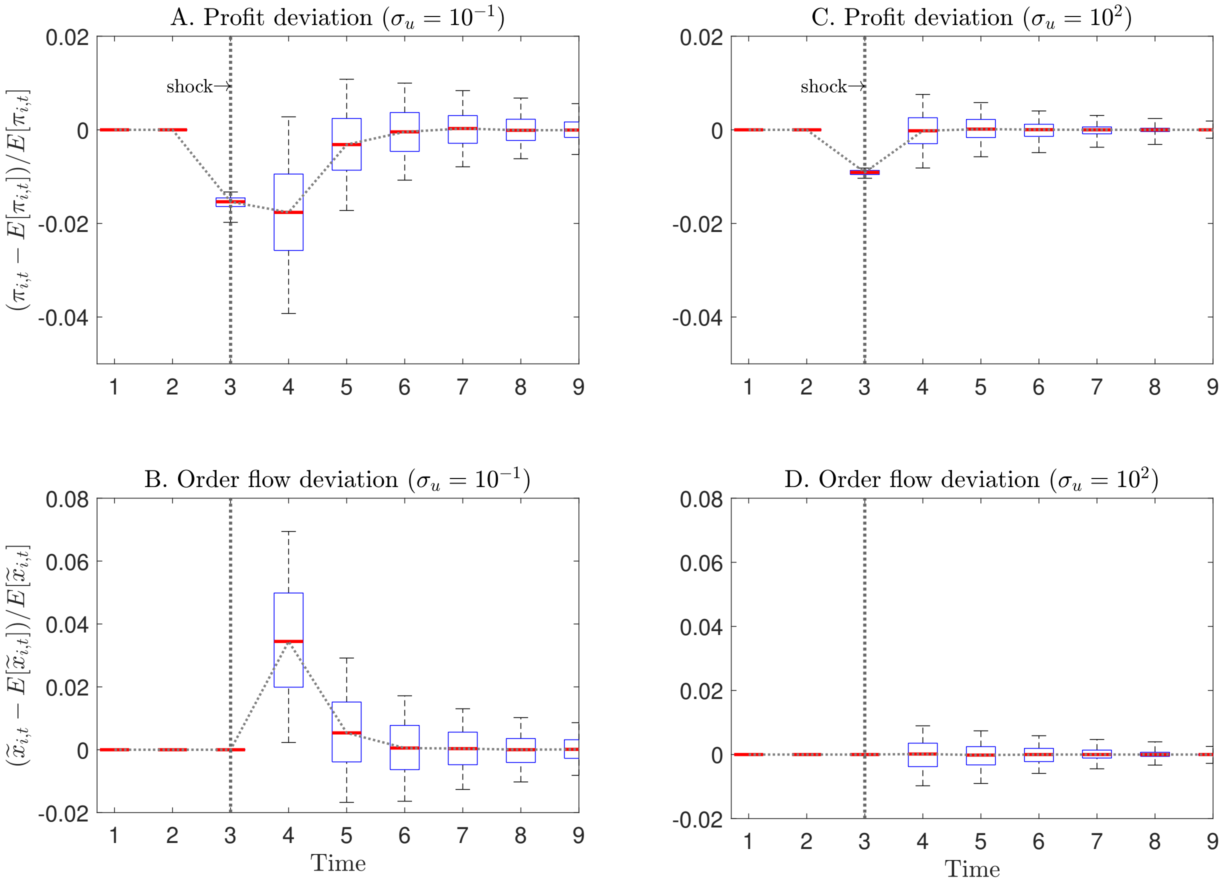
Note: For each simulation session, we compute the average bias according to equation (IA.4.13). We then plot the distribution of average bias based on the tests for the EBE and restricted EBE, respectively, across all $N_{sim} = 1,000$ simulation sessions. Panels A and B represent the environment with high noise trading risk (i.e., $\sigma_u = 10^2$) whereas panels C and D represent the environment with low noise trading risk (i.e., $\sigma_u = 10^{-1}$) The other parameters are set according to the baseline economic environment described in Section 4.2 in the main text.

Figure IA.2: Test results for the experience-based equilibrium and the restricted experience-based equilibrium when $\xi = 5$.

generate information asymmetry to informed AI speculators, which makes grim-trigger strategies infeasible. As a result, informed AI speculators have to adopt price-trigger strategies to collude. In both the model and the simulation experiments with informed AI speculators, the value of σ_u plays a crucial role in determining the level of collusion.

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

When implementing the Q-learning algorithms, we accelerate the convergence speed by computing the continuation value based on the expected value of v_{t+1} in period $t+1$, rather than the realized value of v_{t+1}. This acceleration method is without loss of generality because informed AI speculators can easily learn the distribution of v_t, which is an exogenous state variable. This method also follows the idea in the reinforcement learning literature (e.g., Sutton, 1991; Kearns and Koller, 1999; Kearns and Singh, 2002; Senda, Fujii and Mano, 2006; Azar et al., 2011; Gu et al., 2016). All the results reported in this paper are robust if the reazlied value of v_{t+1} is used to compute the continuation value, but the total computation time will be much longer.



Note: The experiment is similar to that described for the "medium deviation" case in Figure 4 in the main text. Panels A and B the percentage deviation of profit and order flow from their long-run mean for one informed AI speculator, respectively, in the environment with low noise trading risk (i.e., $\sigma_u = 10^{-1}$). Panels C and D focus on the environment with high noise trading risk (i.e., $\sigma_u = 10^2$). 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 $N_{sim} = 1,000$ simulation sessions. The other parameters are set according to the baseline economic environment described in Section 4.2 in the main text.

Figure IA.3: Confidence intervals for the IRF after an exogenous shock to u_t.

4.4 Distribution of IRFs

In Figure IA.3, we conduct the same experiment as that in the "medium deviation" case in panels A to C of Figure 4 in the main text, with $\xi = 500$. However, rather than plotting the average impulse response, we plot the distribution of impulse responses across the $N_{sim} = 1,000$ simulation sessions. Panels A and B show that in the environment with low noise trading risk (i.e., $\sigma_u = 10^{-1}$), the $[25\%, 75\%]$ and $[5\%, 95\%]$ confidence intervals indicate that price-trigger strategies are consistently adopted by informed AI speculators, although the magnitudes of the deviations in prices and trading flows differ significantly across simulation sessions. By contrast, panels C and D show that in the environment with high noise trading risk (i.e., $\sigma_u = 10^2$), the percentage deviation of profit and order flow from their long-run mean is virtually zero for $t \geq 4$ across almost all simulation sessions.

4.5 Identifying Collusion Mechanisms Through IRFs

To distinguish whether a simulation session converges to either mechanism of collusion, we rely on the properties that emerge from the impulse response analysis conducted in Section 5 in the main text. Specifically, conditional on each value of $\log \sigma_u$ along the x-axis in panel A of Figure 2 in the main text, for each of the $N_{sim} = 1,000$ simulation sessions, we construct the impulse response function to an unexpected exogenous shock u_{shock}. For comparisons, across the experiments with different values of $\log \sigma_u$, we calibrate the value of u_{shock} so that the price deviation in $\widetilde{p}_t$ at $t = 3$ is always equal to 1.2%, similar to that in the scenario with "medium deviation" in panel A of Figure 3 in the main text.

Based on the IRF function, we define a simulation session as converging to the steady state in which informed AI speculators achieve collusion via price-trigger strategies if the order flows of both informed AI speculators increase significantly at $t = 4$, as if they are punishing the price deviation observed at $t = 3$. That is, $(\widetilde{x}_{i,t} - \mathbb{E}[\widetilde{x}_{i,t}])/\mathbb{E}[\widetilde{x}_{i,t}] > \overline{x}$ at $t = 4$ for $i = 1, 2$, where $\overline{x}$ is a sufficiently high threshold. We define a simulation session as converging to the steady state in which informed AI speculators achieve collusion via over-pruning bias in learning if the order flows of both informed AI speculators change insignificantly at $t = 4$. That is, $|\widetilde{x}_{i,t} - \mathbb{E}[\widetilde{x}_{i,t}]/\mathbb{E}[\widetilde{x}_{i,t}]| < \underline{x}$ at $t = 4$ for $i = 1, 2$, where $\underline{x}$ is a sufficiently low threshold. For illustration purposes, we set $\overline{x} = 10\underline{x}$ and $\underline{x} = 5 \times 10^{-5}$.

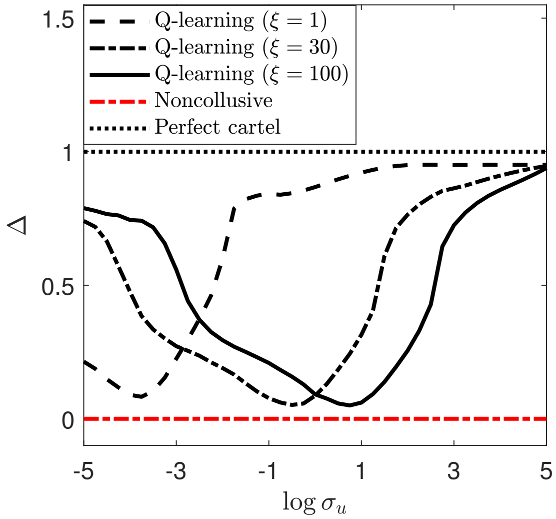
4.6 Robustness of the U-shaped Relationship

Complementary to panel A of Figure 2 in the main text, Figure IA.4 demonstrates that the U-shaped relationship between the average Δ^C and $\log \sigma_u$ is robust across different values of ξ.

4.7 Trading Strategy of Informed AI Speculators

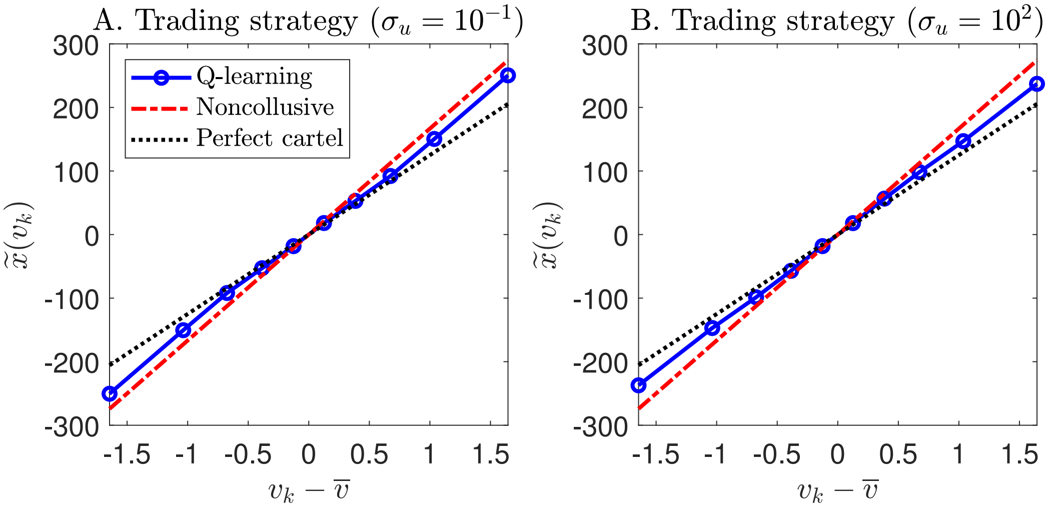
We illustrate informed AI speculators' trading strategies after their Q-learning algorithms converge. Figure IA.5 presents the average trading strategy of informed AI speculators across $N_{sim} = 1,000$ simulation sessions. Panels A and B focus on the environment with low (i.e., $\sigma_u = 10^{-1}$) and high (i.e., $\sigma_u = 10^{2}$) noise trading risk, respectively, with $\xi = 500$. The trading strategy in each simulation session is calculated as $\widetilde{x}(v_k) = \frac{1}{In_pn_v}\sum_{i=1}^{I}\sum_{m=1}^{n_p}\sum_{l=1}^{n_v}\widetilde{x}_i(p_m, v_l, v_k)$, which is the average order flow of I informed AI speculators across all grid points of lagged asset price $\mathbb{P}$ and lagged asset value $\mathbb{V}$, after Q-learning algorithms converge. The dots on the blue solid lines represent the average order flow corresponding to the discrete grid points of $\mathbb{V}$. The black dotted and red dash-dotted lines represent the theoretical benchmarks, $\chi^M(v_k - \overline{v})$ and $\chi^N(v_k - \overline{v})$, in the perfect cartel equilibrium and noncollusive Nash equilibrium, respectively.

It is clear that informed AI speculators learn an optimal trading strategy that is roughly linear in the asset's value after their Q-learning algorithms converge, even though the linearity restriction is not imposed during the learning process. Moreover, the slope of a linear fit for the trading



Note: This figure plots the average Δ^C of all $N_{sim} = 1,000$ simulation sessions as $\log \sigma_u$ varies along the x-axis, for the cases of $\xi = 1, 30$, and 100. The red dash-dotted and black dotted lines represent the theoretical benchmarks of the noncollusive Nash and perfect cartel equilibria, respectively.

Figure IA.4: The U-shaped relationship is robust across environments with different values of ξ.



Note: The trading strategy in each simulation session is calculated as $\widetilde{x}(v_k) = \frac{1}{In_pn_v} \sum_{i=1}^{I} \sum_{m=1}^{n_p} \sum_{l=1}^{n_v} \widetilde{x}_i(p_m, v_l, v_k)$, which is the average order flow of I informed AI speculators across all grid points of lagged asset price $\mathbb{P}$ and lagged asset value $\mathbb{V}$, after Q-learning algorithms converge. The dots on the blue solid lines represent the average order flow corresponding to the discrete grid points of $\mathbb{V}$. Panels A and B focus on the environments with low (i.e., $\sigma_u = 10^{-1}$) and high (i.e., $\sigma_u = 10^{2}$) noise trading risk, respectively, with $\xi = 500$. The other parameters are set according to the baseline economic environment described in Section 4.2 in the main text.

Figure IA.5: The trading strategy of informed AI speculators.

strategy of informed AI speculators, i.e., $\widehat{\chi}^C$, lies between χ^M and χ^N. Thus, the trading strategy learned by informed AI speculators is more conservative than that in the noncollusive Nash equilibrium, which explains why informed AI speculators are able to obtain supra-competitive profits.

4.8 Impacts of Noise Trading Risk

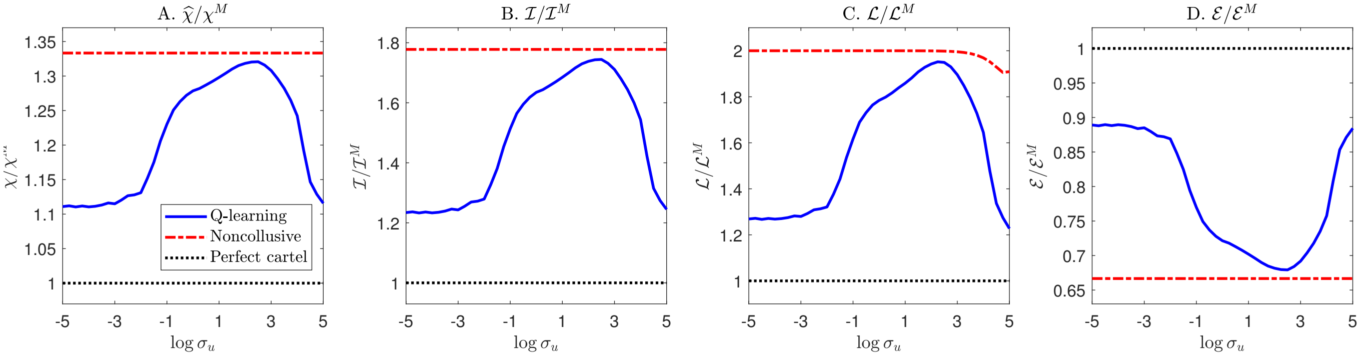
We complement the analysis of noise trading risk in Figure 2 in the main text by studying their impacts on informed AI speculators' trading strategies, price informativeness, market liquidity, and mispricing in financial markets. We show that in the baseline trading environment with $\xi = 500$, AI collusion leads to more conservative trading strategies, lower price informativeness, lower market liquidity, and higher mispricing, with the magnitude depending on the extent to which informed AI speculators collude with each other, which is largely determined by the noise trading risk σ_u.

Panel A of Figure IA.6 plots the average trading policy relative to the theoretical benchmark of the perfect cartel equilibrium, $\widehat{\chi}^C/\chi^M$, across $N_{sim} = 1,000$ simulation sessions as a function of $\log \sigma_u$. The trading policy determines the sensitivity of informed AI speculators' order flow to the asset's value. Consistent with Figure 2 in the main text, $\widehat{\chi}^C/\chi^M$ displays an inverted U-shape as $\log \sigma_u$ increases along the x-axis. By contrast, the relative sensitivity in the two theoretical benchmarks χ^N/χ^M stays roughly unchanged as $\log \sigma_u$ varies.

Panel B of Figure IA.6 plots the market's price informativeness relative to the theoretical benchmark of the perfect cartel equilibrium. By definition, the black dotted line shows that the relative price informativeness in the perfect cartel equilibrium is $\mathcal{I}^M/\mathcal{I}^M \equiv 1$. The red dash-dotted line shows that the ratio of price informativeness in the theoretical benchmark of the noncollusive Nash equilibrium and perfect cartel equilibrium, $\mathcal{I}^N/\mathcal{I}^M$, is greater than 1 and stays unchanged as $\log \sigma_u$ varies. The blue solid line plots the average relative price informativeness, $\mathcal{I}^C/\mathcal{I}^M$, across $N_{sim} = 1,000$ simulation sessions with informed AI speculators. Its value is close to the relative price informativeness in the theoretical benchmark of the non-collusive equilibrium when $\log \sigma_u$ is around 2 due to the lack of collusion. When $\log \sigma_u$ is very small or very large, the relative price informativeness in our simulation experiments with informed AI speculators is significantly lower than that in the theoretical benchmark of the noncollusive Nash equilibrium. The reason is that informed AI speculators place orders in a more conservative manner, with $\widehat{\chi}^C < \chi^N$, as shown in panel A of Figure IA.6.

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

Panel C of Figure IA.6 plots the market liquidity relative to the theoretical benchmark of the perfect cartel equilibrium. The red dash-dotted line shows that the ratio of market liquidity in the theoretical benchmark of the noncollusive Nash equilibrium and perfect cartel equilibrium,



Note: Panels A, B, C, and D plot the average trading policy, price informativeness, market liquidity, and mispricing relative to the theoretical benchmark of the perfect cartel equilibrium, i.e., $\widehat{\chi}/\chi^M$, $\mathcal{I}/\mathcal{I}^M$, $\mathcal{L}/\mathcal{L}^M$, and $\mathcal{E}/\mathcal{E}^M$, respectively, across $N_{sim} = 1,000$ simulation sessions as $\log \sigma_u$ varies. The blue solid line represents the simulation experiments with informed AI speculators; the red dash-dotted and black dotted lines represent the theoretical benchmarks of the noncollusive Nash and perfect cartel equilibrium, respectively. The other parameters are set according to the baseline economic environment described in Section 4.2 in the main text.

Figure IA.6: $\widehat{\chi}/\chi^M$, $\mathcal{I}/\mathcal{I}^M$, $\mathcal{L}/\mathcal{L}^M$, and $\mathcal{E}/\mathcal{E}^M$ for $\log \sigma_u \in [-5, 5]$.

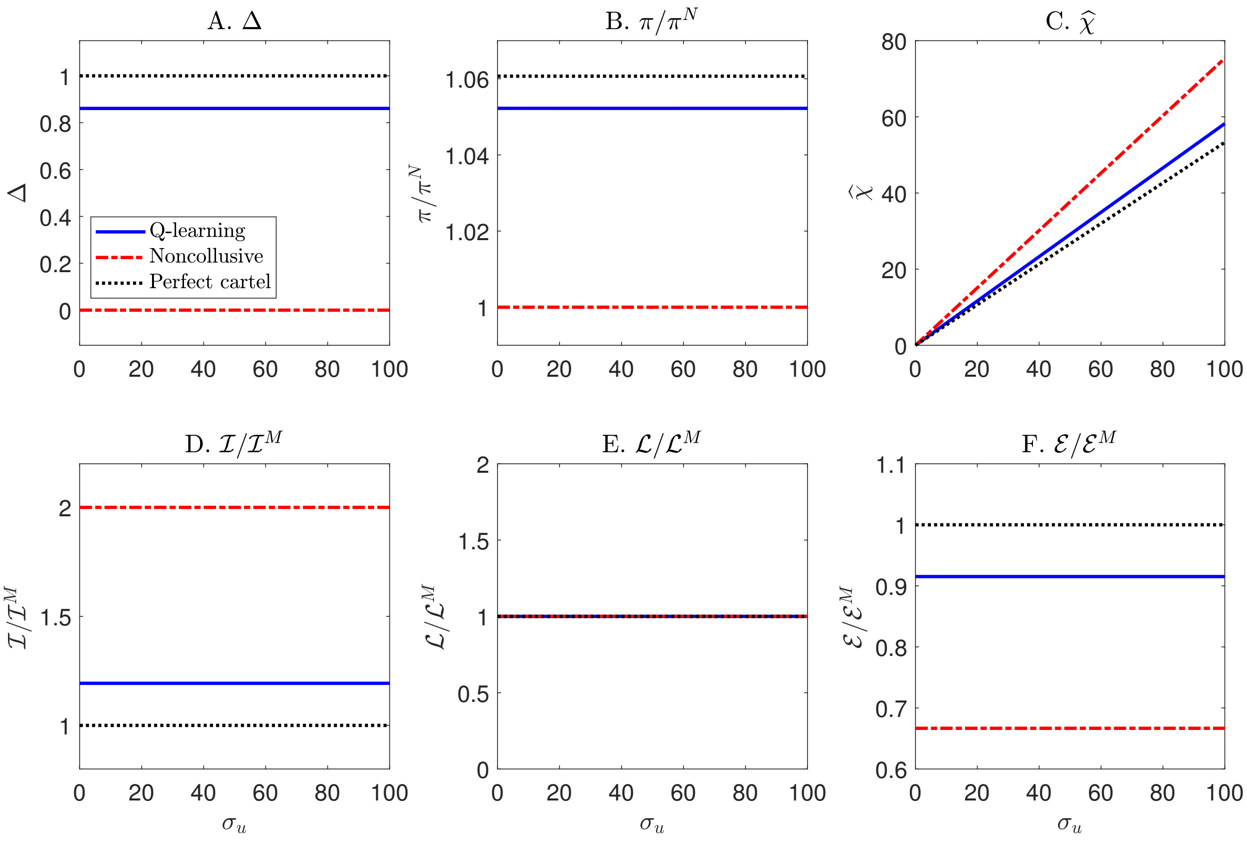
$\mathcal{L}^N/\mathcal{L}^M$, is greater than 1 and decreasing in $\log \sigma_u$.[10] The blue solid line shows that the market liquidity in our simulation experiments with informed AI speculators is higher than that in the theoretical benchmark of the perfect cartel equilibrium and lower than that in the theoretical benchmark of the noncollusive equilibrium. The blue solid line displays an inverted U shape similar to panel A of Figure IA.6, indicating that the market liquidity is closer to the theoretical benchmark of the perfect cartel equilibrium if there is a higher degree of collusion among informed AI speculators.

Panel D of Figure IA.6 plots the average mispricing relative to the theoretical benchmark of the perfect cartel equilibrium. Mispricing is higher in the theoretical benchmark of the perfect cartel equilibrium (the black dotted line) than that in the noncollusive equilibrium (the red dash-dotted line). The blue solid line shows that AI collusion increases mispricing, and the magnitude is larger when there is a higher degree of collusion among informed AI speculators.

4.9 Environments with Efficient Prices

We now study the behavior of informed AI speculators in the economic environment with $\xi = 0$, where information-insensitive investors are absent. Consequently, the market maker sets prices solely to maximize price discovery, meaning $p_t = \mathbb{E}[v_t | y_t]$. This economic environment is similar to Kyle (1985) but features $I = 2$ informed speculators engaging in a repeated trading game. Proposition 3.1 in Section 3 in the main text indicates that, theoretically, tacit collusion cannot be sustained by price-trigger strategies in this environment due to efficient prices. Figure IA.7

[10] This is because $\lambda^N < \lambda^M$ for all $\log \sigma_u$. Intuitively, in the perfect cartel equilibrium, the market maker knows that informed speculators submit orders jointly like a monopoly, and thus the market maker adopts a pricing rule that is more responsive to the combined order flow of informed speculators and the noise trader, i.e., $\gamma^N < \gamma^M$, resulting in $\lambda^N < \lambda^M$. As $\log \sigma_u$ increases, both λ^N and λ^M decline, so that market liquidity defined by equation (IA.4.6) increases.



Note: We consider the economic environment with efficient prices as in Kyle (1985). That is, we set $\xi = 0$, implying that the asset's price p_t is determined to minimize pricing errors, with $p_t = \mathbb{E}[v_t|y_t]$. The blue solid line plots the average values of Δ^C, π^C/π^N, $\widehat{\chi}^C$, $\mathcal{I}^C/\mathcal{I}^M$, $\mathcal{L}^C/\mathcal{L}^M$, and $\mathcal{E}^C/\mathcal{E}^M$ across $N_{sim} = 1,000$ simulation sessions as σ_u varies. The red dash-dotted and black dotted lines represent the theoretical benchmarks of the noncollusive Nash equilibrium and perfect cartel equilibrium, respectively. The other parameters are set according to the baseline economic environment described in Section 4.2 in the main text, except for $\xi = 0$.

Figure IA.7: Implications of noise trading risk in the environment with $\xi = 0$.

presents the average results across $N_{sim} = 1,000$ simulation paths with informed AI speculators. The blue solid lines in Panels A and B show that informed AI speculators can achieve an average Δ^C of approximately 0.86, and their average profit is about 5% higher than that in the theoretical benchmark of the non-collusive equilibrium.

As discussed in Section 5.1 in the main text, collusion in this environment is achieved through over-pruning bias in learning. Similar to the Kyle (1985) model, the profits of informed speculators in the theoretical benchmark of the non-collusive Nash equilibrium and the perfect cartel benchmark are linear in the noise trading risk σ_u. Consequently, as expected, the red dash-dotted and black dotted lines in panels A and B are flat. Moreover, the collusive equilibrium formed by informed AI speculators also has a constant Δ^C and π^C/π^N as σ_u varies along the x-axis, demonstrating a similar scaling property with respect to σ_u. Panel C shows that the sensitivity of informed AI speculators' order flows to the value of the asset $\widehat{\chi}^C$ increases linearly with σ_u. This scaling property with respect to σ_u mirrors that in the theoretical benchmarks of the non-collusive

Nash equilibrium and the perfect cartel benchmark, a property that also holds in the Kyle (1985) model.

Panel D shows that due to collusion, price informativeness in the environment with informed AI speculators is lower than that in the theoretical benchmark of the non-collusive Nash equilibrium, but higher than that in the perfect cartel benchmark. Moreover, as in the Kyle (1985) model, price informativeness remains unchanged as σ_u varies along the x-axis. Panel E shows that market liquidity is always equal to 1, that is, $p_t = \mathbb{E}[v_t|y_t]$. This can be directly seen from equation (IA.4.6). In the absence of information-insensitive investors, the inventory of the market maker is equal to $-y_t \equiv -\sum_{i=1}^{I} x_{i,t} - u_t$. Thus, the sensitivity of the market maker's inventory to noise order flows is always 1, which holds regardless of the level of noise trading risk. Panel F shows that mispricing in the environment with informed AI speculators is higher than that in the theoretical benchmark of the non-collusive Nash equilibrium, but lower than that in the perfect cartel benchmark.

4.10 Computing Q-loss

We test the extent to which an AI speculator's limit strategy after Q-learning algorithms converge constitutes an optimal response to that of the rival informed AI speculators.

To perform this test, in each of the $N_{sim} = 1,000$ simulation sessions, for each informed AI speculator i, we calculate its theoretical Q-matrix under the assumption that its rival informed AI speculators use their limit strategies. Specifically, the theoretical Q-matrix is characterized as follows:

$$Q_i^*(s_t, x_{i,t}) = \mathbb{E}\left[\pi_{i,t}|s_t, x_{i,t}\right] + \rho\mathbb{E}\left[\max_{x'\in\mathbb{X}} Q_i^*(s_{t+1}, x')\middle| s_t, x_{i,t}\right], \tag{IA.4.16}$$

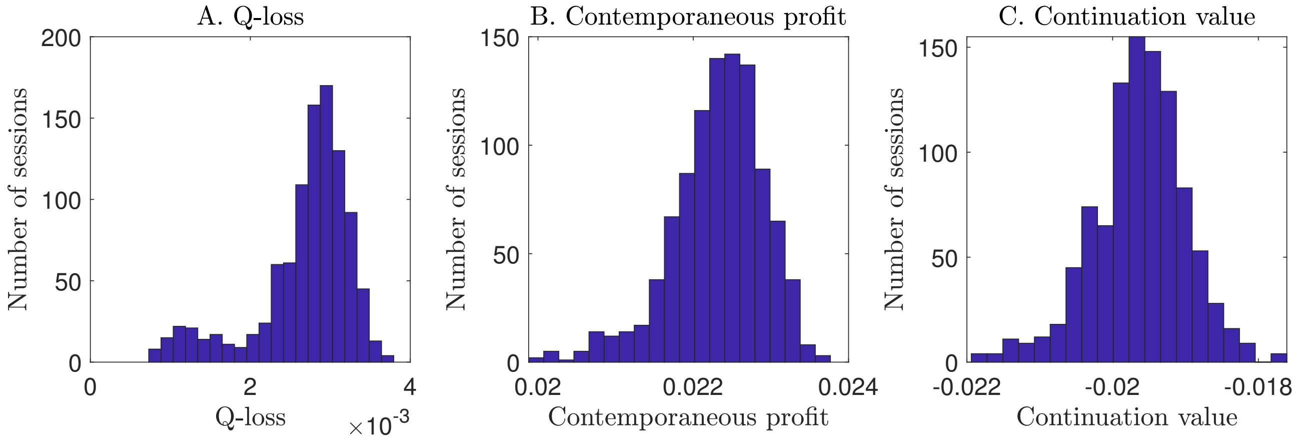
where $s_t = \{p_{t-1}, v_{t-1}, v_t\}$. The first expectation is taken over the noise order flow u_t, and the second expectation is taken over u_t and v_{t+1}. The price p_t and profit $\pi_{i,t}$ are computed by

$$\begin{aligned} p_t =& \widehat{\gamma}_0 + \widehat{\lambda} y_t, \\ \pi_{i,t} =& (v_t - p_t)x_{i,t}, \end{aligned}$$

where $\widehat{\gamma}_0 = \sum_{t=T_c}^{T_c+T} \widehat{\gamma}_{0,t}$ and $\widehat{\lambda} = \sum_{t=T_c}^{T_c+T} \widehat{\lambda}_t$ are simple averages of $\widehat{\gamma}_{0,t}$ and $\widehat{\lambda}_t$ over $T = 100,000$ periods after Q-learning algorithms reach convergence at T_c, with $\widehat{\gamma}_{0,t}$ and $\widehat{\lambda}_t$ estimated by equation (4.2) in the main text. The total order flow is $y_t = x_{i,t} + \sum_{j=1,j\neq i}^{I} \widetilde{x}_j(s_t) + u_t$, where $\widetilde{x}_j(s_t)$ is the limit strategy of informed AI speculator j after Q-learning algorithms converge.

The recursive problem (IA.4.16) is solved using a standard approach of value function iterations based on the same discrete grids, $\mathbb{P} \times \mathbb{V} \times \mathbb{V} \times \mathbb{X}$, as the Q-learning programs. With $\rho < 1$, convergence of value function iterations is guaranteed by the contraction mapping theorem.

After computing the theoretical Q-matrix from equation (IA.4.16), in each state $s \in \mathbb{P} \times \mathbb{V} \times \mathbb{V}$, we pin down the optimal response of informed AI speculator i to its rival's limit strategies, denoted by $x_i^*(s)$. Then, we compute the forfeited payoff of informed AI speculator i for playing its limit



Note: Panel A reports the average Q-loss across all $N_{sim} = 1,000$ simulation sessions. Panels B and C decompose the average Q-loss into the loss in contemporaneous profit and continuation value, based on the first and second terms on the right-hand side of equation (IA.4.16), respectively. We set $\sigma_u = 10^{-1}$. The other parameters are set according to the baseline economic environment described in Section 4.2 in the main text.

Figure IA.8: Q-loss distribution in the environment with low noise trading risk (i.e., $\sigma_u = 10^{-1}$) and significant presence of information-insensitive investors (i.e., $\xi = 500$).

strategy $\widetilde{x}_i(s)$ in state s as follows:

$$Q_{i,loss}(s) = \frac{Q_i^*(s, x_i^*(s)) - Q_i^*(s, \widetilde{x}_i(s))}{Q_i^*(s, x_i^*(s))}, \tag{IA.4.17}$$

which is referred to as the Q-loss. If $\widetilde{x}_i(s) = x_i^*(s)$, the resulting Q-loss would be zero, meaning that informed AI speculator i learned to play the optimal response to the rivals' limit strategies in state s. For any non-optimal responses, the resulting Q-loss would be positive, with a larger magnitude reflecting more ineffective learning. In other words, the Q-loss reflects the deviation of limit strategies from those in a Nash equilibrium.

We calculate the Q-loss along the equilibrium path for all $N_{sim} = 1,000$ simulation sessions. Specifically, starting with an arbitrary state s_0, we use the limit trading policies $\{\widetilde{x}_i(s)\}_{i=1}^{I}$ to simulate a sample path $\{s_t\}_{t=1}^{T_0+T}$, with sufficiently large T_0 and T. The first T_0 periods are dropped as burn-in. In our test, we set $T_0 = 100,000$ and $T = 1,000,000$. We further verify that setting larger values for T_0 or T will not alter any results. For each informed AI speculator $i = 1, ..., I$ and each period $t = T_0 + 1, ..., T_0 + T$, we compute the Q-loss using equation (IA.4.17). Then, we compute the simple average across all I informed AI speculators and T periods to obtain the Q-loss of each simulation session.

Panel A of Figure IA.8 plots the distribution of the Q-losses across $N_{sim} = 1,000$ simulation sessions in the environment with low noise trading risk ($\sigma_u = 10^{-1}$) and significant presence of information-insensitive investors (i.e., $\xi = 500$). The largest and average Q-loss across all sessions are 0.38% and 0.27%, respectively, indicating that the AI speculators' limit strategies are close to optimal strategies. We further decompose the Q-loss into the loss in contemporaneous profit and continuation value as follows:

$$Profit_{i,loss}(s) = \frac{\mathbb{E}[\pi_{i,t}|s, x_i^*(s)] - \mathbb{E}[\pi_{i,t}|s, \widetilde{x}_i(s)]}{Q_i^*(s, x_i^*(s))}, \tag{IA.4.18}$$

$$Cont\_val_{i,loss}(s) = \frac{\rho\mathbb{E}\left[\max_{x' \in \mathbb{X}} Q_i^*(s', x')\big|s, x_i^*(s)\right] - \rho\mathbb{E}\left[\max_{x' \in \mathbb{X}} Q_i^*(s', x')\big|s, \widetilde{x}_i(s)\right]}{Q_i^*(s, x_i^*(s))}. \tag{IA.4.19}$$

Panels B and C of Figure IA.8 show that the loss in contemporaneous profit is positive, with an average value of 2.33%, whereas the loss in continuation value is negative, with an average value of −1.96%. This indicates that given the rivals playing their limit strategies, each informed AI speculator should act more aggressively when placing their trading orders. In this way, the contemporaneous profit would increase, which outweighs the decrease in continuation value (caused by rivals playing punishment strategies after a price trigger). In other words, the results in Figure IA.8 suggest that the punishment strategies learned by informed AI speculators are not sufficiently significant to rule out all profitable deviations.

4.11 Q-Learning Market Maker

In the baseline economic environment, the market maker analyzes historical data to estimate the pricing rule. We now consider the market maker adopting Q-learning algorithms to learn the pricing rule. We find that all the results presented in the main text remain unchanged; they do not depend on whether the market maker determines the pricing rule using statistical learning or Q-learning algorithms.

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

$$p_t = v_t^{MM} + \lambda_t^{MM} y_t, \tag{IA.4.20}$$

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

$$\begin{aligned}\widehat{Q}_{t+1}^{MM}(v_t^{MM}, \lambda_t^{MM}) =& (1 - \alpha^{MM})\widehat{Q}_t^{MM}(s_t, a_t) + \alpha \bigg[(y_t - \xi(v_t^{MM} - \overline{v} + \lambda_t^{MM} y_t))^2 \\ & + \theta(v_t^{MM} + \lambda_t^{MM} y_t - v_t)^2 + \rho^{MM} \min_{v' \in \mathcal{V}, \lambda' \in \Lambda} \widehat{Q}_t^{MM}(v', \lambda')\bigg],\end{aligned} \tag{IA.4.21}$$

where the reward in period t is

$$\begin{aligned}(y_t + z_t)^2 + \theta(p_t - v_t)^2 =& (y_t - \xi(p_t - \overline{v}))^2 + \theta(p_t - v_t)^2 \\ =& (y_t - \xi(v_t^{MM} - \overline{v} + \lambda_t^{MM} y_t))^2 + \theta(v_t^{MM} + \lambda_t^{MM} y_t - v_t)^2.\end{aligned} \tag{IA.4.22}$$

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

To implement the Q-learning algorithm for the market maker, we construct discrete grids 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 $n_{\overline{v}}$ and n_λ equally spaced grid points, i.e., $\mathbb{V} = \{v_1^{MM}, ..., v_{n_{\overline{v}}}^{MM}\}$ and $\Lambda = \{\lambda_1^{MM}, ..., \lambda_{n_\lambda}^{MM}\}$. The parameters v^{MM} and λ^{MM} correspond to the optimal values in the theoretical benchmark of the noncollusive equilibrium. The parameter $\kappa > 0$ ensures that the values of v_t and λ_t chosen by the market maker can be different from the theoretical values, v^{MM} and λ^{MM}.

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

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

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

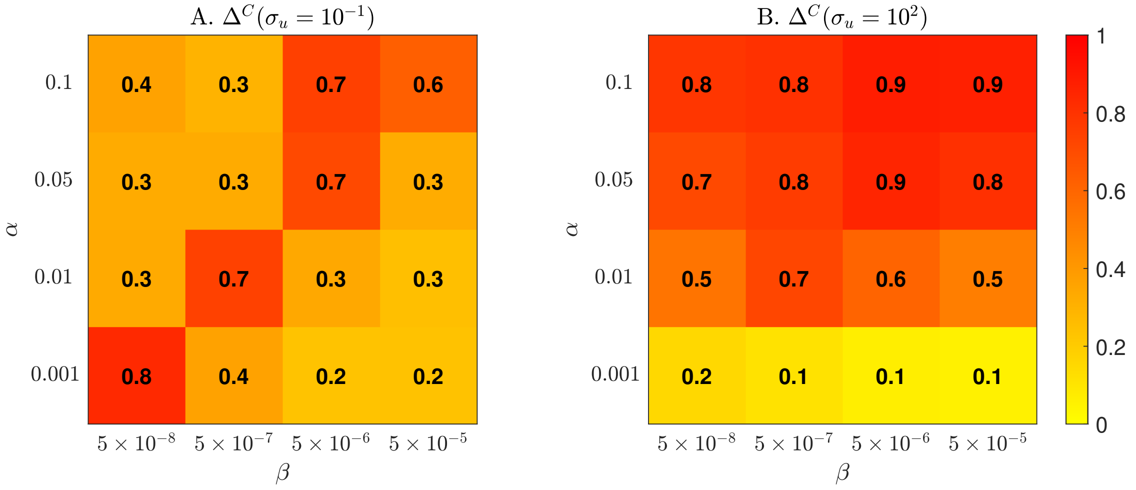
$$\begin{aligned}\widehat{Q}_0^{MM}(v_k^{MM}, \lambda_j^{MM}) = & \frac{1}{1-\rho^{MM}}\left[(1-\xi\lambda_j^{MM})^2((I\chi^N\sigma_v)^2 + \sigma_u^2) + \xi^2(v_k^{MM} - \overline{v})^2\right] \\ & + \frac{\theta}{1-\rho^{MM}}\left[(v_k^{MM} - \overline{v})^2 + (\lambda_j^{MM} I\chi^N - 1)^2\sigma_v^2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

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

4.12 Hyperparameters

Hyperparameters α and β. The behavior of Q-learning algorithms is governed by two key hyperparameters: α, which controls the forgetting rate, and β, which determines the rate at which exploration decays over time. Panels A and B of Figure IA.9 show the average Δ^C for varying α and β in environments with low ($\sigma_u = 10^{-1}$) and high ($\sigma_u = 10^2$) noise trading risks, respectively. In Panel A, where AI collusive equilibrium through price-trigger strategies prevails, the collusive trading profitability Δ^C remains robustly high across hyperparameter combinations but peaks along the diagonal line. This suggests that efficient learning to achieve AI collusive equilibrium through price-trigger strategies requires striking a balance between α (the forgetting rate) and β (the exploration decay rate), which ensures the effectiveness of the exploration-exploitation tradeoff in figuring out the optimal trading strategies. In other words, efficient learning cannot be obtained by tuning an individual hyperparameter value of α or β.

In Panel B, where AI collusive equilibrium through over-pruning bias in learning prevails, the



Note: Panel A shows Δ^C in an environment with low noise trading risk ($\sigma_u = 10^{-1}$), while Panel B displays Δ^C in an environment with high noise trading risk ($\sigma_u = 10^2$). All other parameters follow the baseline economic environment described in Section 4.2.

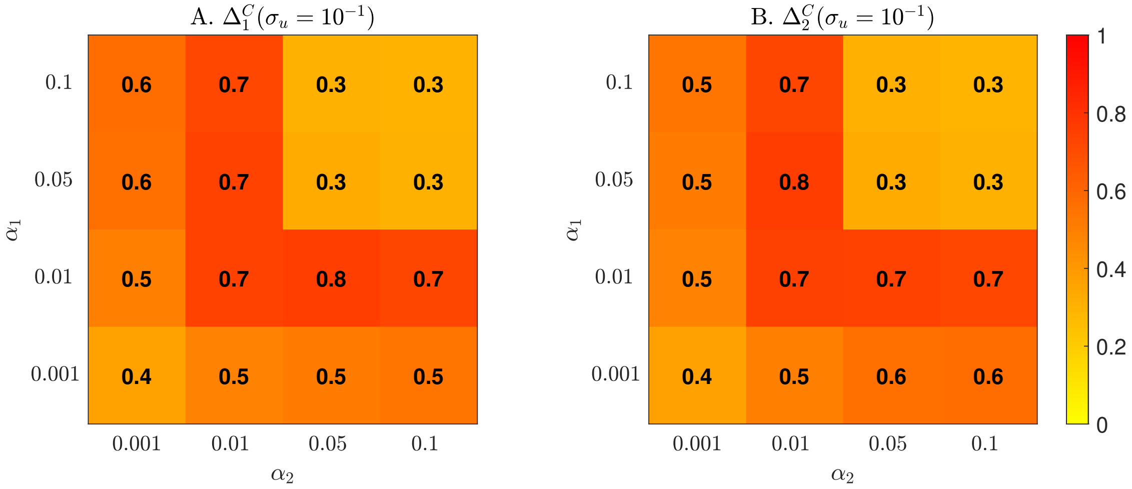
Figure IA.9: Implications of hyperparameters α and β on Δ^C.

collusive trading profitability Δ^C remains robustly high across hyperparameter combinations, as long as α is not very close to zero. Importantly, for a fixed α, the exploration decay rate β has little impact on the average collusive trading profit. In contrast, for a fixed β, the average trading profit Δ^C increases with α. This occurs because a higher α amplifies the over-pruning bias, reducing the algorithm's learning efficiency and reinforcing collusive behavior through this bias.

Heterogeneous Forgetting Rates (α_1 and α_2). The algorithm with a lower α exhibits smaller learning biases but requires more time and computational resources for training. In this context, α can be interpreted as the "intelligence level" of the algorithm: a lower α indicates a more advanced algorithm capable of more accurate learning.

This subsection conducts simulation experiments within the baseline economic environment using standard Q-learning algorithms with heterogeneous, fixed values of α. In Online Appendix 4.13, we extend this approach by introducing a two-tier Q-learning algorithm with an adaptive α, a form of Meta Q-learning. This extension enables informed AI speculators to learn not only the trading strategies corresponding to a given α but also the optimal α values themselves. Our results reveal that informed AI speculators using two-tier Q-learning algorithms can strategically coordinate their choices of α at high levels (i.e., low "intelligence levels") to maximize collective benefits. Notably, the intuition behind the tacit collusion on α observed in two-tier Meta Q-learning algorithms is already evident in the simulation experiments with heterogeneous, fixed α values presented here.

Focusing on the baseline calibration, we allow the two informed AI speculators to employ Q-learning algorithms with varying intelligence levels, represented by distinct values of α. Specifically, each informed AI speculator i adopts an algorithm whose forgetting rate is α_i, with

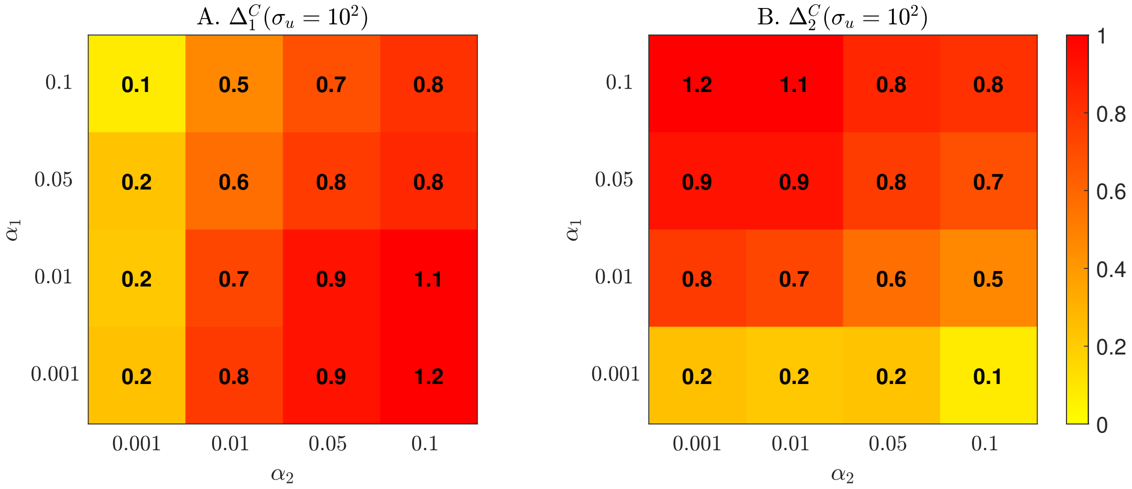


Note: Panels A and B show Δ_1^C and Δ_2^C, respectively, in the baseline environment with $\sigma_u = 10^{-1}$.

Figure IA.10: The Impact of heterogeneous α_1 and α_2 on Δ^C when $\sigma_u = 10^{-1}$.

$\alpha_i = 0.001, 0.01, 0.05$ and 0.1 for $i = 1, 2$. Panels A and B of Figure IA.10 plot the average Δ_1^C and Δ_2^C for informed AI speculators 1 and 2, respectively, in the environment with low noise trading risks (i.e., $\sigma_u = 10^{-1}$). Below, we highlight several key insights that emerge from these results. First, the substantial average collusive trading profits sustained by price-trigger strategies remain remarkably robust, even in the presence of significant heterogeneity in the algorithms adopted by informed AI speculators. Second, adopting algorithms with overly low intelligence levels, such as $(\alpha_1, \alpha_2) = (0.001, 0.001)$, is suboptimal. Obviously, both speculators achieve significantly higher profits when using $(\alpha_1, \alpha_2) = (0.01, 0.01)$ compared to employing algorithms with higher intelligence levels, such as $(\alpha_1, \alpha_2) = (0.001, 0.001)$. This echos a key observation of Figure IA.9 — efficient learning to achieve AI collusive equilibrium through price-trigger strategies requires striking a balance between α and β, rather than solely tuning one hyperparameter value. In the experiments of Figure IA.10, the value of β is fixed at its baseline calibration level (i.e., $\beta = 5 \times 10^{-7}$), and $\alpha = 0.001$ appears too low and fail to achieve the necessary balance with β, resulting in inefficient learning. Third, adopting algorithms with overly low intelligence levels, such as $(\alpha_1, \alpha_2) = (0.05, 0.05)$ or $(\alpha_1, \alpha_2) = (0.1, 0.1)$, is also suboptimal. With $\beta = 5 \times 10^{-7}$, these values of α fail to achieve the necessary balance between α and β, leading to inefficient learning, as shown in Panel A of Figure IA.9.

Panels A and B of Figure IA.11 plot the average Δ_1^C and Δ_2^C for informed AI speculators 1 and 2, respectively, in the environment with high noise trading risks (i.e., $\sigma_u = 10^2$). Several key insights emerge from these results. First, as in Figure IA.10, the substantial average collusive trading profits sustained by over-pruning bias in learning remain remarkably robust, even with significant heterogeneity in the algorithms adopted by informed AI speculators. Second, the patterns in Figure IA.11 fundamentally differ from those in Figure IA.10, further highlighting the distinction between the underlying mechanisms driving AI collusive equilibrium under high



Note: Panels A and B show Δ_1^C and Δ_2^C, respectively, in the baseline environment with $\sigma_u = 10^2$.

Figure IA.11: The Impact of heterogeneous α_1 and α_2 on Δ^C when $\sigma_u = 10^2$.

versus low noise trading risk scenarios. Third, an informed AI speculator has a strong incentive to adopt an algorithm with a high intelligence level (i.e., low α) when the other speculator uses an algorithm with a low intelligence level (i.e., high α). Importantly, however, both speculators achieve significantly higher profits when using (α_1, α_2) such that $\alpha_i \geq 0.01$ for $i = 1, 2$ compared to employing algorithms with higher intelligence levels, such as $(\alpha_1, \alpha_2) = (0.001, 0.001)$.

The simulation findings on the strong potential for coordination at high levels of α (i.e., low intelligence levels) among informed AI speculators are fundamentally consistent with the general equilibrium effects in active management described by Stambaugh (2020). His model shows that when all managers lack the ability to select positive-alpha stocks, collective profits remain high. However, as a small fraction of managers gains skill, their profits rise at the expense of less skilled managers. If many managers become skilled, profits for all of them would decline due to strengthened price corrections and the resulting reduced alpha. Similarly, Dugast and Foucault (2024) find that improved manager skills, driven by lower information costs or new datasets, reduce average performance as asset prices become more informative.

4.13 Two-Tier Meta Q-Learning Algorithms

Algorithms. Each informed AI speculator i employs a two-tier Meta Q-learning algorithm. In the lower tier, the speculator updates the Q-function $\widehat{Q}_{i,t}(s_t, x_{i,t})$, where $s_t = \{p_{t-1}, v_{t-1}, v_t\}$ is the state and $x_{i,t}$ is the order flow, using a Q-learning algorithm with an adaptive forgetting rate $\alpha_{i,t}$. The lower-tier algorithm follows the structure in Section 4.1 in the main text, with the addition of the adaptive $\alpha_{i,t}$. In the upper tier, the speculator learns the Q-function $\widehat{Q}^u_{i,t}(s^u_{i,t}, \alpha_{i,t})$, where $s^u_{i,t}$ is the state and $\alpha_{i,t}$ is the chosen forgetting rate, optimizing it to improve the lower-tier learning process.

To ensure profits stabilize for any given $\alpha_{i,t}$ in the upper tier, the lower-tier Q-learning algorithm runs for an extended period. Consequently, updates to $\alpha_{i,t}$ occur less frequently than updates to $x_{i,t}$ in the lower tier. Specifically, each informed AI speculator i adjusts $\alpha_{i,t}$ only after completing a lower-tier training epoch of T periods, where T is a large integer. Let $\tau = 1, 2, ...$ denote these training epochs, with epoch τ spanning periods $(\tau - 1)T + 1$ to τT. Within each training epoch τ, the upper-tier Q-function $\widehat{Q}^u_{i,t}(s^u_{i,t}, \alpha_{i,t})$ and action $\alpha_{i,t}$ for each informed AI speculator i remain fixed from $(\tau - 1)T + 1$ to $\tau T - 1$. Updates to $\widehat{Q}^u_{i,t}(s^u_{i,t}, \alpha_{i,t})$ and $\alpha_{i,t}$ occur only at the end of the epoch, at $t = \tau T$. The updated choice of $\alpha_{i,\tau T}$ follows the standard exploration-exploitation procedure.

The recursive learning equation for the upper-tier Q-learning algorithm, updated at the end of each epoch τ, is specified as follows:

$$\widehat{Q}^u_{i,(\tau+1)T}(s^u_{i,\tau T}, \alpha_{i,\tau T}) = (1 - \alpha^u)\widehat{Q}^u_{i,\tau T}(s^u_{i,\tau T}, \alpha_{i,\tau T}) + \alpha^u \left[\pi^u_{i,\tau T} + \rho^u \max_{\alpha' \in \mathcal{A}} \widehat{Q}^u_{i,\tau T}(s^u_{i,(\tau+1)T}, \alpha') \right], \tag{IA.4.23}$$

for $\tau = 1, 2, ...$ In equation (IA.4.23), $\pi^u_{i,\tau T}$ is the reward in the training epoch τ, given by $\pi^u_{i,\tau T} = \frac{1}{T}\sum_{t=(\tau-1)T+1}^{\tau T}(v_t - p_t)x_{i,t}$, which is the average trading profit over the last T periods, from $(\tau - 1)T + 1$ to τT. The parameters α^u and ρ^u are the forgetting rate and the subjective discount rate for the upper tier Q-learning algorithm. For tractability, we choose the state variable $s^u_{i,\tau T} = \{\pi^u_{i,(\tau-1)T}\}$, which is the reward in the previous training epoch. The choice of $\alpha_{i,\tau T}$ is made as follows:

$$\alpha_{i,\tau T} = \begin{cases} \text{argmax}_{\alpha' \in \mathcal{A}} \widehat{Q}^h_{i,\tau T}(s^u_{i,\tau T}, \alpha'), & \text{with prob. } 1 - \varepsilon^u_\tau, \quad \text{(exploitation)} \\ \widetilde{\alpha} \sim \text{uniform distribution on } \mathcal{A}, & \text{with prob. } \varepsilon^u_\tau. \quad \text{(exploration)} \end{cases} \tag{IA.4.24}$$

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

Simulation Results. The two-tier Q-learning algorithm takes a substantially longer time to converge because there are experimentations on both $\alpha_{i,t}$ and $x_{i,t}$. For the upper-tier algorithm, we consider the following parameter values: $\alpha^u = 0.05$, $\beta^u = 10^{-4}$, and $\rho^u = 0.95$. Each training epoch has a total of $T = 10,000,000$ periods. The convergence criterion requires the decisions of $\alpha_{i,t}$ to stay unchanged for $100,000$ consecutive training epochs. For tractability, we choose three grids for the choice of $\alpha_{i,t}$, with $\mathcal{A} = \{0.001, 0.01, 0.1\}$. The parameters and grids for the lower-tier Q-learning algorithm are similar to those described in Section 4 in the main text.

Table IA.I summarizes the number of simulation sessions that converge to each pair of (α_1, α_2) after algorithms converge. In the environment with low noise trading risk (i.e., $\sigma_u = 10^{-1}$), across the $N_{sim} = 1,000$ simulations sessions, 90.7% of the sessions converge to the best equilibrium with $(\alpha_1, \alpha_2) = (0.01, 0.01)$, which maximizes both informed AI speculators' profits. The remaining 9.3% of the sessions converge to either $(\alpha_1, \alpha_2) = (0.1, 0.01)$ or $(\alpha_1, \alpha_2) = (0.01, 0.1)$, which yield similar profits for informed AI speculators compared with $(\alpha_1, \alpha_2) = (0.01, 0.01)$, as shown in

Table IA.I: Adaptive forgetting rates after the convergence of two-tier Q-learning algorithms.

	Low noise trading risk (i.e., $\sigma_u = 10^{-1}$)	High noise trading risk (i.e., $\sigma_u = 10^2$)
$(0.1, 0.1)$	0	0.028
$(0.1, 0.01)$ or $(0.01, 0.1)$	0.093	0.194
$(0.01, 0.01)$	0.907	0.716
$(0.001, 0.001)$	0	0.002
Others	0	0.06

Note: This table reports the proportion of the $N_{sim} = 1,000$ independent simulation sessions that converge to each pair of (α_1, α_2) after the two-tier Q-learning algorithms converge.

Figure IA.10 in the main text. These results suggest that our two-tier Q-learning algorithm enables the two informed AI speculators to learn to play the optimal equilibrium.

Turning to the environment with high noise trading risk (i.e., $\sigma_u = 10^2$). As shown in Figure IA.11 in the main text, the two informed AI speculators face a situation that resembles the prisoner's dilemma. Specifically, given informed AI speculator i's choice of α_i, informed AI speculator j can gain by adopting the smallest $\alpha_j = 0.001$. However, both informed AI speculators would not make much profit if they reach the unqiue Nash equilibrium of $(\alpha_1, \alpha_2) = (0.001, 0.001)$ of a one-shot game. Instead, both of them would attain supra-competitive profits if neither informed AI speculator adopts a forgetting rate of 0.001, that is, when both informed AI speculators adopt unadvanced algorithms to trade. In theory, the equilibria with high values of α can only be sustained in a repeated game. In our simulation experiments, we find that across the $N_{sim} = 1,000$ simulations sessions, 2.8% of the sessions converge to the equilibrium with $(\alpha_1, \alpha_2) = (0.1, 0.1)$, 19.4% of the sessions converge to either $(\alpha_1, \alpha_2) = (0.1, 0.01)$ or $(\alpha_1, \alpha_2) = (0.01, 0.1)$, and 71.6% of the sessions converge to the equilibrium with $(\alpha_1, \alpha_2) = (0.01, 0.01)$. There are only 0.2% of the sessions converging to the equilibrium with $(\alpha_1, \alpha_2) = (0.001, 0.001)$, even though this is the unique Nash equilibrium in a one-shot game.[11] Our results indicate that in the environment with high noise trading risk, the two informed AI speculators are able to learn to adopt less advanced algorithms (i.e., high α), as if they are implicitly coordinating with each other. This sort of coordination allows both informed AI speculators to obtain supra-competitive profits.

References

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

Abreu, Dilip, David Pearce, and Ennio Stacchetti. 1986. "Optimal Cartel Equilibria with Imperfect Monitoring." *Journal of Economic Theory*, 39(1): 251–269.

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

[11] Complementary to this result, we also find that if one informed AI speculator's α is exogenously fixed at 0.001, the other informed AI speculator will always learn to set its α at 0.001. This implies that although a unilateral deviation by setting $\alpha = 0.001$ could boost self-profit in the short run, it will not be profitable in the long run because the peer informed AI speculator will also learn to set $\alpha = 0.001$.

Azar, Mohammad Gheshlaghi, Remi Munos, Mohammad Ghavamzadeh, and Hilbert J. Kappen. 2011. "Speedy Q-Learning." *International Conference on Neural Information Processing Systems (NIPS)*, 2411 – 2419.

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

Brown, Zach Y., and Alexander MacKay. 2023. "Competition in Pricing Algorithms." *American Economic Journal: Microeconomics*, 15(2): 109–156.

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

Cartea, Álvaro, Patrick Chang, José Penalva, and Harrison Waldon. 2022. "The Algorithmic Learning Equations: Evolving Strategies in Dynamic Games." Working papers.

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

Dou, Winston Wei, David Pollard, and Harrison H. Zhou. 2012. "Estimation in Functional Regression for General Exponential Families." *The Annals of Statistics*, 40(5): 2421–2451.

Dou, Winston Wei, Itay Goldstein, and Yan Ji. 2024. "AI-Powered Trading, Algorithmic Collusion, and Price Efficiency." University of Pennsylvania Working Papers.

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

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." *Review of Financial Studies*, 35(8): 3867–3921.

Dugast, Jérôme, and Thierry Foucault. 2024. "Equilibrium Data Mining and Data Abundance." *Journal of Finance*, 80(1): 211–258.

Fershtman, Chaim, and Ariel Pakes. 2012. "Dynamic Games with Asymmetric Information: A Framework for Empirical Work." *Quarterly Journal of Economics*, 127(4): 1611–1661.

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, 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." *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." *Quarterly Journal of Economics*, 138(4): 2327–2389.

Gu, Shixiang, Timothy Lillicrap, Ilya Sutskever, and Sergey Levine. 2016. "Continuous deep Q-learning with model-based acceleration." *International Conference on International Conference on Machine Learning (ICML)*, 48: 2829 – 2838.

Hall, Peter, and Joel L. Horowitz. 2007. "Methodology and convergence rates for functional linear regression." *The Annals of Statistics*, 35(1): 70 – 91.

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

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference and prediction.* . 2 ed., Springer.

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

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

Kearns, Michael, and Daphne Koller. 1999. "Efficient reinforcement learning in factored MDPs." *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2: 740–747.

Kearns, Michael, and Satinder Singh. 2002. "Near-Optimal Reinforcement Learning in Polynomial Time." *Machine Learning*, 49: 209–232.

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., and Wei Xiong. 2001. “Contagion as a Wealth Effect.” *Journal of Finance*, 56(4): 1401–1440.

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

Mildenstein, Eckart, and Harold Schleef. 1983. “The Optimal Pricing Policy of a Monopolistic Marketmaker in the Equity Market.” *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.

Possnig, Clemens. 2024. “Reinforcement Learning and Collusion.” Working papers.

Rotemberg, Julio J, and Garth Saloner. 1986. “A Supergame-Theoretic Model of Price Wars during Booms.” *American Economic Review*, 76(3): 390–407.

Senda, Kei, Shinji Fujii, and Syusuke Mano. 2006. “Acceleration of Reinforcement Learning by Using State Transition Probability Model.”

Stambaugh, Robert F. 2020. “Skill and Profit in Active Management.”

Sutton, Richard S. 1991. “Dyna, an integrated architecture for learning, planning, and reacting.” *ACM SIGART Bulletin*, 2(4): 160 – 163.

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.