AI-Powered Trading, Algorithmic Collusion, and Price Efficiency

Winston Wei Dou Itay Goldstein Yan Ji *

July 8, 2024

Abstract

The integration of algorithmic trading with reinforcement learning, known as AI-powered trading, has significantly impacted capital markets. This study employs a theoretical laboratory characterized by information asymmetry and imperfect competition, where informed AI speculators serve as the subjects of our simulation experiments. It explores how AI technology impacts market power, information rents, price informativeness, market liquidity, and mispricing. Our findings show that informed AI speculators can autonomously learn to sustain collusive supra-competitive profits without any form of agreement, communication, intention, or any interactions that might violate traditional antitrust regulations. AI collusion robustly emerges from two distinct mechanisms: one through price-trigger strategies ("artificial intelligence") when price efficiency and noise trading risk are both low, and the other through self-confirming bias in learning ("artificial stupidity") under other conditions.

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

*Dou: University of Pennsylvania (wdou@wharton.upenn.edu) and NBER; Goldstein: University of Pennsylvania (itayg@wharton.upenn.edu) and NBER; Ji: Hong Kong University of Science and Technology (jiy@ust.hk). We thank Kerry Back, Snehal Banerjee, Hui Chen, Antoine Didisheim, Itamar Drechsler, Maryam Farboodi, Slava Fos, Cary Frydman, Paolo Fulghieri, Vincent Glode, Joao Gomes, Mark Grinblatt, Ming Guo, Tim Johnson, Chris Jones, Scott Joslin, Larry Harris, Zhiguo He, David Hirshleifer, Jerry Hoberg, Harrison Hong, Mariana Khapko, Leonid Kogan, Pete Kyle, Tse-Chun Lin, Deborah Lucas, Ye Luo, Semyon Malamud, Andrey Malenko, George Malikov, Albert Menkveld, Jonathan Parker, Lasse Pedersen, Josh Pollet, Paul Romer, Nick Roussanov, Tom Sargent, Antoinette Schoar, Hyun-Song Shin, Daniel Sokol, Rob Stambaugh, Yannan Sun, Eric Talley, Anton Tsoy, Quentin Vandeweyer, Jiang Wang, Neng Wang, Liyan Yang, Yilin Yang, Jacob Yunger, David Zhang, Xiaoyan Zhang, and seminar and conference participants at AisanFA, ASU Sonoran Winter Finance Conference, BIS, BI-SHoF Conference, Boston College, CFTRC, CUFE, Duke/UNC Asset Pricing Conference, FINRA, Fudan, George Mason, HKU, HKUST, HK Conference for Fintech, AI, and Big Data in Business, Imperial College, Jackson Hole Finance Conference, Johns Hopkins Carey Finance Conference, Melbourne Asset Pricing Meeting, MIT, MFA, Nordic Fintech Symposium, NYU/Penn Law and Finance Conference, OECD, Olin Finance Conference at WashU, Oxford, PKU/PHBS Sargent Institute Macro-Finance Workshop, QES Global Quant and Macro Investing Conference, QRFE Workshop on Market Microstructure, Fintech and AI, Rice University, SHUFE, Toronto Macro/Finance Conference, Tsinghua PBCSF, Tsinghua SEM, UIUC, University of Macau, University of Minnesota, University of Toronto, USC, WFA, and Wharton for their comments. Dou is grateful for the financial supports from the Golub Faculty Scholar Award at Wharton.

1 Introduction

The integration of algorithmic trading with reinforcement learning (RL) algorithms, often termed AI-powered trading, poses new regulatory challenges and has the potential to fundamentally reshape capital markets.[1] With Nasdaq receiving SEC approval for an RL-based, AI-driven order type, the momentum for AI integration in trading continues to build. Leading digital trading platforms are endorsing RL-based AI trading bots, and major hedge funds, along with investment powerhouses, are adopting AI technologies. This trend has led policymakers, regulators, and financial market supervisors worldwide to make AI a regulatory priority.[2]

The U.S. Securities and Exchange Commission (SEC) has recently issued warnings about the potential for AI collusion that could undermine competition and market efficiency. The concern is that AI algorithms might autonomously optimize themselves to cooperatively benefit a select few sophisticated speculators at the expense of other investors. SEC Chair Gary Gensler has emphasized this concern, noting that there is evidence of machines in high-frequency trading starting to exhibit cooperative behavior independently of human intervention or interaction.

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

The economics of AI collusion in trading can be intuitively understood as follows. On one hand, consider a trading environment where subgame perfect collusive Nash equilibria theoretically

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

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

exist for rational-expectations agents, supported by price-trigger strategies as introduced by Green and Porter (1984). In this environment, even without direct monitoring of trading behaviors, agents can develop collusive incentives. This is achieved by allowing non-collusive competition to occur when market prices diverge from the expected collusive level beyond a certain threshold. If the trading environment is not overly disrupted by noise trading flows, AI algorithms have the capacity to interact and learn, ultimately achieving a steady state, within which they engage in collusive trading based on a price-trigger strategy. On the other hand, in a trading environment where subgame perfect collusive Nash equilibria do not theoretically exist, AI algorithms cannot learn to sustain collusion through price-trigger strategies. Instead, they may converge to a steady state characterized by an experience-based equilibrium, as introduced by Fershtman and Pakes (2012), or a self-confirming equilibrium, as introduced by Fudenberg and Levine (1993). These equilibrium concepts are fundamentally connected and are weaker than Nash equilibrium. They allow for potentially incorrect or biased off-equilibrium valuations or beliefs, which align tightly with the learning and trading behaviors of AI algorithms. Valuations and beliefs may be accurate along the equilibrium path, as this is more commonly observed, but can be inaccurate off the equilibrium path, unless there is sufficient exploration of non-optimal actions (e.g., Fudenberg and Kreps, 1988, 1995; Cho and Sargent, 2008). Crucially, these incorrect off-equilibrium valuations or beliefs are not necessarily inconsistent with observed outcomes along the equilibrium path.

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

demonstration.

In this paper, we adopt a streamlined theoretical framework as our laboratory. Building upon the seminal work of Kyle (1985), we extend this framework in two novel ways. First, our model incorporates multiple informed speculators within a repeated-trading context. Second, we introduce a continuum of atomistic, information-insensitive investors who collectively create a downward-sloping demand curve, along with the inventory cost concerns of market makers. These factors together introduce price inefficiency, in contrast to the efficient pricing in the Kyle (1985) baseline. Within each trading period, agents execute a single transaction. The sequence of events for each period unfolds as follows: Initially, the fundamental value of the asset is determined. Subsequently, a continuum of noise traders collectively places an order flow, which is independent of the asset's fundamental value. The variance of such an aggregate noise trading flow encapsulates the noise trading risk (Long et al., 1990). This noise trading risk is a crucial characteristic of the trading environment. Each oligopolistic informed speculator is aware of the fundamental value but remains uninformed about the noise trading flow when determining his or her optimal trading strategy. The market maker, in turn, sets the market price with the goal of minimizing the weighted average of inventory costs and pricing errors. In doing so, the market maker also takes into account the price elasticity of the information-insensitive investors' demand. This price elasticity represents another critical characteristic of the trading environment.

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

To ascertain whether informed AI speculators' behavior exhibits collusion sustained by price-

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

As a noteworthy theoretical contribution, we establish a novel result on the impossibility of collusion under information asymmetry. We demonstrate that informed speculators are unable to achieve collusive outcomes through price-trigger strategies in certain conditions. This includes scenarios where market prices are already efficient, accurately reflecting the asset's fundamental value, especially when information-insensitive investors have high price elasticity of demand, thereby playing a minimal role in price formation. Another scenario precluding collusion is when the noise trading risk is excessively high. This novel result illuminates a mechanism distinct from existing theories on the impossibility of collusion under information asymmetry in the context of product market competition (Abreu, Milgrom and Pearce, 1991; Sannikov and Skrzypacz, 2007). Intuitively, sustaining price-trigger collusion requires two conditions: first, monitoring necessitates high price informativeness, and second, maintaining informational rents requires a low price impact of informed trading. These two conditions cannot be simultaneously met when price efficiency or noise trading risk is high.

Furthermore, as an additional theoretical contribution, we illustrate that in scenarios where information-insensitive investors, exhibiting low price elasticity of demand, significantly influences price formation, market prices can become inefficient. In such cases, tacit collusion among informed speculators can be sustained through price-trigger strategies. The success of these

strategies is contingent on the degree of price efficiency and the level of noise trading risk in the market. We find that price-trigger strategies can only sustain collusion in markets with both low price efficiency and low noise trading risk. Additionally, we show that collusion capacity increases, market liquidity decreases, price informativeness decreases, and mispricing increases, when the number of informed speculators drops, the level of noise trading risk decreases, or the subjective discount rate factor increases.

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

Finally, we elaborate further on the two distinct mechanisms behind AI collusion across various trading environments. The first mechanism, known as "algorithmic collusion through price-trigger strategies," involves a form of collusion driven by "artificial intelligence." In this scenario, informed AI speculators have the capability to learn and implement price-trigger strategies effectively. This price-trigger strategy enables the AI speculators to sustain collusion and reach a steady state resembling a subgame perfect Nash equilibrium, despite not following it exactly. Such a scenario can only occur if both price efficiency and noise trading risk are low. Leveraging simulation experiments, we provide direct evidence that sizable price deviations trigger aggressive trading flows similar to those in a non-collusive Nash equilibrium, which diminishes the trading profits of all informed AI speculators. While the underlying mechanisms through which AI speculators learn to conduct the price-trigger trading strategy, thereby achieving algorithmic collusion, may differ from those behind how humans would learn to coordinate using price-trigger trading strategies, the resulting patterns exhibit notable similarities. At the heart of these mechanisms, whether involving AI or human speculators, the threat of punishment effectively acts as a deterrent, discouraging individual speculators from violating the collusive agreement. Closely aligned with the theoretical predictions of a collusive Nash equilibrium sustained by price-trigger strategies with rational-expectations agents, as the number or impatience of speculators decreases, the extent of achievable collusion increases. This leads to reduced market liquidity, diminished price informativeness, and increased mispricing.

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

by AI-powered trading, where algorithmic collusion through price-trigger strategies is prevalent, achieving perfect price informativeness becomes unattainable.

The second mechanism, known as "algorithmic collusion through self-confirming bias in learning," involves a form of collusion driven by "artificial stupidity." Despite the learning biases originating from intrinsic features of the RL algorithms, informed AI speculators might still achieve and sustain supra-competitive profits. In the context of RL learning, the emergence of a learning bias is directly linked to inconsistencies in statistical learning. These inconsistencies often stem from an asymmetric effect of exploitation on learning, especially when noise trading risk is excessive. This inherently asymmetric effect of exploitation leads informed speculators to under-react to their private information in their learned trading strategies, compared to the optimal strategy in a non-collusive equilibrium setting. Consider a scenario in which an RL-based AI speculator explores a trading strategy that aggressively responds to private information and receives a positive signal about the asset's fundamental value. If a substantial and positive noise trading flow occurs, this could result in significant losses for the AI speculator. Consequently, the RL algorithm is unlikely to revisit and update its understanding of this state-strategy pair sufficiently, consistently deeming this strategy as suboptimal for the given state. This means the initial adverse effect on the Q-function at the state-strategy pair due to such a shock is unlikely to be mitigated in subsequent iterations. Conversely, if a substantial and negative noise trading flow occurs, it could lead to significant gains for the AI speculator. In this fortunate case, the RL algorithm is more likely to revisit and thoroughly understand the performance of this state-strategy pair, adequately exploiting it, and thus, the initial beneficial effect on the Q-function at this pair may be averaged out, which even leads to accurate estimations of Q-function at this state-strategy pair. Such severe asymmetric learning outcomes from large positive and negative noise trading flows can lead AI speculators to generally under-react to their private information in their learned trading strategies.

This learning bias steers informed AI speculators toward a steady state where trading behaviors can be accurately characterized by an experience-based equilibrium, as introduced by Fershtman and Pakes (2012), or a self-conforming equilibrium, as introduced by Fudenberg and Levine (1993). In contrast to the Nash equilibrium, these equilibrium concepts are weaker because they permit players to hold incorrect (or biased) off-equilibrium valuations or beliefs. This concept

of equilibrium is motivated by the idea that noncooperative equilibria should be interpreted as outcomes of a learning process, where players form beliefs based on their past experiences. While beliefs can generally be correct along the equilibrium path of play due to its frequent observation, they are not necessarily correct off the equilibrium path. Correct beliefs off the equilibrium path require players to engage in sufficient experimentation with non-optimal actions, as suggested in works by Fudenberg and Kreps (1988), Fudenberg and Kreps (1995), and Cho and Sargent (2008).

These two types of AI collusion, while both generating supra-competitive trading profits, can exhibit opposite collusive behaviors as trading environments vary. In AI collusion through price-trigger strategies, a decrease in noise trading risk or an increase in the subjective discount rate factor leads to increased collusion capacity. This results in reduced market liquidity, diminished price informativeness, and increased mispricing. In contrast, AI collusion through self-confirming bias in learning shows that an increase in the subjective discount rate factor has little impact on the collusive experience-based equilibrium. Moreover, unlike price-trigger strategies, an elevation in noise trading risk enhances the potential for collusion due to a more pronounced learning bias, leading to reduced market liquidity, diminished price informativeness, and increased mispricing. Notably, in the scenario with price-trigger strategies, the collusive, supra-competitive trading profit of informed AI speculators primarily comes from trading against information-insensitive investors. Conversely, in the case of AI collusion through self-confirming bias in learning, these profits are significantly, if not entirely, derived from trading against noise traders.

Homogenization is instrumental, though not necessary, for AI collusion to be achieved, regardless of the mechanism through which it occurs. Homogenization can emerge when speculators use similar foundational models, effectively forming a type of hub-and-spoke conspiracy.[3] Johnson and Sokol (2021) emphasize the prevalence of this type of AI collusion in the context of e-commerce platforms, observing that many retailers adopt similar or even identical AI pricing algorithms. Specifically, anti-competitive effects may emerge when multiple competitors use the same AI pricing algorithm supplied by a common service provider, who serves as the hub. In the financial markets, informed speculators often rely on similar foundational models for their AI-powered trading systems. This practice, whether intentional or not, can result in a significant

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

degree of homogenization, a phenomenon documented by Bommasani et al. (2022), among others.

Furthermore, homogenization can also emerge due to the autonomous learning of two-layer AI-powered trading systems. Although adopting superior algorithms can disrupt the collusion created by self-confirming bias in learning, it is likely that no AI speculator would choose to gain an advantage by using superior algorithms due to the nature of AI collusion. Intuitively, if one speculator adopts a superior algorithm, it could render the trading strategies of other AI speculators unprofitable, thereby compelling them to adopt equally or more advanced algorithms. This could spark a race towards algorithmic advancement, ultimately leading to an equilibrium where trading profitability is minimal for every AI speculator. Consequently, AI speculators autonomously learn to adopt similarly basic algorithms in equilibrium. To illustrate this point, we consider a simple extension of the baseline Q-learning algorithms, wherein informed AI speculators are able to learn both the key parameter that governs the sophistication of their Q-learning algorithms and their trading strategies based on the AI-chosen Q-learning algorithm. Our simulation experiments robustly demonstrate that informed AI speculators may collectively opt for less advanced algorithms. This occurs despite the potential for increased self-profit that could come from unilaterally choosing a more advanced algorithm while others' algorithms remain fixed.

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

strategies. Expanding upon this, our paper extensively broadens the AI experimental framework, moving from a scenario of perfect information and a static demand curve to one imbued with asymmetric information and a strategically-determined demand scheme. We characterize the various types of AI algorithmic collusion, whether occurring through price-trigger strategies or through self-confirming bias in learning, across diverse market environments.

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

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

2 AI-Powered Trading Algorithms

The traditional rule-based algorithmic trading system executes orders rigidly according to protocols predefined by human quantitative strategists. These rules are typically derived from technical analysis and statistical models. In contrast, AI-powered trading employs RL algorithms to dynamically adjust and autonomously optimize trading strategies in real-time.

The RL algorithm, a pivotal technique in AI, forms the foundation of numerous successful AI algorithms, like "AlphaGo," demonstrating the superiority of RL-backed AI over human cognitive abilities in areas such as securities trading and other complex tasks. RL algorithms are model-free machine learning techniques that learn autonomously through trial-and-error experimentation, without relying on typical assumptions, such as the multi-agent system being on an equilibrium path or agents having knowledge of the true state and payoff distributions at equilibrium. The basic rationale behind RL algorithms centers on the principle that actions yielding higher rewards historically are more likely to be selected in the future, compared to those that have led to lesser rewards. By interacting with its environment and experimenting with different actions, the agent incrementally learns an optimal policy. Through continuous rounds of exploration and experimentation, it refines its strategy to prefer actions that offer the greatest long-term benefits, even without any knowledge of the environment beforehand. This iterative process enables the agent to progressively enhance its decision-making approach, consistently steering towards actions that maximize the cumulative rewards based on its gathered experiences.

While RL encompasses different variants (e.g., Watkins and Dayan, 1992; Sutton and Barto, 2018), we choose to focus on Q-learning for several reasons. First, Q-learning serves as a foundational framework for numerous dynamically sophisticated RL algorithms, upon which many recent AI breakthroughs are built.[4] However, it is important to note that AI trading algorithms currently in use may not exclusively rely on Q-learning principles. Second, Q-learning holds substantial popularity among computer scientists in practical applications. Third, Q-learning algorithms possess simplicity and transparency, offering clear economic interpretations, in contrast to the black-box nature of many machine learning and AI algorithms.

[4]Q-learning and these dynamically sophisticated RL algorithms are typically employed in complex scenarios, where actions lead to state transitions, and each action taken in a state affects future states and rewards. In contrast, multi-armed bandit algorithms, which represent another category of RL algorithms, are employed in simpler settings where actions do not depend on previous ones and do not prompt state transitions based on the actions taken.

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

2.1 Bellman Equation and Q-Function

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

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

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

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

curly brackets of the Bellman equation (2.1):

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

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

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

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

2.2 Q-Learning Algorithm

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

The iterative experimentation of agent i starts from an arbitrary initial agent-i Q-matrix, denoted by $\widehat{Q}_{i,0}$, and updates its estimated Q-matrix $\widehat{Q}_{i,t}$ recursively as follows:

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

where $\alpha \in [0, 1]$ captures the forgetting rate, s_t is the state that the iteration t concentrates on, s_{t+1} is randomly drawn from the Markovian transition probabilities conditional on the current state s_t, the chosen action x_i of agent i, and the collective actions of all other agents within the system.

Here, $\widehat{Q}_{i,t}(s,x)$ is the estimated Q-matrix of agent i in the t-th iteration, and $\pi_{i,t}$ is the payoff in the t-th iteration, corresponding to agent i's choice of action $x_{i,t}$ and all other agents' actions.

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

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

$$\widehat{Q}_{i,t_\tau}(s,x) \approx \sum_{h=0}^{\tau-1} \alpha(1-\alpha)^h \left[\pi_{i,t_{\tau-h}} + \rho \max_{x' \in \mathcal{X}} \widehat{Q}_{i,t_{\tau-h}} \left(s_{t_{\tau-h}+1}, x' \right) \right], \tag{2.5}$$

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

2.3 Experimentation

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

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

To determine the mode, we employ the simple ε-greedy method. As outlined in equation (2.6), during the t-th iteration, agent i engages in the exploration and exploitation modes with exogenous probabilities ε_t and $1 - \varepsilon_t$, respectively. In the exploitation mode, agent i chooses its action to maximize the current state's Q-value based on past experience, given by $x_{i,t} = \text{argmax}_{x \in \mathcal{X}} \widehat{Q}_{i,t}(s_t, x)$. Conversely, in the exploration mode, agent i randomly chooses its action $\widetilde{x}$ from the set of all possible values in $\mathcal{X}$, each with equal probability.[5] Essentially, the exploration mode guides the Q-learning algorithm to experiment with suboptimal actions based on the current Q-matrix estimation, $\widehat{Q}_{i,t}$. As t approaches infinity, the pre-specified exploration probability ε_t monotonically decreases to zero. Sufficient exploration is crucial for accurately approximating the true Q-matrix, requiring many attempts of all actions in all states, especially in complex environments. However, this comes with a tradeoff: extensive exploration not only increases computational costs but can also introduce noise, impeding learning when multiple agents interact.

3 Model and Laboratory Design

To set up the laboratory for our simulation experiments, we develop a model that incorporates only the minimal set of ingredients necessary to capture the economic context of securities trading and elucidate key novel insights. Our model extends the influential framework of Kyle (1985), emphasizing the financial market as an information-aggregating mechanism, where information asymmetry and strategic trading with endogenous asset demand are crucial in price formation.

Specifically, we introduce two minimal deviations from the standard Kyle (1985) model. First, we shift our focus to oligopolistic informed speculators in a repeated trading environment,

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

instead of a monopolistic informed speculator in a one-period trading environment. Second, we incorporate information-insensitive investors (e.g., Kyle and Xiong, 2001; Vayanos and Vila, 2021) and the inventory cost concerns of market makers, which together introduce price inefficiency, as opposed to the efficient pricing in the Kyle (1985) baseline. We emphasize that information-insensitive investors in our model do not need to be subject to behavioral bias; instead, they can be entirely rational but simply unresponsive to short-term information.

Merging theoretical rigor with practical relevance, this model acts as a laboratory for exploring the effects of AI-powered trading on price formation, focusing on its implications for market liquidity, price informativeness, and mispricing due to algorithmic collusion. The model's theoretical findings serve as a benchmark for characterizing AI collusion in the experiments.

3.1 Economic Environment

Model Setup. Time is discrete, indexed by $t = 1, 2, ...$, and runs forever. There are $I \geq 2$ risk-neutral informed speculators, indexed by $i \in \{1, \cdots, I\}$, a representative noise trader, a representative information-insensitive investor, and a market maker. The environment is stationary, and all exogenous shocks are independent and identically distributed across periods.

In each period t, an asset trades with its fundamental value v_t realized at period's end, distributed as $N(\overline{v}, \sigma_v^2)$, where we set $\overline{v} \equiv \sigma_v \equiv 1$ for simplicity.[6] Trading profits are calculated after v_t is known. The noise trader's order flow u_t follows $N(0, \sigma_u^2)$, with σ_u capturing noise trading risk. Noise traders are needed to incorporate meaningful information asymmetry. Each informed speculator i knows v_t perfectly but does not observe u_t when submitting their orders; they understand that their order flow $x_{i,t}$ impacts the asset's market price p_t by shifting the market-clearing condition and revealing information. Specifically, informed speculator i solves:

$$V_i(s_t) = \max_{x_{i,t}} \mathbb{E}\left[(v_t - p_t)x_{i,t} + \rho V_i(s_{t+1}) | s_t, x_{i,t}\right], \tag{3.1}$$

where the state variable s_t encapsulates the relevant information necessary for informed speculator i's trading strategy $x_{i,t}$, with s_t including variables such as $\{v_t, v_{t-1}, p_{t-1}, y_{t-1}, z_{t-1}\}$, along with other historical variables as needed, $V_i(s)$ is the optimal value function of speculator i, $\rho \in (0, 1)$

[6] For conciseness, the notations $\overline{v}$ and σ_v will be omitted in this manuscript when not needed for comprehension.

is the subjective discount rate factor, and $y_t = \sum_{i=1}^{I} x_{i,t} + u_t$ is the total order flow from both informed speculators and the noise trader.

The quantity z_t in (3.1) is the demand of information-insensitive investors, who have a collective demand curve:

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

The same specification is adopted by Kyle and Xiong (2001), who justify it through the optimal portfolio choice made by a rational yet information-insensitive investor under certain assumptions.[7] Similarly, in our model, information-insensitive investors do not need to be subject to behavioral bias; instead, they can be entirely rational but simply unresponsive to short-term information. The logic behind specification (3.2) is straightforward: the information-insensitive investor, focusing on the ex-ante expected fundamental value $\overline{v}$, buys more as $p_t - \overline{v}$ becomes more negative, interpreting this as the asset being undervalued. The fundamental idea of introducing information-insensitive investors with exogenous net demand curves in the framework of a noisy rational expectations equilibrium is to efficiently capture relevant institutional frictions and preferences. This approach has been commonly adopted in the literature (e.g., Hellwig, Mukherji and Tsyvinski, 2006; Goldstein, Ozdenoren and Yuan, 2013).

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 from informed speculators and the noise trader, represented by y_t, as well as the order flow schedule z_t of information-insensitive investors for any selected market price p_t. However, the market maker cannot distinguish between order flows from informed speculators and the noise trader. The market maker sets the market price p_t to jointly minimize inventory and pricing errors according to the following objective function:

$$\min_{p_t} \mathbb{E}\left[(y_t + z_t)^2 + \theta(p_t - v_t)^2 \middle| y_t\right], \tag{3.3}$$

[7] To derive the functional-form of the aggregate demand curve of information-insensitive investors, one approach is to assume CARA utility maximization without any learning or strategic trading, as detailed in Online Appendix 2.1. 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). These investors can be rational, although they do not learn fundamental information through market price p_t as rational-expectations uninformed investors in Grossman and Stiglitz (1980) and Kyle (1989) do.

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 understanding of the behavior of informed speculators in equilibrium.

The market maker's objective function, detailed in (3.3), accounts for both inventory costs and challenges related to asymmetric information. To clear the market, the market maker assumes the position $-(y_t + z_t)$, incurring inventory costs represented by $(y_t + z_t)^2$. This quadratic formulation, chosen for its simplicity, is consistent with established literature, such as 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. The first-order condition leads to

$$p_t = \frac{\xi}{\xi^2 + \theta} y_t + \frac{\xi^2}{\xi^2 + \theta} \overline{v} + \frac{\theta}{\xi^2 + \theta} \mathbb{E}[v_t|y_t]. \tag{3.4}$$

The impact of the pricing error term is minimal in practice, and our results remain unchanged with $\theta = 0$. However, we choose to treat θ as a tiny, universally fixed positive constant in both our theoretical and simulation analyses. By fixing θ, we do not subject it to comparative-static variation as an environmental parameter in our theoretical and experimental frameworks. With $\theta > 0$, our theoretical framework or laboratory setup becomes more conceptually coherent, providing two meaningful extreme benchmarks, unlike when $\theta = 0$. Specifically, when $\theta > 0$, 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).[8]

Model Interpretation. Here, we provide a specific interpretation of the model, offering an economic context for the simulation experiments involving AI-powered trading algorithms, although other interpretations are possible. At the beginning of each period, a different asset, such as a short-lived derivative contract that expires at the end of the period, is traded. The short-lived derivatives are typically close-to-maturity options and futures. The fundamental value v_t represents the intrinsic value of the short-lived derivative contract, which is realized at the end

[8] More discussions are detailed in Online Appendix 2.1.1.

of period t (i.e., at maturity).

Informed speculators, typically quant-based hedge funds and quantitative trading firms, specialize in privately extracting valuable trading signals about intrinsic value using their advanced data and technological advantages. These traders are usually the leading quantitative players in the derivatives market, equipped with the necessary advanced trading infrastructure to establish AI-powered trading systems and exploit information rents through complex strategic trading strategies.

Information-insensitive investors are those who remain unresponsive to short-term fundamental information in the market of close-to-maturity derivatives and make trading decisions based solely on market price patterns. These investors are typically retail investors who employ technical analysis, which aims to trade based on patterns of market prices (e.g., Lo and MacKinlay, 1999; Lo, Mamaysky and Wang, 2000; Chen, Peng and Zhou, 2024). The demand specification (3.2) captures the essence of certain technical analysis strategies, assuming that a positive spread $p_t - \overline{v}$ indicates overbought conditions with prices likely to fall, whereas a negative spread $p_t - \overline{v}$ indicates oversold conditions with prices likely to rise. These investors can also include institutional entities such as pension funds, insurance companies, and mutual funds. They might purchase close-to-maturity derivatives and hold them until expiration to hedge existing positions or adjust their exposure to specific near-term event risks without having to buy or sell the underlying assets.

Noise traders are market participants who make trading decisions based on factors unrelated to the fundamental information or any private signals about the intrinsic value of derivative securities. Their decisions stem from reasons beyond fundamental or technical analysis, including liquidity needs, portfolio rebalancing, psychological factors, or random speculation.

3.2 Theoretical Benchmarks

Non-Collusive Nash Equilibrium. Informed speculators do not internalize the impact of their trading on others' profits. In the non-collusive Nash equilibrium (N), each informed speculator i seeks to maximize expected profit by solving $x^N(v_t) = \text{argmax}_{x_i \in \mathcal{X}} \mathbb{E}[(v_t - p^N(y_t))x_i | v_t]$, taking into account their private information v_t, assuming that other informed speculators follow trading strategy $x^N(v_t)$, and understanding and internalizing their own price impact on the equilibrium price $p^N(y_t)$ as $p^N(y_t) = \overline{v} + \lambda^N y_t$, where $y_t = x_i + (I-1)x^N(v_t) + u_t$ and λ^N is dependent on

market parameters and the equilibrium trading strategy $x^N(v_t)$. Here, we focus on linear trading strategies $x^N(v_t) \equiv \chi^N(v_t - \overline{v})$ in the equilibrium. Details are in Online Appendix 2.1.

Perfect Cartel Benchmark. Informed speculators operate collectively as a monopoly, strategically coordinating their total order flow before allocating trades among themselves according to predetermined agreements. In the perfect cartel equilibrium (M), the cartel seeks to maximize expected profit by solving $x^M(v_t) = \text{argmax}_{x \in \mathcal{X}} \mathbb{E}[(v_t - p^M(y_t))x|v_t]$, taking into account the private information v_t, and understanding and internalizing its price impact on the equilibrium price $p^M(y_t)$ as $p^M(y_t) = \overline{v} + \lambda^M y_t$, where $y_t = Ix + u_t$ and λ^M is dependent on market parameters and the equilibrium trading strategy $x^M(v_t)$. Here, we focus on linear trading strategies $x^M(v_t) \equiv \chi^M(v_t - \overline{v})$. Details are in Online Appendix 2.1.

Collusive Equilibrium. A collusive equilibrium is descriptively categorized by examining the behavior of agents both on and off the equilibrium path, rather than focusing on the mechanisms through which it is achieved. These mechanisms may include punishment-based trigger strategies, agreed upon by agents either explicitly or implicitly, as well as other mechanisms that sustain collusion without requiring agreement or communication among agents. Below, we define the notion of collusive equilibrium.

Definition 3.1. *A collusive equilibrium is generically defined by two key properties: (i) all agents achieve supra-competitive profits, and (ii) there are short-term gains for agents who deviate from on-path equilibrium actions at the expense of others.*

Collusive Nash Equilibrium Sustained by Price-Trigger Strategies. Tacit collusion can arise in a subgame perfect Nash equilibrium, sustained by punishment-based trigger strategies. In securities trading, information asymmetry and noise trading risk complicate tacit collusion, primarily due to challenges in monitoring each other's trading actions. Nonetheless, under certain circumstances, collusion can be sustained through the so-called "price-trigger strategies," which allow informed speculators to infer others' order flows from market prices, thereby upholding collusive incentives.[9]

[9]The study of tacit collusion through grim-trigger strategies, initiated by Fudenberg and Maskin (1986) and Rotemberg and Saloner (1986), has been significantly advanced in recent research. This includes studies on its impact on asset pricing (e.g., Opp, Parlour and Walden, 2014; Dou, Ji and Wu, 2021*a*,*b*; Dou, Wang and Wang, 2023).

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. Within our framework, 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 formally describe the collusive Nash equilibrium, sustained through price-trigger strategies, as a subgame perfect Nash equilibrium in repeated games. Our focus narrows to the symmetric linear collusive Nash equilibrium scenario. Consider a situation where, at time t, informed speculators are within a collusion regime. In coordination, they adopt an identical trading strategy denoted by $x^C(v) \equiv \chi^C(v - \overline{v})$, understanding and internalizing that the equilibrium market price, $p^C(y_t)$, will be $p^C(y_t) = \overline{v} + \lambda^C y_t$, with $y_t = Ix^C(v_t) + u_t$. This reflects their understanding of the dependence of λ^C on market parameters and the equilibrium trading strategy $x^C(v)$. Once $v_t > \overline{v}$ and the observed market price p_t is above the price-trigger strategy threshold $q_+(v_t) \equiv \mathbb{E}\left[p^C(y_t)|v_t\right] + \lambda^C \sigma_u \omega$, that is, $p_t > q_+(v_t)$, they revert to the non-collusive Nash equilibrium at $t+1$ with a probability η. Once in the non-collusive regime, they will remain in this regime with the same probability η in each subsequent period up to period $t+T$. Similarly, once $v_t < \overline{v}$ and the observed market price p_t is below the price-trigger strategy threshold $q_-(v_t) \equiv \mathbb{E}\left[p^C(y_t)|v_t\right] - \lambda^C \sigma_u \omega$, that is, $p_t < q_-(v_t)$, they may revert to the non-collusive Nash equilibrium at $t+1$ with probability η, following the same probabilistic rule up to period $t+T$ as described above. Here, the constants $\eta \in [0,1]$, $\omega > 0$, and $T \geq 1$ are part of the implicit agreement among informed speculators. The rationale behind these price-trigger strategies in capital markets is straightforward: excessive deviations of market prices from the anticipated collusive equilibrium price level suggest potential cheating by other informed speculators. More details about the collusive price-trigger strategy are in Online Appendix 2.1.

The following proposition highlights the impossibility of achieving a collusive Nash equilib-

rium through price-trigger strategies in an environment that closely resembles the standard Kyle (1985) benchmark (when ξ is small), where efficient prices prevail, or the noise trading risk is excessive (when σ_u is large). The proof is in Online Appendix 2.3.

Proposition 3.1 (Impossibility of Collusion Through Price-Trigger Strategies)**.** *When ξ is small or σ_u is large, sustaining a collusive Nash equilibrium through price-trigger strategies becomes impossible.*

Sustaining coordination through price-trigger strategies hinges on two critical conditions: (i) the informativeness of prices must be high enough to allow for adequate monitoring, a point underscored by both Abreu, Milgrom and Pearce (1991) and Sannikov and Skrzypacz (2007); and (ii) the price impact of the informed speculators' order flows must be sufficiently low to permit the attainment of significant informational rents. However, in environments where σ_u is large, price informativeness is low, making adequate monitoring for collusion impossible. More importantly, the environments with small ξ closely resemble the standard Kyle benchmark (Kyle, 1985), where efficient prices prevail. In this environment, when price informativeness is high, the price impact of the informed speculators' order flows must also be high. As a result, the two necessary conditions (i) and (ii) cannot hold simultaneously. We underscore the intrinsic value of this theoretical result, offering novel economic insights that distinguish it from established theories on the impossibility of collusion under information asymmetry, as posited by Abreu, Milgrom and Pearce (1991) and Sannikov and Skrzypacz (2007). By illustrating the challenges of achieving collusion in environments with efficient prices, our findings enhance the understanding of market dynamics and the impact of information asymmetry on collusion strategies.

The proposition presented below demonstrates that collusion sustained by price-trigger strategies is feasible when information-insensitive investors significantly influence price formation – that is, when ξ is large, resulting in less efficient prices – and the noise trading risk, σ_u, is low. The proof is in Online Appendix 2.4.

Proposition 3.2 (Existence of Collusion Through Price-Trigger Strategies)**.** *When ξ is large and σ_u is small, a collusive Nash equilibrium sustained by price-trigger strategies exists.*

If ξ is large and σ_u is small, the market maker primarily sets prices to minimize inventory costs, not pricing errors. This leads to a low price impact from informed trading, even in environments with low noise trading risks. The reduced price impact encourages informed speculators to place

large orders, enhancing price informativeness. Consequently, conditions (i) and (ii) required for sustaining collusion through price-trigger strategies can be met simultaneously.

To discern whether informed speculators trade in a tacitly collusive manner based on observable outcomes, we derive testable properties of collusion. The proof of the following proposition is in Online Appendix 2.6.

Proposition 3.3 (Supra-Competitive Nature of Collusion)**.** *Let π^C, π^N, and π^M represent the expected profits of informed speculators in the price-trigger collusive, non-collusive, and perfect cartel equilibria, respectively. It holds that*

$$\Delta^C \equiv \frac{\pi^C - \pi^N}{\pi^M - \pi^N} \in (0, 1].^{10} \tag{3.5}$$

The price informativeness, market liquidity, and mispricing are measured, respectively, by

$$\mathcal{I} = \log\left[\frac{\text{var}(x_t)}{\text{var}(u_t)}\right], \quad \mathcal{L} = \left[\frac{\partial |m_t|}{\partial u_t}\right]^{-1}, \quad \text{and} \quad \mathcal{E} = \left|\frac{p_t - \mathbb{E}[v_t|y_t]}{\mathbb{E}[v_t|y_t] - \overline{v}}\right|, \tag{3.6}$$

where x_t, z_t, u_t, and $m_t \equiv -(y_t + z_t)$ are the total order flow of informed speculators, information-insensitive investors, noise traders, and market makers, respectively, and p_t is the market price of the asset. In the following proposition, we examine how Δ^C, π^C, $\mathcal{I}^C$, $\mathcal{L}^C$, and $\mathcal{E}^C$ vary across various market structures and information environments in the collusive Nash equilibrium sustained by price-trigger strategies. The proof is in Online Appendix 2.6.

Proposition 3.4 (Effects of Market Structures and Information Environments)**.** *If the price-trigger collusive Nash equilibrium exists and ξ is sufficiently large, the following properties hold:*

$$\sigma_u \uparrow, \quad \rho \downarrow, \quad \text{or} \quad \xi \downarrow \quad \Longrightarrow \quad \Delta^C \downarrow, \quad \mathcal{I}^C/\mathcal{I}^M \uparrow, \quad \mathcal{L}^C/\mathcal{L}^M \uparrow, \quad \text{and} \quad \mathcal{E}^C \downarrow,$$

where C and M represent the price-trigger collusive Nash and the perfect cartel equilibrium, respectively. Furthermore, when I is sufficiently large, Δ^C and $\mathcal{E}^C$ are monotonically decreasing in I, and thus $\mathcal{I}^C/\mathcal{I}^M$ and $\mathcal{L}^C/\mathcal{L}^M$ are monotonically increasing in I.

Collusive Experience-Based Equilibrium. Proposition 3.1 shows that achieving a collusive Nash equilibrium through price-trigger strategies is impossible when ξ is small or σ_u is large. However,

[10] Clearly, a greater Δ^C signifies higher collusion capacity. We adopt Δ^C as a measure for collusion capacity, following Calvano et al. (2020). Similar measures are also used in empirical studies like Dou, Wang and Wang (2023).

in these cases, a form of collusive trading behavior may still emerge as an observed outcome of an experience-based equilibrium, as conceptualized by Fershtman and Pakes (2012), which is closely related to the notion of self-confirming equilibrium (e.g., Fudenberg and Levine, 1993; Battigalli et al., 2015). Compared to the Nash equilibrium, these alternative notions of equilibrium are considered weaker because they allow players to hold incorrect (or biased) beliefs or evaluations about outcomes off the equilibrium path. Unlike the Nash equilibrium, in these weaker equilibria, players' actions are shaped by what they have learned from past experiences. While beliefs or evaluations along the equilibrium path of play are correct (i.e., consistent with the evidence generated by equilibrium strategies) due to frequent and recurrent observation, beliefs or evaluations off the path are not necessarily correct unless players engage in sufficient experimentation with non-optimal actions (e.g., Fudenberg and Kreps, 1988, 1995; Cho and Sargent, 2008).

Specifically, an experience-based equilibrium features (i) a recurrent Markovian state process, (ii) strategies optimized for potentially incorrect outcome evaluations, and (iii) behaviors that yield expected discounted future net cash flows, which are consistent with the evaluations of their outcomes on the equilibrium path. Crucially, the conditions of this equilibrium do not require that players' evaluations of outcomes from off-equilibrium strategies align with the actual distribution of outcomes, meaning that off-equilibrium evaluations of outcomes can be significantly biased. Instead, the sole restriction is that players' beliefs and consequent evaluations of outcomes must be consistent with the evidence derived from strategies employed within the equilibrium. In our next proposition, we illustrate that a collusive equilibrium, with self-confirming bias in learning, can consistently arise as an experience-based equilibrium. The proof of the proposition is in Online Appendix 2.7.

Proposition 3.5 (Existence of Collusion Through Homogenized Self-Confirming Bias)**.** *A collusive experience-based equilibrium, where supra-competitive trading profits ($\Delta^C > 0$) are achieved, can be sustained for all $\xi > 0$ and $\sigma_u > 0$. In this equilibrium, informed speculators uniformly undervalue aggressive trading strategies, perpetuating an incorrect system of outcome evaluation that remains uncorrected. As a result, they employ the trading strategy $x^C(v_t) \equiv \chi^C(v_t - \overline{v})$, with $\chi^C < \chi^N$, indicating a lower level of trading aggressiveness based on private information compared to that in the non-collusive Nash equilibrium.*

Experience-based equilibrium offers a more flexible framework than Nash equilibrium. While our focus is on equilibria with homogenized self-confirming bias in learning, there are numerous

experience-based equilibria characterized by heterogeneous learning biases among informed speculators. Our emphasis on an experience-based equilibrium with homogeneous learning biases in Proposition 3.5 is driven by the following consideration: homogenization is a key feature of AI technologies in financial markets, as emphasized by many regulatory entities. Homogenization of AI applications may occur through collusion in trading algorithm selection among multi-tier AIs or through the adoption of common foundational machine learning models. Furthermore, our simulation experiments, discussed in Sections 4 through 6, demonstrate that a certain level of homogenization in self-confirming bias in learning – though not requiring strong homogenization – is critical for enabling AI-powered trading algorithms to converge to a collusive experience-based equilibrium, thereby sustaining supra-competitive trading profits.

4 Simulation Experiments on AI Trading Algorithms

Theoretical benchmarks discussed in Section 3 detail the conditions necessary for a collusive Nash equilibrium, sustained by price-trigger strategies, to emerge. They also outline when a collusive experience-based equilibrium, sustained by self-confirming bias in learning, exists. Despite this, it remains uncertain whether autonomous, model-free AI algorithms can learn to maintain tacit collusion during trading, thereby achieving supra-competitive trading profits in alignment with these theoretical benchmarks. Furthermore, it is unclear which type of collusive equilibrium these AI trading algorithms will ultimately reach and maintain in a steady state.

As a proof-of-concept illustration, in this section, we design simulation experiments to investigate the capability of Q-learning algorithms to attain tacit collusion under asymmetric information and the endogenous strategic asset demand curve of a market maker, without the overt acts of communication or agreements typically seen in competition law infringements (Harrington, 2018).

4.1 Algorithms as Experimental Subjects

Informed AI Speculators. We now examine the behavior of algorithms as subjects in the experimental lab, as detailed in Section 3.1. Specifically, in these experiments, we replace the theoretical agents known as "informed speculators" in the model with Q-learning algorithms, as outlined in Section 2. To reinforce the key qualitative message derived from these experiments, we employ

the simplest and most naive Q-learning algorithm. The dimensionality of the vector of state variables, s_t, is particularly crucial for determining the capacity of Q-learning algorithms, with the high-dimensionality challenge typically tackled by using deep learning techniques.[11] To highlight the key insight, maintain numerical tractability, and ensure transparency, we intentionally choose the smallest possible set of state variables in $s_t \equiv \{p_{t-1}, v_{t-1}, v_t\}$ that can theoretically capture the information advantage of informed speculators and facilitate tacit collusion sustained by price-trigger strategies.[12] Put simply, we equip the informed AI speculator with private information v_t for trading in period t and a one-period memory of the historical market price p_{t-1} and asset value v_{t-1}, which enables the tracking of historical data for decision making in period t. We could expand informed AI speculator i's state variables in s_t with its own lagged order flow $x_{i,t-1}$ and a longer memory for lagged asset prices, values, and order flows. In our simulation experiments, we observe that enlarging the state variable s_t augments the degree of tacit collusion among informed AI speculators, leading to higher trading profits. Thus, our deliberate choice to solely incorporate p_{t-1}, v_{t-1}, and v_t as state variables sets a stringent bar for the Q-learning algorithms to reach tacit collusion within our economic environment. We discuss the implications of alternative choices of state variables in Online Appendix 1.4.

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

$$z_{t-\tau} = \xi_0 - \xi_1 p_{t-\tau}, \quad \text{and} \quad v_{t-\tau} = \gamma_0 + \gamma_1 y_{t-\tau} + \epsilon_{t-\tau}, \tag{4.1}$$

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

[11] Reinforcement learning algorithms, enhanced by deep learning techniques to address the challenge of high dimensionality, form the backbone of many successful real-world AI applications, such as "AlphaGo."

[12] Intuitively, by tracking both p_{t-1} and v_{t-1}, rather than just p_{t-1}, informed AI speculators can better assess the likelihood of deviation by other informed speculators in period $t-1$ by comparing p_{t-1} against v_{t-1}.

policy using a plug-in procedure:

$$\widehat{p}_t(y) = \widehat{\gamma}_{0,t} + \widehat{\lambda}_t y \;\text{ with }\; \widehat{\lambda}_t = \frac{\theta\widehat{\gamma}_{1,t} + \widehat{\xi}_{1,t}}{\theta + \widehat{\xi}_{1,t}^2}, \tag{4.2}$$

where θ is defined in (3.3). Thus, the market maker is adaptive using simple statistical models. To show robustness of our results, we also consider the economic environment where the market maker determines the pricing rule with rational expectations or adopts Q-learning algorithms to learn the trading strategy of informed AI speculators (see Online Appendix 1.11). Across all scenarios, the results are consistently similar.

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

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

(1) In period t, each informed AI speculator i independently enters exploration mode with a probability of ε_t or exploitation mode with a probability of $1 - \varepsilon_t$. Then, based on the mode it is in, each informed AI speculator i submits its own order flow $x_{i,t}$, as specified in (2.6).

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

(3) The market maker analyzes the historical data $\mathcal{D}_t \equiv \{v_{t-\tau}, p_{t-\tau}, z_{t-\tau}, y_{t-\tau}\}_{\tau=1}^{T_m}$ and estimates the optimal pricing rule $\widehat{p}_t(y)$ according to (4.2). Upon observing $y_t = \sum_{i=1}^{I} x_{i,t} + u_t$, the market price is set at $p_t = \widehat{p}_t(y_t)$.

(4) Observing p_t, information-insensitive investors submit their aggregate order flow z_t in accordance with (3.2). Each informed AI speculator i realizes its profits $\pi_{i,t} = (v_t - p_t)x_{i,t}$.

(5) At the beginning of period $t+1$, the state variable for each informed AI speculator transitions from $s_t = \{p_{t-1}, v_{t-1}, v_t\}$ to $s_{t+1} = \{p_t, v_t, v_{t+1}\}$, where v_{t+1} is randomly drawn from $N(\overline{v}, \sigma_v^2)$ and is independent of any other variables. Following this, each informed AI speculator i updates its Q-matrix for the specific state-action pair $(s_t, x_{i,t})$ in accordance with the recursive update of $\widehat{Q}_{i,t+1}(s_t, x_{i,t})$ outlined in (2.4).

4.2 Numerical Specifications

We now detail the numerical specifications for our simulation experiments. This includes the discretization of the state and action spaces, the initialization of Q-matrices, the selection of parameters, and the criteria for convergence.

Discretization of State and Action Spaces. We approximate the normal distribution $N(\overline{v}, \sigma_v)$ using a sufficiently larger number of n_v grid points, $\mathbb{V} = \{v_1, \cdots, v_{n_v}\}$, with equal probabilities across the grids. Specifically, the probability of each grid point is $\mathcal{P}_k = 1/n_v$. The locations of grid points are chosen based on $v_k = \overline{v} + \sigma_v \Phi^{-1}((2k-1)/(2n_v))$ for $k = 1, \cdots, n_v$, where Φ^{-1} is the inverse cumulative density function of a standard normal distribution.[13]

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

[13] All results remain robust when alternative methods are used to discretize the state variable v_t.

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

Initial Q-Matrix and States. We initialize the Q-matrix at $t = 0$ using the discounted payoff that would accrue to informed AI speculator i if the other informed AI speculators randomize their actions uniformly over the grid points defined by $\mathbb{X}$.[14] Moreover, we consider a zero order flow from the noise trader, corresponding to the expected value of the distribution $N(0, \sigma_u^2)$. Specifically, for each informed AI speculator $i = 1, \cdots, I$, we set its initial Q-matrix $\widehat{Q}_{i,0}$ at $t = 0$ as follows:

$$\widehat{Q}_{i,0}(s, x) = \frac{1}{(1-\rho)n_x} \sum_{x_{-i} \in \mathbb{X}} \left[v - (\overline{v} + \lambda^N(x + (I-1)x_{-i})) \right] x,$$

for $s = (p, v, v) \in \mathbb{P} \times \mathbb{V} \times \mathbb{V}$ and $x \in \mathbb{X}$. The initial states of our simulation, $s_0 = \{p_{-1}, v_{-1}, v_0\}$, are randomized uniformly over $\mathbb{P} \times \mathbb{V} \times \mathbb{V}$.

Specification of Exploration Rates. We adopt an exponentially time-declining state-dependent exploration rate for informed AI speculators,

$$\varepsilon_{t(v)} = e^{-\beta t(v)}, \tag{4.3}$$

[14] Using different initial values for the Q-matrix does not significantly alter the results. For example, another strategy is to use optimistic initial values, initializing the Q-matrix with high values that subsequent iterations tend to reduce. This approach helps Q-learning algorithms to explore all actions multiple times early on, resulting in early improvement in estimated action values. Thus, setting optimistic initial values is roughly equivalent to promoting thorough exploration early in the learning phase and exploitation later.

where the parameter $\beta > 0$ governs the speed that informed AI speculators' exploration rate diminishes over time and the variable $t(v)$ captures the number of times that the exogenous state $v \in \mathbb{V}$ has occurred in the past. The specification (4.3) implies that initially, Q-learning algorithms are almost always in the exploration mode, choosing actions randomly. However, as time passes, Q-learning algorithms gradually switch to the exploitation mode. The variable $t(v)$ implies that the exploration rate is state dependent, which ensures that informed AI speculators can sufficiently explore their actions for all grid points in $\mathbb{V}$.

Parameter Choice. The parameters used in our numerical experiments are categorized into four groups based on their roles. "Environment parameters" describe the underlying economic environment and, importantly, none of these values is known to the informed AI speculators and the market maker. "Preference parameters" encompass the discount rate for informed AI speculators and the weight assigned to the pricing error term by the adaptive market maker. "Discretization parameters" detail the methods used to discretize the system for numerical simulation, such as the number of discrete grid points and parallel simulation sessions. The "hyperparameters" are crucial for controlling the machine learning process. Below, we describe the choices of parameter values for the baseline experiments.

We begin with the environment parameters. We normalize $\overline{v} = 1$ and $\sigma_v = 1$ across all experiments, without loss of generality. In the baseline economic environment, we set $I = 2$ and $\xi = 500$, and consider two values of σ_u, with $\sigma_u = 10^{-1}$ and $\sigma_u = 10^2$ representing environments with low and high noise trading risk, respectively. In Section 5 and Online Appendix 1.9, we extensively study the implications of different values for these parameters.

The preference parameters are chosen to make the experiments relevant to the high-frequency trading settings in reality. We fix the value of θ at 0.1 throughout our simulation experiments to capture the primary concern of market makers with inventory cost management in these settings. Additionally, we set ρ at a relatively high level, $\rho = 0.95$, to reflect the high trading frequency. We explore the implications of varying ρ values in Section 5.

We now turn to the discretization parameters. We use $n_v = 10$ grid points to approximate the normal distribution of v_t. Under our discretization, the standard deviation of v_t is $\widehat{\sigma}_v =$

$\sqrt{\sum_{k=1}^{N} \mathcal{P}(v_k)(v_k - \overline{v})^2} = 0.938$, which is close to the theoretical value $\sigma_v = 1$.[15] We set $\iota = 0.1$ so that informed AI speculators can go beyond the theoretical bounds of order flows by 10%. We choose $n_x = 15$ and $n_p = 31$. These grid points are sufficiently dense to capture the economic mechanism we are interested in.[16] All results remain robust when choosing larger values for n_v, n_x, n_p, or ι, provided that the hyperparameters, α and β, are adjusted accordingly to ensure sufficiently good learning outcomes. However, the use of denser grids increases the time required for Q-learning algorithms to fully converge. We set $T_m = 10,000$ to allow market makers to accumulate enough time-series data to estimate their optimal pricing rules effectively. While increasing T_m does not alter any quantitative results, it does increase the computational burden.

Finally, we discuss our choice of the hyperparameters α and β. The hyperparameters that control the learning process of Q-learning algorithms are set at $\alpha = 0.01$ and $\beta = 5 \times 10^{-7}$. All results are robust to choosing different values of α and β so long as they are in the reasonable range that ensures sufficiently good learning outcomes. Our baseline choice of $\beta = 5 \times 10^{-7}$ implies that any action $x \in \mathbb{X}$ is visited purely by random exploration by $\frac{n_v}{n_x} \frac{1}{1 - \exp(-5 \times 10^{-7})} \approx 1,333,333$ times on average before exploration completes. In Sections 5.6 and 6.1, we conduct experiments with varying values of α and β. We also explore scenarios where informed AI speculators adopt different values of α. In Section 6.2, we develop a two-tier Q-learning algorithm that enables informed AI speculators to learn and optimally choose α.

Convergence. We adopt a stringent criterion for convergence, requiring that all informed AI speculators' optimal strategies remain unchanged for $1,000,000$ consecutive periods in a single session. Additionally, all $N_{sim} = 1,000$ independent parallel simulation sessions must continue running until every session meets this convergence criterion. The number of periods required to reach convergence varies considerably across experiments, influenced by the specific choices of environment parameters and hyperparameters. Additionally, even within the same experiment, the number of periods needed can differ significantly across the $N_{sim} = 1,000$ simulation sessions due to the path of realized values of random variables. Across all experiments we conducted, the

[15] In the remainder of this paper, the non-collusive Nash equilibrium and perfect cartel equilibrium are computed using $\widehat{\sigma}_v$, to ensure consistency with the discretization scheme of v_t used in the simulation experiments.

[16] Our choice of $n_p \approx 2n_x$ ensures that, all else equal, a one-grid point change in one informed AI speculator's order will result in a change in price p_t over the grid defined by $\mathbb{P}$. If the grid defined by $\mathbb{P}$ is coarser, informed AI speculators will not be able to detect small deviations of peers even in the absence of noise, which in turn lowers the possibility of algorithmic collusion through price-trigger strategies.

range of periods needed to achieve convergence spans from approximately 20 million to about 50 billion.[17]

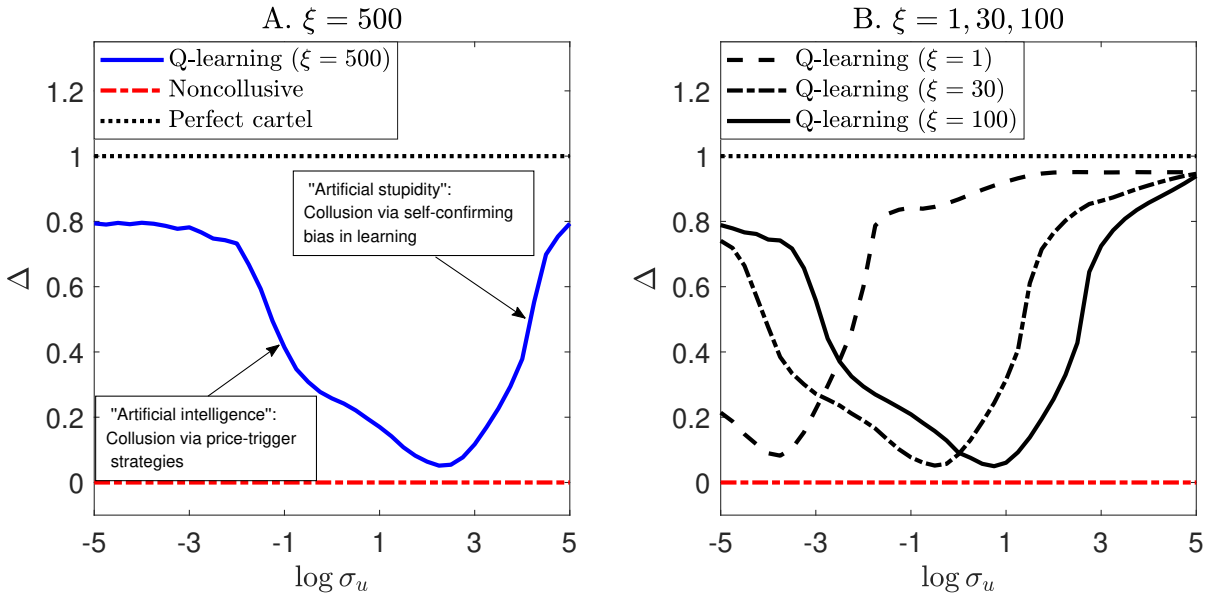
5 Impact of AI on Trading Equilibrium: Experimental Outcomes

In this section, we present the outcomes of simulation experiments that explore the behavior of AI-powered trading algorithms within the theoretical laboratory framework established in Sections 3 and 4. In Subsection 5.1, we vary the level of noise trading risk to demonstrate its U-shaped relation with the equilibrium supra-competitive trading profitability of informed AI speculators. In Subsection 5.2, we demonstrate that in environments characterized by low price efficiency and low noise trading risk, informed AI speculators can learn to employ price-trigger strategies to achieve and maintain collusive, supra-competitive trading profits, aligned with the theoretical benchmark set forth by Proposition 3.2. Conversely, in the same subsection, we show that in environments with high noise trading risk—even with low price efficiency—informed AI speculators struggle to sustain collusive supra-competitive trading profits using price-trigger strategies. Instead, they consistently achieve supra-competitive profits through their homogenized learning bias, a steady state best described by an experience-based equilibrium as in Proposition 3.2. In Subsection 5.3, we elaborate on the intuition behind AI collusion through these two distinct mechanisms. In Subsections 5.4, 5.5, and 5.6, we explore how the number of informed AI speculators, the subjective discount rate, and the hyperparameters impact the trading profitability of AI speculators and market efficiency. Each analysis is conducted under both scenarios: one involving AI collusion through price-trigger strategies and the other involving self-confirming bias in learning.

5.1 U-Shaped Profitability in AI-Driven Collusive Trading

The theoretical benchmarks established in Section 3.2 indicate that an AI-driven collusive trading equilibrium can robustly emerge through two different mechanisms. The dominant mechanism is contingent upon the risk of noise trading, captured by σ_u, and the efficiency of prices, governed

[17] To accelerate computations, our programs are written in C++, using $-O2$ to optimize the compiling process. The C++ program operates on a high-powered computing server cluster equipped with a total of 400 CPU cores. Depending on the total number of iterations required to reach convergence, completing all N_{sim} simulation sessions in one experiment can take anywhere from 1 minute to 6 hours.



Note: The blue solid line plots the average Δ^C across $N_{sim} = 1,000$ simulation sessions as $\log \sigma_u$ varies along the x-axis. The red dash-dotted and black dotted lines represent the theoretical benchmarks of the noncollusive Nash and perfect cartel equilibria, respectively. The other parameters are set according to the baseline economic environment described in Section 4.2.

Figure 1: Δ^C for $\log \sigma_u \in [-5, 5]$.

by ξ^{-1}. According to Propositions 3.1 through 3.4, we expect the AI-driven collusive trading profitability Δ^C to lie between 0 and 1, displaying a U-shaped relationship with noise trading risk. Panel A of Figure 1 illustrates this by plotting the average Δ^C as $\log \sigma_u$ varies from -5 to 5 along the x-axis. The black dotted line represents the theoretical benchmark for the perfect cartel ($\Delta^M \equiv 1$), and the red dash-dotted line represents the benchmark for the non-collusive Nash equilibrium ($\Delta^N \equiv 0$). The blue solid line represents the steady-state collusive capacity Δ^C reached by AI trading algorithms considered in Section 4.1. It is shown that a collusive equilibrium with significant supra-competitive profits arises when noise trading risk σ_u is either low or high. Notably, when the noise trading risk is low, collusion capacity Δ^C decreases with σ_u, whereas when the noise trading risk is high, collusion capacity Δ^C increases with σ_u. This suggests that two different mechanisms drive the collusive equilibrium at high and low levels of noise trading risk. Specifically, when the noise trading risk is low (i.e., $\log \sigma_u \leq 2$), informed AI speculators use price-trigger strategies to maintain collusion and achieve supra-competitive profits. The inverse relationship between the average Δ^C and σ_u is consistent with the theoretical benchmark for a collusive Nash equilibrium sustained through price-trigger strategies, as described in Proposition 3.4. This situation is referred to as "collusion through artificial intelligence." Conversely, when

the noise trading risk is high (i.e., $\log \sigma_u \geq 3$), informed AI speculators achieve supra-competitive profits through self-confirming bias in learning. The positive relationship between the average Δ^C and σ_u reflects a fundamental property of RL algorithms: the self-confirming bias in learning becomes more pronounced as the noise trading risk σ_u increases (see the discussion in Section 5.2 and Online Appendix 3.1). This situation is referred to as "collusion through artificial stupidity."

Panel B of Figure 1 demonstrates that the U-shaped relationship between the average Δ^C and $\log \sigma_u$ is robust across different values of ξ. In our baseline calibration, $\xi = 500$, the sensitivity of order flow z_t to price level p_t may seem large. information-insensitive investors typically hold a substantial amount of the asset already, so $\xi = 500$ can reflect a very low price elasticity of asset demand, consistent with the estimation by Koijen and Yogo (2019).

5.2 Impulse Response Evidence of Two Mechanisms for AI Collusion

In this subsection, we use impulse response analyses to provide direct evidence on the two mechanisms for AI collusion across different trading environments, as suggested by the theoretical benchmarks and the consistent U-shaped relationship between Δ^C and $\log \sigma_u$ in Figure 1. Furthermore, we elaborate on how RL algorithms reach and sustain these two different forms of collusive equilibrium.

We begin by showing that, in scenarios with low noise trading risk, informed AI speculators can learn to sustain collusive, supra-competitive trading profits through price-trigger strategies without any form of agreement, communication, or even intent. This equilibrium resembles the collusive Nash equilibrium sustained by price-trigger strategies, described in Propositions 3.1 and 3.2, although it does not fully adhere to subgame perfect Nash requirements.[18] Conversely, in scenarios with high noise trading risk, informed AI speculators still manage to maintain collusive, supra-competitive trading profits, but through a different mechanism, which is referred to as self-confirming bias in learning, as described in Proposition 3.5.[19]

To assess if informed AI speculators learn to employ price-trigger strategies similar to those of rational-expectations agents in the collusive Nash equilibrium, as described in Propositions 3.1

[18] Our tests indicate that this equilibrium is approximately Nash, meaning that no local deviation is preferred.

[19] In both scenarios, the equilibrium is proven to be an experience-based equilibrium, based on the formal tests proposed by Fershtman and Pakes (2012). Details of these tests are provided in Online Appendix 1.2. This is unsurprising, as experience-based equilibrium is a broader concept that includes subgame perfect Nash equilibrium as a special case.

and 3.2, we analyze the impulse response function in the steady state. Specifically, we evaluate how the trained informed AI speculators respond to an exogenous shock to the noise order flow, which directly impacts the market price of the asset through the endogenous pricing rule of the market maker. At $t = 0$, every single one of the $N_{sim} = 1,000$ simulation paths has converged. Simultaneously, the market price of the asset, p_t, is determined by the market maker's adaptive pricing rule, which responds to the random variables v_t and u_t along each simulation path, independently across N_{sim} different parallel simulation paths.

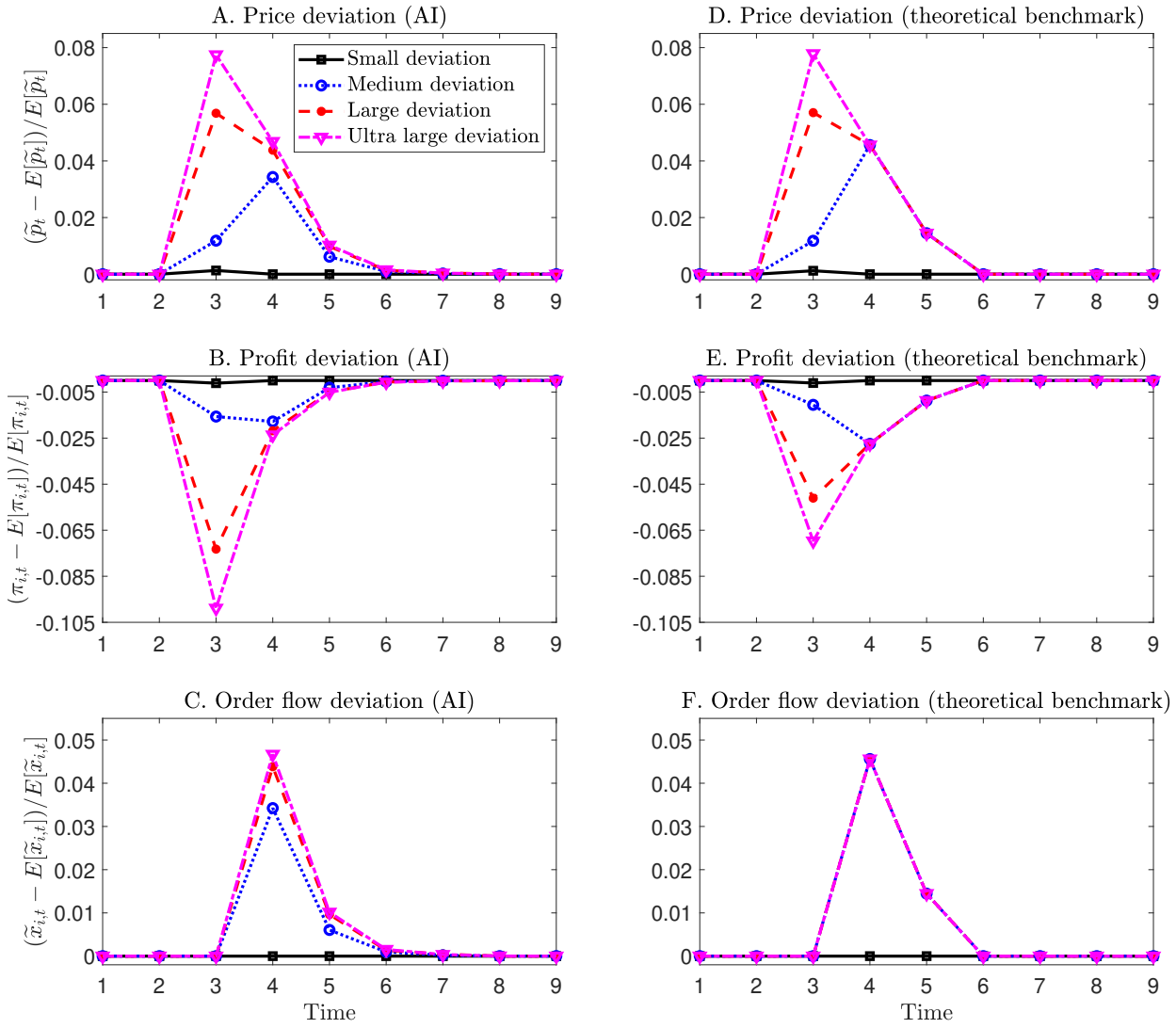
By our design, at $t = 3$, an unexpected exogenous shock u_{shock} is introduced to the noise order flow u_t, affecting all N_{sim} simulation paths simultaneously and uniformly. This shock is intended to adversely impact the trading profits of informed AI speculators, with $u_{\text{shock}} > 0$ if $v_t > \overline{v}$ and $u_{\text{shock}} < 0$ if $v_t < \overline{v}$. Consequently, the market price p_t unexpectedly rises if $v_t > \overline{v}$ and unexpectedly decreases if $v_t < \overline{v}$, with the extent of the price change determined by the magnitude of the noise trading risk shock u_{shock}.

Panels A through C of Figure 3 illustrate the impulse response dynamics in a scenario with low noise trading risk ($\sigma_u = 10^{-1}$), while panels D through F of the same figure depict those in a scenario with high noise trading risk ($\sigma_u = 10^{2}$). Each impulse-response curve in a panel represents the average impulse response dynamics across N_{sim} independent simulation paths.[20] The confidence band of path-by-path impulse response dynamics across N_{sim} simulation paths is presented in Online Appendix 1.6.

Low Noise Trading Risk ($\sigma_u = 10^{-1}$). In environments with low noise trading risk, specifically with $\sigma_u = 10^{-1}$, across N_{sim} parallel simulation paths, the average value of Δ^C is about 0.75, and the average trading profit of informed AI speculators is about 10% higher than that in the non-collusive equilibrium.

We consider exogenous shocks of different magnitudes. In the scenario with "small deviation," $|u_{\text{shock}}|$ is roughly 0.25% of the average magnitude of informed AI speculators' order flow $|x_{i,t}|$, generating a small impact on the asset's price p_t at $t = 3$. In the scenario with "medium deviation," "large deviation," and "ultra large deviation," $|u_{\text{shock}}|$ is about 2.5%, 11.5%, and 15.0% of the

[20] Each of the N_{sim} simulation paths averages 10,000 simulations to smooth out the randomness of v_t and u_t, ensuring a reasonable comparison with the impulse response analysis based on the deterministic model of Calvano et al. (2020), which has no information asymmetry and focuses on a non-stochastic economic environment.



Note: All the plots are for the scenario of low noise trading risk with $\sigma_u = 10^{-1}$. Panels A through C show the IRF following a uniform exogenous shock u_{shock} in simulation experiments using Q-learning algorithms. Panels D through F present the corresponding IRFs based on the theoretical benchmark, where rational-expectations speculators achieve a subgame perfect Nash equilibrium characterized by collusion sustained through price-trigger strategies.

Figure 2: IRF following uniform exogenous shock u_{shock} for $\sigma_u = 10^{-1}$ under Q-learning (left) or theoretical benchmark (right).

average magnitude of informed AI speculators' order flow $|x_{i,t}|$, respectively, resulting in much larger changes in p_t. These larger deviations may trigger the non-collusive reversion punishment, whereas the small deviation may not.

Panel A of Figure 2 illustrates the percentage deviation of the asset's price from its long-run mean, defined as $(\widetilde{p}_t - \mathbb{E}[\widetilde{p}_t])/\mathbb{E}[\widetilde{p}_t]$, where $\widetilde{p}_t = (p_t - \overline{v}) \times \text{sgn}(v_t - \overline{v})$ and $\text{sgn}(\cdot)$ is the sign function. The sign function ensures that $\widetilde{p}_t > 0$. Due to the exogenous shock, the asset's price

deviates from its long-run mean at $t = 3$, with the size of the deviation increasing with the magnitude of the exogenous shock u_{shock}. Panel B of Figure 2 shows the percentage deviation of average profits from their long-run mean for each informed AI speculator, defined as $(\pi_{i,t} - \mathbb{E}[\pi_{i,t}])/\mathbb{E}[\pi_{i,t}]$. This plot demonstrates that at $t = 3$, the price deviation reduces the informed AI speculator's profits, with the impact increasing with the magnitude of the percentage price deviation shown in Panel A. Panel C of Figure 2 shows the percentage deviation of order flow from its long-run mean for each informed AI speculator, defined as $(\widetilde{x}_{i,t} - \mathbb{E}[\widetilde{x}_{i,t}])/\mathbb{E}[\widetilde{x}_{i,t}]$, where $\widetilde{x}_{i,t} = x_{i,t} \times \text{sgn}(v_t - \overline{v})$. The sign function ensures that $\widetilde{x}_{i,t} > 0$. The deviation of order flows is zero at $t = 3$ because price deviation only occurs until $t = 3$.

We now shift our focus to the response of informed AI speculators following the price deviation at $t = 3$. Starting with $t = 4$, the responses exhibit two defining features of price-trigger strategies, as outlined in the theoretical benchmark: (i) there is, on average, no response if the price deviation at $t = 3$ is small (i.e., "small deviation" captured by the black solid curve), and (ii) if the price deviation at $t = 3$ is sufficiently large, AI speculators respond with roughly the same aggressive trading strategies, regardless of the deviation's magnitude (i.e., "medium deviation," "large deviation," and "ultra large deviation," captured by the blue dotted, red dashed, and purple dot-dashed curves, respectively).

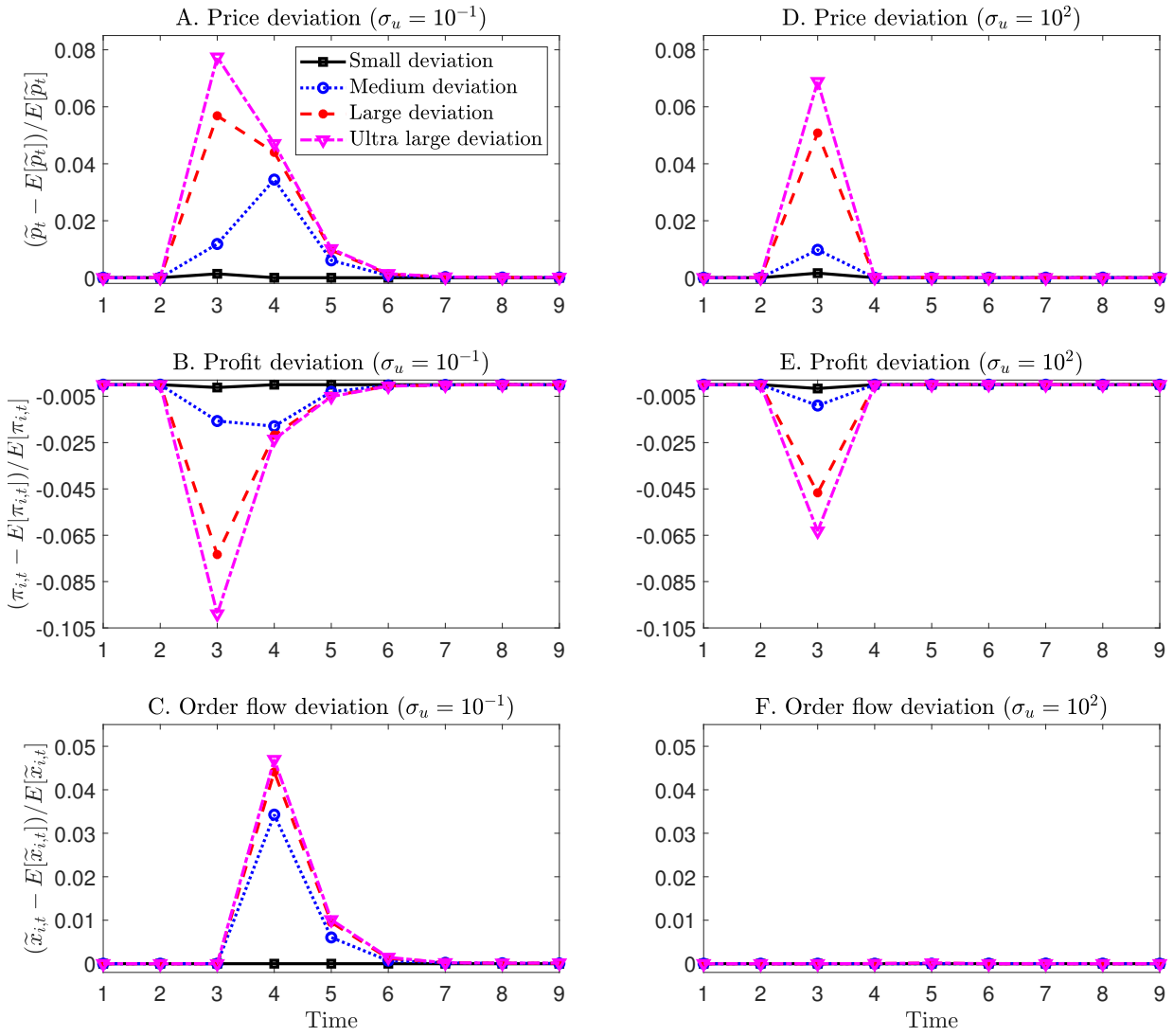
Interestingly, Panel A of Figure 2 shows that for the cases of large and ultra large price deviations, the percentage deviations of the asset's price at $t = 4$ decrease relative to the previous period but remain substantially higher than the long-run mean. For the case of medium deviation, the percentage deviation of the asset's price at $t = 4$ is higher than in the previous period. Crucially, the price deviations have similar magnitudes in the cases of medium, large, and ultra large deviations due to the similar magnitude of order flow deviations at $t = 4$, as shown in Panel C. In contrast, for the case of small deviation, both the asset's price and informed AI speculators' profits revert to the long-run mean at $t = 4$.

Although the Q-learning algorithms only trace the one-period lagged market price p_{t-1} and fundamental value v_{t-1} for their decisions at period t, the punishment can last more than one period. Panels A through C of Figure 2 show that informed AI speculators continue exerting punishment at $t = 5$, though it is significantly weaker on average than at $t = 4$. By $t = 6$, they typically stop punishing, and trading behavior along with the market price begins reverting to

the long-run mean. This pattern indicates that informed AI speculators sustain the collusive equilibrium through price-trigger strategies with a punishment scheme that usually lasts two periods.

To provide further direct evidence that the behavior of informed AI speculators in equilibrium closely resembles a subgame perfect Nash equilibrium characterized by collusion sustained by price-trigger strategies, we plot the impulse responses for AI-powered trading in the left column of Figure 2 alongside the corresponding theoretical benchmarks, as described in Propositions 3.1 and 3.2, in the right column of the same figure. Specifically, in Panels D through F of Figure 2, we plot the impulse response of an exact price-trigger strategy to the same exogenous shock to noise trading risk, u_{shock}. For a meaningful and informative comparison, when plotting Panels D through F, we choose the same magnitudes of price deviations at $t = 3$ as those in the simulation experiments in Panels A to C. Moreover, all overlapping parameters take the same values as in the simulation experiments. The parameters (T, ω, η) are unique to the price-trigger strategy punishment scheme and do not apply to the Q-learning simulations. We set $T = 2$ to match the two-period punishment observed in the Q-learning experiments, $\omega = 2.826$ to achieve an average profitability Δ^C around 0.75, and $\eta = 0.327$ to match the average order flow deviation of the "Ultra large deviation" case at $t = 4$ in the Q-learning simulations. This side-by-side comparison reveals a strong similarity between AI-powered trading and the corresponding theoretical benchmarks in collusion through price-trigger strategies: (i) with a small price deviation at $t = 3$ (the black solid line), informed speculators do not change their order flows in subsequent periods, in both the model and simulations; (ii) with medium, large, and ultra large deviations (represented by the other three lines), informed speculators increase their order flows by roughly the same magnitude at $t = 4$, regardless of the different magnitudes of price deviations at $t = 3$; and, (iii) the average strength of the punishment decays at roughly the same rate in both the model and simulation experiments.

To determine if the price-trigger strategy by informed AI speculators in Panels A through C of Figure 2 is responsible for the collusive, supra-competitive trading profitability observed in Figure 1 for low noise trading risk, we need to disable the AI speculators' ability to use market prices as a monitoring tool. This can be achieved by removing the lagged market price p_{t-1} from the state variable s_t for decisions at period t. Indeed, we find that even in environments with both



Note: All the plots are from simulation experiments using Q-learning algorithms. Panels A through C show the IRF following a uniform exogenous order flow deviation $x_{i,\text{shock}}$ for the scenario of low noise trading risk ($\sigma_u = 10^{-1}$). Panels D through F present the corresponding IRFs for the scenario of high noise trading risk ($\sigma_u = 10^{2}$).

Figure 3: IRF following uniform exogenous shock u_{shock} for scenarios $\sigma_u = 10^{-1}$ (left) and $\sigma_u = 10^{2}$ (right).

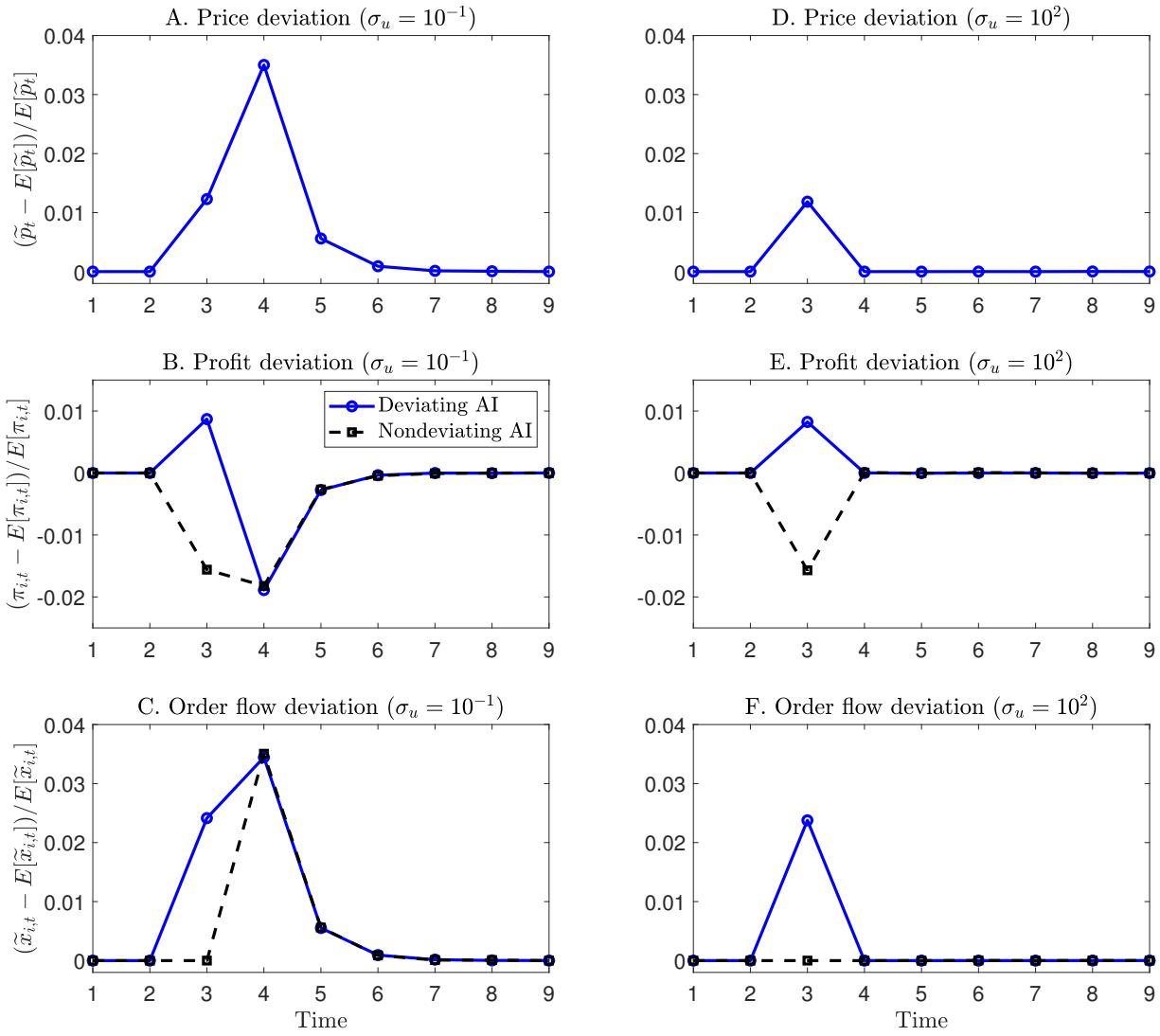
low price efficiency and noise trading risk, the collusion capacity, captured by Δ^C, is roughly zero, regardless of the price efficiency level (ξ^{-1}) and noise trading risk level (σ_u).

High Noise Trading Risk ($\sigma_u = 10^2$). We now examine whether the collusive, supra-competitive trading profitability observed when the noise trading risk σ_u is large in Figure 1 is also attributable to price-trigger strategies, as it is in the scenario where the noise trading risk σ_u is low. The setup of simulation experiments in Figure 3 is the same as that in Figure 2 with Panels A through

C exactly identical for a straightforward comparison. In Panels D through F of Figure 3, we investigate the average IRF over the $N_{sim} = 1,000$ simulation paths in the environment with high noise trading risk (i.e., $\sigma_u = 10^2$). This side-by-side comparison, particularly contrasting Panel C with Panel F of Figure 3, reveals that informed AI speculators do not respond at all to the exogenous shock to noise trading flow (u_{shock}) when noise trading risk is high, let alone respond according to price-trigger strategies. This finding is consistent with the theoretical result of Proposition 3.1, which states that a collusive Nash equilibrium sustained through price-trigger strategies does not exist in an environment with high noise trading risk.

How do informed AI speculators still achieve and sustain supra-competitive profits, despite being unable to learn and use price-trigger strategies? We demonstrate that informed AI speculators can still establish a collusive equilibrium as described in Definition 3.1, particularly a collusive experience-based equilibrium sustained by self-confirming bias in learning. To illustrate this, we study the IRF following a unilateral deviation of one informed AI speculator in both scenarios of low ($\sigma_u = 10^{-1}$) and high ($\sigma_u = 10^2$) noise trading risk in Figure 4. Specifically, we exogenously force one informed AI speculator, labeled as i, to make a one-period deviation from its learned optimal strategy at $t = 3$, uniformly across all of the $N_{sim} = 1,000$ simulation paths. This one-period deviation at $t = 3$ is directed to increase the contemporaneous trading profit of the deviating speculator. Specifically, we exogenously increase the order flow of the deviating speculator by $x_{i,\text{shock}}$ if $v_t > \overline{v}$ and reduce its order flow by $x_{i,\text{shock}}$ if $v_t < \overline{v}$.

Panels A through C of Figure 4 show the IRF following the unilateral deviation of AI speculator i (blue solid curve) at $t = 3$ for the scenario of low noise trading risk with $\sigma_u = 10^{-1}$. Panel C illustrates the exogenous forcing of AI speculator i to deviate by trading more aggressively, while the other AI speculator (black dashed curve) maintains their trading behavior. As shown in Panel A, this aggressive trading by AI speculator i pushes up the market price p_t at $t = 3$. Panel B shows that the deviating AI speculator (blue solid curve) gains greater profits, while the non-deviating AI speculator (black dashed curve) loses profits at $t = 3$. According to Definition 3.1, these IRF results reinforce the findings of Figure 1, demonstrating that informed AI speculators can interact and learn to sustain a collusive equilibrium in environments of low noise trading risk. More importantly, the responses of informed AI speculators to this unilateral deviation in the subsequent periods starting from $t = 4$ further reinforce the findings of Figures 2 and



Note: All the plots are from simulation experiments using Q-learning algorithms. Panels A through C show the IRF following a uniform exogenous order flow deviation $x_{i,\text{shock}}$ for the scenario of low noise trading risk ($\sigma_u = 10^{-1}$). Panels D through F present the corresponding IRFs for the scenario of high noise trading risk ($\sigma_u = 10^{2}$).

Figure 4: IRF following unilateral deviation in trading order flows $x_{i,\text{shock}}$ for $\sigma_u = 10^{-1}$ (left) and $\sigma_u = 10^{2}$ (right).

3, demonstrating that the collusive equilibrium is indeed sustained by price-trigger strategies, resembling the behavior of a subgame perfect Nash equilibrium. Specifically, at $t = 4$, Panel C shows that both AI speculators respond with equally aggressive trading behavior, on average, as punishment for the deviation. As shown in Panel B, this results in both AI speculators losing profits at $t = 4$ due to the skyrocketing market price.

In contrast, Panels D through F of Figure 4 show the IRF following the unilateral deviation of AI speculator i (blue solid curve) at $t = 3$ for the scenario of high noise trading risk with

$\sigma_u = 10^2$. Panel F shows AI speculator i being forced to trade more aggressively, while the other AI speculator (black dashed curve) maintains their usual trading behavior. Panel D shows that this aggressive trading by AI speculator i pushes up the market price p_t at $t = 3$. Similar to Panel B, Panel E shows that the deviating AI speculator (blue solid curve) gains greater profits, while the non-deviating AI speculator (black dashed curve) loses profits at $t = 3$. According to Definition 3.1, these IRF results reinforce the findings of Figure 1, demonstrating that informed AI speculators robustly reach a collusive equilibrium in environments with high noise trading risk. Although the immediate reactions at $t = 3$ mirror those in the low noise trading risk environment shown in Panels A through C, the subsequent responses of informed AI speculators are completely different. Specifically, the deviating AI speculator automatically reverts to the mean trading order flow while the non-deviating AI speculator's behavior remains unaffected, as shown in Panel F. This occurs consistently, even though the deviating AI speculator takes advantage of the non-deviating AI speculator at $t = 3$, as shown in Panel E. This provides direct evidence that the collusive equilibrium in the high noise trading risk scenario is not a Nash equilibrium, and the persistent self-confirming bias in learning cannot be altered by new trial-and-error observations from a single period. Instead, the equilibrium is shown to be an experience-based equilibrium with self-confirming bias in learning, according to the formal tests proposed by Fershtman and Pakes (2012).[21]

Role of Informative-Insensitive Investors. In Figures 1 through 4, we examine the impact of different levels of noise trading risk σ_u on informed AI speculators' trading equilibrium, including their collusion capacity and the mechanisms behind AI collusion, with $\xi = 500$ kept fixed. As suggested by the theoretical benchmarks, such as Propositions 3.1 and 3.2, collusion through price-trigger strategies requires inefficient prices caused by strong presence of information-insensitive investors who absorb the trading order flows of informed AI speculators. Thus, when noise trading risk is low, the collusive, supra-competitive trading profits of informed AI speculators through "artificial intelligence" are primarily derived from trading against information-insensitive investors. Specifically, in our simulation experiments with Q-learning algorithms for the scenario $\sigma_u = 10^{-1}$, each informed AI speculator gains approximately 54 on average, which is derived

[21] For detailed information on the tests for experience-based equilibrium, refer to Online Appendix 1.2.

from the loss of information-insensitive investors, roughly 108, as the average trading profit of noise traders and market makers is nearly zero. By contrast, when noise trading risk is high, the collusive, supra-competitive trading profits of informed AI speculators through "artificial stupidity" are derived not only from trading against information-insensitive investors but also from trading against noise traders. Specifically, in our simulation experiments with Q-learning algorithms for the scenario $\sigma_u = 10^2$, each informed AI speculator gains approximately 54 on average, which is derived not only from the loss of information-insensitive investors (roughly 88) but also from the loss of noise traders (roughly 20), with the average trading profit of market makers still nearly zero. Notably, these results are consistent with the recent empirical findings of Chen, Peng and Zhou (2024), suggesting that the profits of AI-powered trading primarily arise from trading against the technical analysis sentiment of retail investors, as information-insensitive investors can be interpreted as retail investors who employ technical analysis in our model. The contrast between scenarios $\sigma_u = 10^{-1}$ and $\sigma_u = 10^2$ further demonstrates the distinct mechanisms behind AI collusion. To highlight this difference, we conducted additional simulation experiments with Q-learning algorithms for the scenario $\sigma_u = 2.5 \times 10^2$. These experiments show that information-insensitive investors can trade alongside informed AI speculators in the same direction when noise traders make substantial order flows in the losing direction. In this case, each informed AI speculator gains approximately 54.5, and information-insensitive investors gain roughly 16 on average, derived from the loss of noise traders (roughly 125), with the average trading profit of market makers remaining nearly zero.

Furthermore, according to the theoretical benchmarks, such as Propositions 3.1 and 3.2, the existence of collusive equilibrium through price-trigger strategies requires ξ to be sufficiently large, resulting in sufficiently inefficient prices. Although we focus on examining the role of noise trading risk σ_u on AI collusion in Figures 1 through 4, we show in Online Appendix 1.9 that when ξ is low, collusion through self-confirming bias in learning, rather than through price-trigger strategies, arises robustly regardless of the level of noise trading risk σ_u.

5.3 Intuition Behind AI Collusion

Collusion Through Price-Trigger Strategies When σ_u Is Low. We first elaborate the intuition behind collusion through price-trigger strategies in the scenario of low noise trading risk. Specifi-

cally, we provide insight into how RL algorithms learn from their interactions to achieve collusion sustained by price-trigger strategies. Although this price-trigger strategy appears to be based on logical thinking and the deviation-and-punishment causal response, it is actually driven by the RL algorithms autonomously learning to make optimal decisions based solely on pattern recognition.

For illustrative purposes, consider the baseline case with $\sigma_u = 0$ and $v_t \equiv \overline{v}$. In this scenario, the state variable is $s_t = p_{t-1}$. Intuitively, when at least one AI speculator trades aggressively, denoted by x_H, the price is driven up to a high level, denoted by p_H. Conversely, if both AI speculators trade conservatively, denoted by x_L, the price remains low, denoted by p_L.

To grasp the key idea, assume that both AI speculators adopt conservative trading behavior, x_L, as their optimal strategy at time t when $s_t = p_L$. This means the Q-function initially has a higher value at (p_L, x_L) than at (p_L, x_H) for both AI speculators. If there is no exploration at all, the system will remain in conservative trading, x_L, with a consistently low price, p_L. Consequently, aggressive trading, x_H, or a high price, p_H, will never occur on the equilibrium path. Without exploration, price-trigger strategies cannot be learned or implemented.

Suppose one exploration occurs at t, causing one AI speculator to deviate from x_L to x_H, shifting the state from p_L to p_H. If another exploration occurs in the subsequent period $t+1$, prompting one AI speculator to choose x_H, the state remains at p_H at $t+2$. The AI speculator choosing x_H gains more profits at the expense of the other, who chooses x_L. Exploitation will ensure that this combination of trading behaviors persists, keeping the state at p_H in subsequent periods, until both AI speculators find it optimal to trade aggressively, x_H, in state p_H.

As a result, the optimal value function $\max_{x' \in \mathcal{X}} \widehat{Q}_{i,t}(p_H, x')$ conditional on the state p_H is low. Therefore, if ρ is sufficiently high, the recursive relation (2.4) indicates that conservative trading behavior x_L can be preferred over x_H in the state p_L, even though deviating to x_H leads to higher one-period trading profits with the other speculator remaining at x_L. This is because AI speculators understand that choosing x_H would shift the state to p_H and result in a low continuation value, $\rho \max_{x' \in \mathcal{X}} \widehat{Q}_{i,t}(p_H, x')$. This, in turn, justifies the initial assumption about the Q-function.

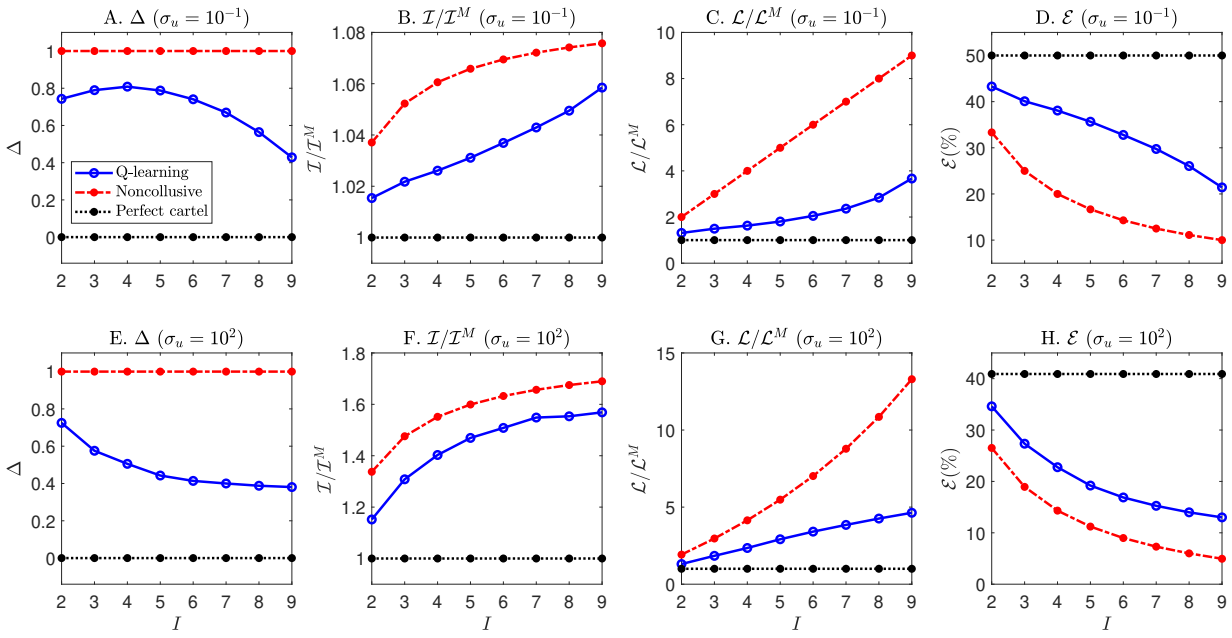
From the intuition explained above, it is evident that exploration is crucial not only for approximating the true Q-function but also for enabling informed AI speculators to learn and sustain collusion through price-trigger strategies in low noise trading environments. Exploration allows

RL algorithms to acquire off-equilibrium-path information sets, facilitating the implementation of price-trigger strategies. These strategies help maintain collusive, supra-competitive trading profits through the threat of "punishment."

Collusion Through Self-Confirming Bias When σ_u Is High. As σ_u increases to a high level, the mechanism behind the price-trigger strategies explained above becomes invalid because the state variable p_t is primarily driven by noise trading flows u_t rather than by the trading behavior of informed AI speculators. However, a distinct mechanism for AI collusion begins to emerge. We now elaborate on the intuition behind collusion through self-confirming bias in learning in the scenario of high noise trading risk. Specifically, we provide insight into how RL algorithms learn from their interactions to achieve collusion sustained by self-confirming bias, where AI speculators undervalue aggressive trading strategies, perpetuating an incorrect system of outcome evaluation that remains uncorrected.

Aggressive trading behavior, x_H, is prone to disastrous profit outcomes when large noise trading flows move in the same direction, leading the algorithm to label x_H as disastrous. Consequently, exploitation prevents the algorithm from revisiting x_H, resulting in a persistent undervaluation. Conversely, x_H can also yield exceptional profits when noise trading flows move in the opposite direction, prompting the algorithm to label x_H as favorable. This prompts the algorithm to learn about x_H repeatedly, leading to an unbiased valuation. However, the asymmetric effect of exploitation on learning aggressive trading behavior ultimately results in the persistent dominance of the undervaluation effect. This systematic undervaluation leads informed AI speculators to settle on conservative trading strategies in the steady state, preventing them from revisiting aggressive strategies and correcting their evaluation of off-equilibrium-path outcomes.

From the intuition explained above, unlike the mechanism behind collusion through price-trigger strategies when σ_u is low, exploration is not as crucial for collusion through self-confirming bias in learning when σ_u is high. Instead, exploitation, a defining characteristic of RL algorithms alongside exploration, alone plays a vital role.



Note: The blue solid line plots the average values of π^C/π^M, $\mathcal{I}^C/\mathcal{I}^M$, $\mathcal{L}^C/\mathcal{L}^M$, and $\mathcal{E}^C$ across $N_{sim} = 1{,}000$ simulation sessions as the number of informed AI speculators I varies. The red dash-dotted and black dotted lines represent the theoretical benchmarks of the noncollusive Nash and perfect cartel equilibria, respectively. Panels A to D represent the environment with low noise trading risks (i.e., $\sigma_u = 10^{-1}$); Panels E to H represent the environment with high noise trading risks (i.e., $\sigma_u = 10^2$). The other parameters are set according to the baseline economic environment described in Section 4.2.

Figure 5: Implications of the number of informed AI speculators.

5.4 Effect of the Number of Informed AI Speculators (I)

To study how the number of informed AI speculators affects their trading strategies, we increase I from 2 to 6 in the baseline economic environment. Panels A to D of Figure 5 focus on the environment with low noise trading risks (i.e., small σ_u). Panel A shows that as I increases, the relative profit π^C/π^M decreases from 0.97 to 0.87, indicating a decline in the extent of collusion among informed AI speculators. Moreover, panels B to D show that as I increases, the relative price informativeness $\mathcal{I}^C/\mathcal{I}^M$ and market liquidity $\mathcal{L}^C/\mathcal{L}^M$ increase whereas the magnitude of mispricing $\mathcal{E}^C$ decreases. These patterns are consistent with the prediction of the model, which suggests that informed speculators are less able to collude through price-trigger strategies when the number of informed speculators I increases (see Proposition 3.4).

For comparisons, in panels E to H of Figure 5, we focus on the environment with high noise trading risks (i.e., $\sigma_u = 10^2$). In this environment, informed AI speculators achieve supra-competitive profits due to homogenized learning biases, as discussed in Subsection 5.2. We

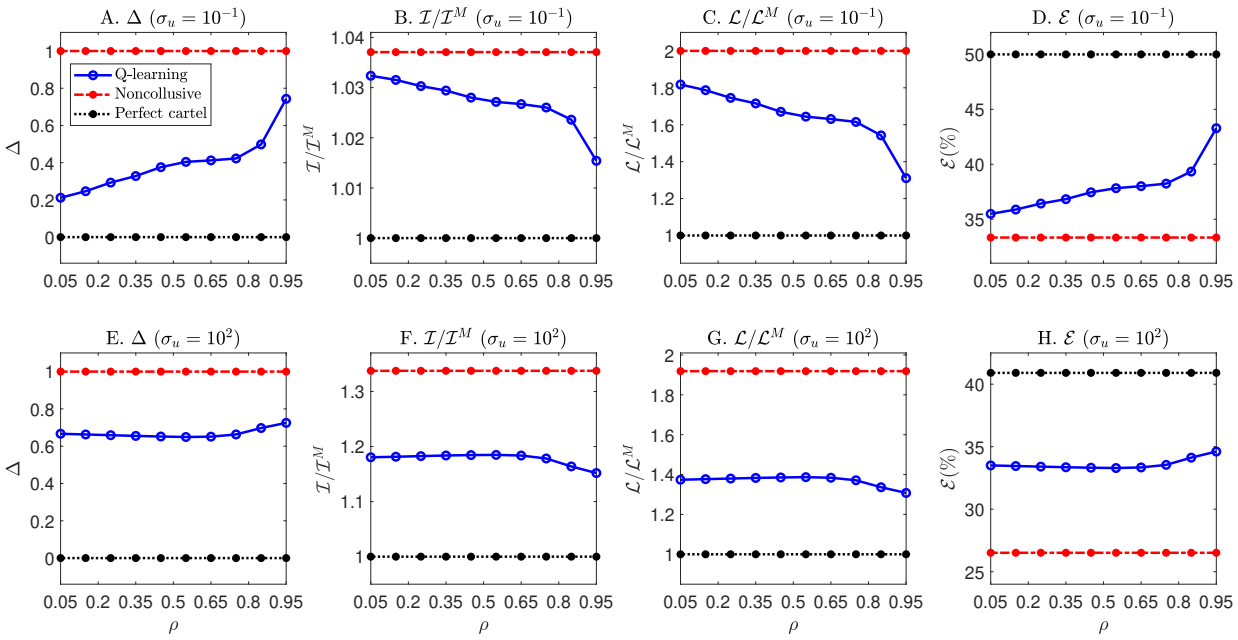
show that as I increases, the relative profit π^C/π^M decreases, the relative price informativeness $\mathcal{I}^C/\mathcal{I}^M$ and market liquidity $\mathcal{L}^C/\mathcal{L}^M$ increase whereas the relative mispricing $\mathcal{E}^C$ decreases. These patterns are remarkably similar to those in the environment with low noise trading risks (panels A to D of Figure 5), despite the difference in the underlying mechanisms that result in collusion. They suggest that the coordination through homogenized learning biases becomes more difficult to achieve when there are more informed AI speculators in the market. Intuitively, in the environment with high noise trading risks, the equilibrium degree of collusion is determined by the interaction of two counterveiling forces. One is the magnitude of learning biases, which is the mechanism that results in collusion. The other is the deviation gain from the collusive experience-based equilibrium. A larger deviation gain makes it more difficult for informed AI speculators to reach the collusive equilibrium because in the process of exploration (which, in essence, generates deviation behavior), AI speculators will more likely learn noncollusive trading strategies. As the number of informed AI speculators I increases, the deviation gain becomes larger, but the magnitude of learning biases remain unchanged,[22] reducing the capacity to collude.

5.5 Effect of Subjective Discount Rate (ρ)

To study how the subjective discount rate affects informed AI speculators' trading strategies, we vary ρ in the baseline economic environment. Panels A to D of Figure 6 focus on the environment with low noise trading risks (i.e., $\sigma_u = 10^{-1}$). Panel A shows that as ρ increases, the relative profit π^C/π^M increases. Moreover, panels B to D show that as ρ increases, the relative price informativeness $\mathcal{I}^C/\mathcal{I}^M$ and market liquidity $\mathcal{L}^C/\mathcal{L}^M$ decline whereas the relative mispricing $\mathcal{E}^C$ increases. These patterns are consistent with the prediction of the model, which suggests that informed speculators are able to collude on higher profits through price-trigger strategies as the subjective discount rate ρ increases (see Proposition 3.4).

Turning to the environment with high noise trading risks (i.e., $\sigma_u = 10^2$), panels E to H of Figure 6 show that as ρ increases, the relative profit π^C/π^M is roughly unchanged for $\rho \leq 0.65$ and increases slightly for $\rho > 0.65$. The relative price informativeness $\mathcal{I}^C/\mathcal{I}^M$, market liquidity

[22] As I increases, individual informed AI speculators' trading flows x_i decrease. However, in equation (IA.6) in Online Appendix 3.1, the trading flow x_i proportionally affects every term. Thus, the decrease in x_i does not affect the importance of the term $\alpha\lambda x_i \sum_{\tau=0}^{T}(1-\alpha)^{\tau} u_{t(T-\tau)}$, which causes learning biases, relative to other terms in the equation. This is why the magnitude of learning biases does not depend on I.



Note: The blue solid line plots the average values of π^C/π^M, $\mathcal{I}^C/\mathcal{I}^M$, $\mathcal{L}^C/\mathcal{L}^M$, and $\mathcal{E}^C$ across $N_{sim} = 1,000$ simulation sessions as the subjective discount rate ρ varies. The red dash-dotted and black dotted lines represent the theoretical benchmarks of the noncollusive Nash and perfect cartel equilibria, respectively. Panels A to D represent the environment with low noise trading risks (i.e., $\sigma_u = 10^{-1}$); Panels E to H represent the environment with high noise trading risks (i.e., $\sigma_u = 10^2$). The other parameters are set according to the baseline economic environment described in Section 4.2.

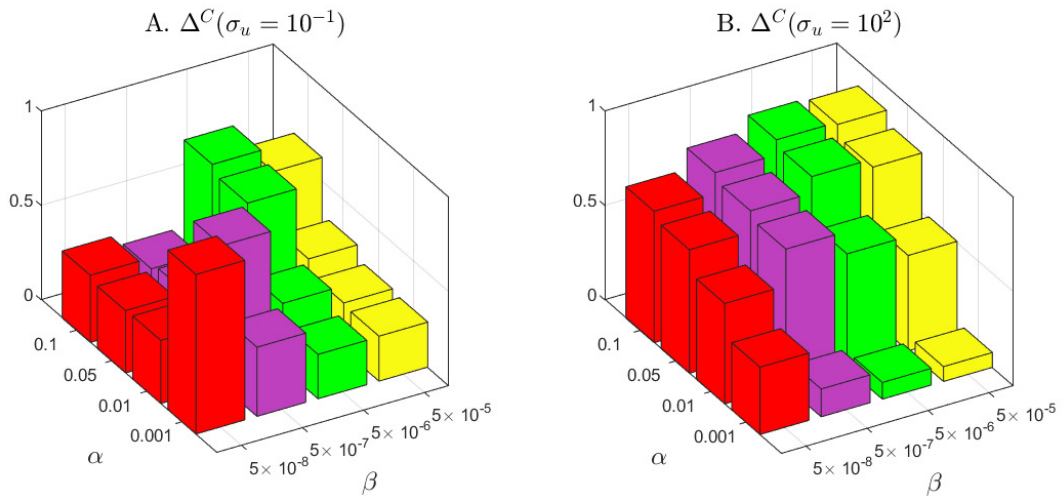
Figure 6: Implications of the subjective discount rate.

$\mathcal{L}^C/\mathcal{L}^M$, and the magnitude of mispricing $\mathcal{E}^C$ also stay roughly unchanged as ρ increases. The insignificant impact of ρ in this environment is due to the algorithmic property that ρ does not significantly affect the magnitude of learning biases (see Online Appendix 3.1). Because collusion is achieved through homogenized learning biases in this environment, the degree of collusion would be insensitive to ρ if the magnitude of learning biases does not change much with ρ.

5.6 Hyperparameters

The implementation of Q-learning algorithms is determined by the two key hyperparameters α and β, where the former determines the forgetting rate and the latter determines the decaying speed of exploration. We now study how α and β affect informed AI speculators' trading strategies in the baseline economic environment.

Panel A of Figure 7 plots the average Δ^C in the environment with low noise trading risks (i.e., $\sigma_u = 10^{-1}$) for different values of α and β. It is shown that, to make the learning process more



Note: Panel A plots Δ^C in the environment with low noise trading risks ($\sigma_u = 10^{-1}$); panel B plots Δ^C in the environment with high noise trading risks ($\sigma_u = 10^2$). The other parameters are set according to the baseline economic environment described in Section 4.2.

Figure 7: Implications of hyperparameters α and β on Δ^C.

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

For comparisons, panel B of Figure 7 plots the average Δ^C in the environment with high noise trading risks (i.e., $\sigma_u = 10^2$) for different values of α and β. Holding β unchanged at each value of

$\{5 \times 10^{-8}, 5 \times 10^{-7}, 5 \times 10^{-6}, 5 \times 10^{-5}\}$, it is shown that the value of Δ^C declines monotonically as α decreases. This is because, as discussed in Online Appendix 3.1, a lower α reduces learning biases. As a result, it will be more difficult for informed AI speculators to learn collusive trading strategies through homogenized learning biases in the environment with high noise trading risks.

6 Coordinated Choice of Q-Learning Algorithms

In this section, we study the trading profits of informed AI speculators when they adopt Q-learning algorithms with different values of the forgetting rate α. The algorithm with a lower α has smaller learning biases but its training takes longer time and more computation power. We can think of α as capturing the "intelligence level" of the algorithm: the algorithm is more advanced if it has a lower α.

In Subsection 6.1, we conduct simulation experiments in the baseline economic environment using the standard Q-learning algorithms with a fixed α. In Subsection 6.2, we extend the Q-learning algorithm to a two-tier Q-learning algorithm with an adaptive α. This algorithm allows informed AI speculators to learn both the choice of α and the trading strategies associated with each α. We show that informed AI speculators can learn to coordinately choose the values of their α for mutual benefits.

6.1 Heterogeneous Forgetting Rates

Focusing on the baseline economic environment, we allow the two informed AI speculators to adopt Q-learning algorithms with different intelligence levels, as represented by different values of α. We show that when noise trading risks are low (i.e., $\sigma_u = 10^{-1}$), both informed AI speculators' profits are maximized when they choose the same value of α that best matches with the value of β, confirming the result presented in panel A of Figure 7. By contrast, when noise trading risks are high (i.e., $\sigma_u = 10^2$), informed AI speculators face a situation resembling the prisoner's dilemma when they are allowed to choose their algorithms' α.

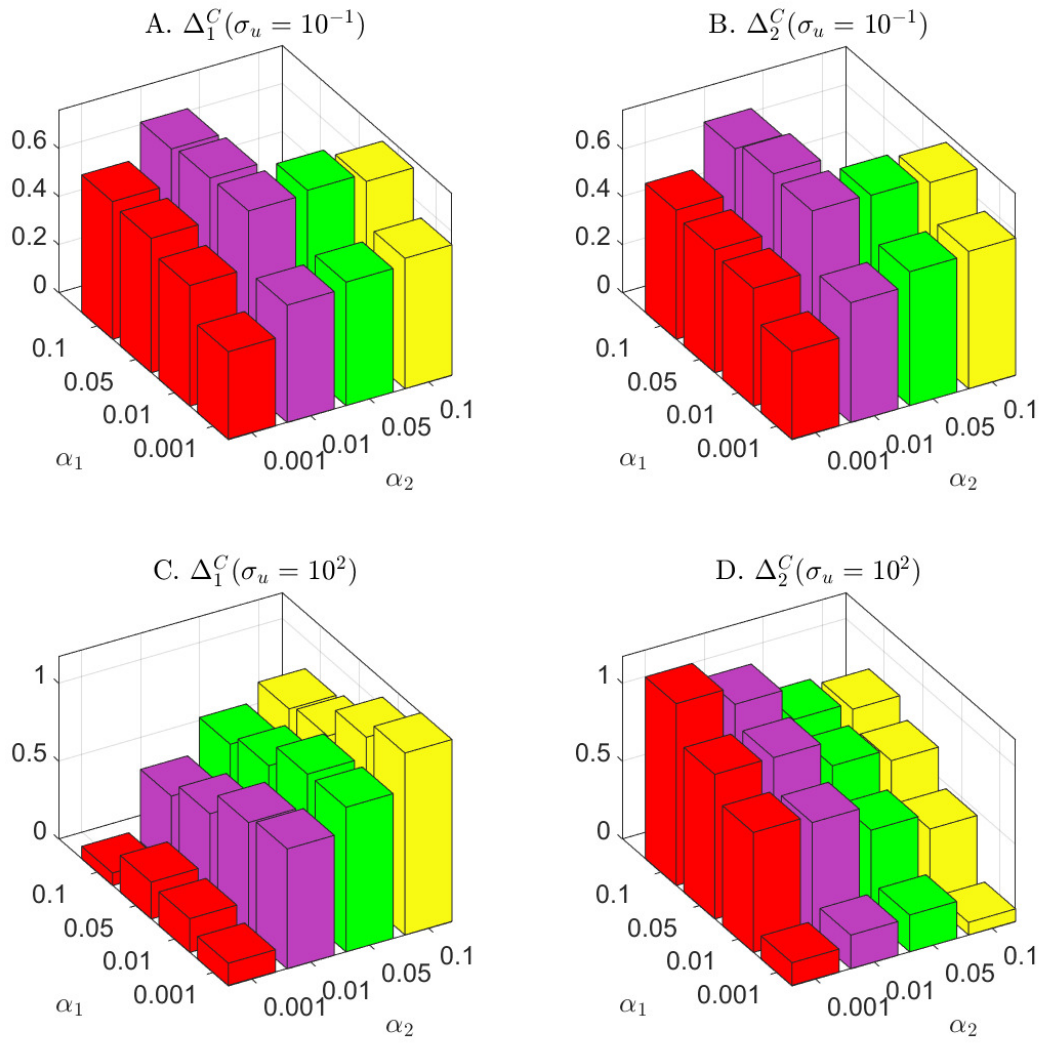
Specifically, each informed AI speculator i adopts an algorithm whose forgetting rate is α_i, with $\alpha_i = 0.001, 0.01, 0.05$ and 0.1 for $i = 1, 2$. Panels A and B of Figure 8, plot the average Δ_1^C and Δ_2^C for informed AI speculators 1 and 2, respectively, in the environment with low noise trading risks

(i.e., $\sigma_u = 10^{-1}$). Both informed AI speculators' profits are maximized when $(\alpha_1, \alpha_2) = (0.01, 0.01)$, which result in $(\Delta_1^C, \Delta_2^C) = (0.743, 0.743)$. Importantly, neither informed AI speculator has the incentive to choose a different α because it would reduce self-profit. The intuition for this result follows the discussions for panel A of Figure 7. Given $\beta = 5 \times 10^{-7}$, setting $\alpha = 0.01$ helps informed AI speculators to learn price-trigger strategies to achieve best collusive outcomes in the environment with low noise trading risks.

By contrast, panels C and D of Figure 8 focus on the environment with high noise trading risks (i.e., $\sigma_u = 10^2$). It is shown that for any combination of (α_1, α_2), the informed AI speculator with a lower α_i attains a higher average Δ_i^C than the other informed AI speculator. Moreover, for $i = 1, 2$ and $j \neq i$, holding informed AI speculator i's α_i unchanged at each value of $\{0.001, 0.01, 0.05, 0.1\}$, as the other informed AI speculator j's α_j decreases, the average Δ_i^C for informed AI speculator i decreases and the average Δ_j^C for informed AI speculator j increases.

These results suggest that if the two informed AI speculators can freely choose their algorithms' α, the situation facing them resembles a prisoner's dilemma in the environment with high noise trading risks. Given the peer's algorithm choice, upgrading the algorithm by setting a lower α can increase its own profit while reduces its peer's profit. Intuitively, because the value of α determines the magnitude of learning biases, the more advanced algorithm has smaller learning biases than the less advanced algorithm. As discussed in Section 5.2, learning biases induce informed AI speculators to adopt more collusive trading strategies with small order flows. Therefore, the informed AI speculator with a less advanced algorithm would adopt a more collusive trading strategy than the one with a more advanced algorithm. This means that the informed AI speculator with a more advanced algorithm tends to choose larger order flows than its peer, enabling it to obtain more profit than its peer. However, if both informed AI speculators adopt advanced algorithms with similarly low values of α, the profits for both of them will be very low (e.g., $(\Delta_1^C, \Delta_2^C) = (0.308, 0.308)$ when $(\alpha_1, \alpha_2) = (0.001, 0.001)$). On the flip side, both informed AI speculators can obtain supra-competitive profits if they both adopt unadvanced algorithms with similarly high values of α (e.g., $(\Delta_1^C, \Delta_2^C) = (0.806, 0.806)$ when $(\alpha_1, \alpha_2) = (0.1, 0.1)$). This confirms the previous results in panel B of Figure 7 that collusion is achieved through homogenized learning biases in the environment with high noise trading risks.

The results we observe bear similarity with the general equilibrium effects in active man-



Note: We allow the two informed AI speculators to adopt Q-learning algorithms with different values of the forgetting rate, denoted by α_1 and α_2 for informed AI speculators 1 and 2, respectively. Panels A and B plot Δ_1^C and Δ_2^C in the baseline economic environment with low noise trading risks (i.e., $\sigma_u = 10^{-1}$). Panels C and D plot those in the environment with high noise trading risks (i.e., $\sigma_u = 10^2$). The other parameters are set according to the baseline economic environment described in Section 4.2.

Figure 8: Profit gains when informed AI speculators use different values of α.

agement characterized by Stambaugh (2020). According to his model, if all managers lack the ability to select positive-alpha stocks, they can collectively achieve high profits. When a small fraction of managers gains more skill, it results in increased profits for the skilled ones, while the less skilled managers experience a decline in their profits. However, if a large proportion of managers becomes more skilled, the profits for all managers start to diminish. This decline

is due to a shrinking alpha magnitude, caused by more substantial price corrections in general equilibrium. Interestingly, the total profit of the active management industry typically decreases whenever any of the managers become more skilled. In a recent work, Dugast and Foucault (2024) derive a similar result by showing that improvements in the skills of active asset managers, due to lower information processing costs or the proliferation of new datasets, can reduce their average performance as asset prices become more informative.

6.2 Adaptive Forgetting Rates

In this subsection, we extend the Q-learning algorithm to a two-tier Q-learning algorithm whereby informed AI speculators learn both the choice of α and the trading strategies associated with each α. Our two-tier Q-learning algorithm essentially allows machines to learn to set an adaptive forgetting rate.

Based on the two-tier Q-learning algorithm, we show that in the baseline economic environment with low noise trading risks (i.e., $\sigma_u = 10^{-1}$), informed AI speculators can easily learn the optimal choice of α, which maximizes their profits. In the environment with high noise trading risks (i.e., $\sigma_u = 10^2$), which is inherently a situation resembling the prisoner's dilemma, informed AI speculators can learn to choose α for mutual benefits, enabling both to obtain supra-competitive profits. Specifically, informed AI speculators learn to choose high values of α, as if they are implicitly coordinating with each other, despite the fact that choosing a low value of α unilaterally may boost self-profit. This result implies that a collusive equilibrium with unadvanced algorithms (i.e., high α) may arise endogenously due to the optimal decisions of informed AI speculators.

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

For any given choice of $\alpha_{i,t}$ in the upper tier, it is necessary to ensure that the lower tier

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

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

$$\widehat{Q}^u_{i,(\tau+1)T}(s^u_{i,\tau T}, \alpha_{i,\tau T}) = (1-\alpha^u)\widehat{Q}^u_{i,\tau T}(s^u_{i,\tau T}, \alpha_{i,\tau T}) + \alpha^u \left[\pi^u_{i,\tau T} + \rho^u \max_{\alpha' \in \mathcal{A}} \widehat{Q}^u_{i,\tau T}(s^u_{i,(\tau+1)T}, \alpha')\right], \quad (6.1)$$

for $\tau = 1, 2, ...$ In equation (6.1), $\pi^u_{i,\tau T}$ 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:

$$\alpha_{i,\tau T} = \begin{cases} \text{argmax}_{\alpha' \in \mathcal{A}}\, \widehat{Q}^h_{i,\tau T}(s^u_{i,\tau T}, \alpha'), & \text{with prob. } 1-\varepsilon^u_\tau, \quad \text{(exploitation)} \\ \widetilde{\alpha} \sim \text{uniform distribution on } \mathcal{A}, & \text{with prob. } \varepsilon^u_\tau. \quad \text{(exploration)} \end{cases} \quad (6.2)$$

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: $\alpha^u = 0.1$, $\beta^u = 10^{-4}$, and $\rho^u = 0.95$. Each training epoch has a total of $T = 10,000,000$ periods. The convergence criterion requires the decisions of

Table 1: Adaptive forgetting rates after the convergence of two-tier Q-learning algorithms.

	$(0.001, 0.001)$	$(0.01, 0.01)$	$(0.1, 0.1)$	Others
Low noise trading risks (i.e., $\sigma_u = 10^{-1}$)	0	957	2	41
High noise trading risks (i.e., $\sigma_u = 10^2$)	0	710	272	18

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

$\alpha_{i,t}$ 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.

Table 1 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 risks (i.e., $\sigma_u = 10^{-1}$), across the $N_{sim} = 1,000$ simulations sessions, 957 sessions converge to the best equilibrium with $(\alpha_1, \alpha_2) = (0.01, 0.01)$, which maximizes both informed AI speculators' profits, as shown in panels A and B of Figure 8. This suggests 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 risks (i.e., $\sigma_u = 10^2$). As shown in panels C and D of Figure 8, 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 by coordinately reaching the equilibrium with $(\alpha_1, \alpha_2) = (0.01, 0.01)$ or $(\alpha_1, \alpha_2) = (0.1, 0.1)$, that is, by adopting unadvanced algorithms to trade. In theory, these two equilibria with high values of α can only be sustained in a repeated game. In our simulation experiments, we find that across the $N_{sim} = 1,000$ simulations sessions, 272 sessions converge to the equilibrium with $(\alpha_1, \alpha_2) = (0.1, 0.1)$, and 710 sessions converge to the equilibrium with $(\alpha_1, \alpha_2) = (0.01, 0.01)$. There does not exist a single simulation session that converges to the equilibrium with $(\alpha_1, \alpha_2) = (0.001, 0.001)$, even though this is the unique Nash equilibrium in a one-shot game.[23] Our results indicate that in the environment

[23]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 $\alpha = 0.001$.

with high noise trading risks, 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

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

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

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

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

Battigalli, Pierpaolo, Simone Cerreia-Vioglio, Fabio Maccheroni, and Massimo Marinacci. 2015. "Self-Confirming Equilibrium and Model Uncertainty." *American Economic Review*, 105(2): 646–77.

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

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

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

Chen, Shuaiyu, Lin Peng, and Dexin Zhou. 2024. "Wisdom or Whims? Decoding Investor Trading Strategies with Large Language Models." Zicklin School of Business, Baruch College Working Papers.

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

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

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

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

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

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

Fershtman, Chaim, and Ariel Pakes. 2012. "DYNAMIC GAMES WITH ASYMMETRIC INFORMATION: A FRAMEWORK FOR EMPIRICAL WORK." *The Quarterly Journal of Economics*, 127(4): 1611–1661.

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

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

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

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

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

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

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

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

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

Grossman, Sanford J., and Joseph E. Stiglitz. 1980. "On the Impossibility of Informationally Efficient Markets." *The American Economic Review*, 70(3): 393–408.

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

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

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

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

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

Koijen, Ralph S. J., and Motohiro Yogo. 2019. "A Demand System Approach to Asset Pricing." *Journal of Political Economy*, 127(4): 1475–1515.

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

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

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

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

Lo, Andrew W., and A. Craig MacKinlay. 1999. *A Non-Random Walk Down Wall Street.* Princeton University Press.

Lo, Andrew W., Harry Mamaysky, and Jiang Wang. 2000. "Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation." *The Journal of Finance*, 55(4): 1705–1765.

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

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

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

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

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

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

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

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

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

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

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

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

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