Do AI-Algorithmic Traders Lead to Market Instability? A Multi-Agent Reinforcement Learning Approach*

Yang Fan† Markus Pelger‡ Xintong Yu§

This draft: August 2026
First draft: October 2024

Abstract

AI-powered algorithmic trading can destabilize financial markets through a mechanism fundamentally different from algorithmic collusion. In a multi-agent deep reinforcement learning framework, informed AI traders endogenously learn manipulative strategies, pump-and-dump schemes and strategic under-correction of mispricing, that exploit the learning of less informed AI traders and generate persistent local price bubbles. Manipulation requires no coordination: a single informed trader suffices, and its victims are fully adaptive algorithms rather than mechanical rules. Instability is most severe when information asymmetry is large and few uninformed traders hold substantial market power. The strategies emerge under deep reinforcement learning but not under tabular Q-learning.

Keywords: algorithmic trading, reinforcement learning, multi-agent systems, market microstructure, price bubbles, market manipulation, artificial intelligence, market stability.
JEL Classification: D43, C73, G10, G14, L13

*We thank seminar and conference participants at Stanford University and at the CUHK FinTech Conference 2026, INFORMS Annual Meeting 2024, Annual Meeting 2025 of the European Financial Management Association (EFMA), Annual Meeting 2025 of the Society for Industrial and Applied Mathematics (SIAM) on Financial Mathematics and Engineering, Journal of Investment Management (JOIM) 2025 Conference on "AI in Finance," and the Stanford ICME Research Symposium for helpful comments. We thank the International Monetary Fund for including our research results in their chapter on Advances in Artificial Intelligence: Implications for Capital Market Activities in International Monetary Fund (2024). We thank Qube Research & Technologies for generous research funding and helpful comments.

†Stanford University, Department of Management Science & Engineering, Email: ynf@stanford.edu.
‡Stanford University, Department of Management Science & Engineering, Email: mpelger@stanford.edu.
§Columbia University, Columbia Business School, Email: xyu29@gsb.columbia.edu.

1 Introduction

The rapid adoption of artificial intelligence in financial markets is fundamentally reshaping how trading decisions are made. Algorithmic trading already accounts for the majority of U.S. equity trading volume, with common estimates in the range of 60–75% (U.S. Securities and Exchange Commission, 2020), and market participants are increasingly deploying AI-based systems that go far beyond traditional rule-based strategies.[1] While traditional algorithmic trading relies on static rules defined by humans, a new generation of AI-powered trading algorithms autonomously optimizes strategies through self-learning and trial-and-error interaction with the market. Deep reinforcement learning (deep RL; Sutton and Barto, 2018), the technology behind breakthroughs such as AlphaGo (Silver et al., 2016), raises the possibility that trading algorithms develop strategies whose complexity exceeds human cognitive capacity. This technological shift poses a fundamental question for financial economists and regulators: *What happens when AI algorithms trade against each other in financial markets?*

In this paper, we provide evidence that the interaction of *heterogeneous* AI traders can destabilize markets through a mechanism that is fundamentally different from the algorithmic collusion documented in the existing literature. Using a multi-agent deep reinforcement learning framework, we show that informed AI traders endogenously develop manipulative strategies that exploit the learned inference of less informed AI traders, and that shape what those traders learn in the first place. These strategies, "pump-and-dump" schemes and strategic under-correction of price disparities, are not programmed. They emerge from the learning process itself. The resulting price dynamics feature *local price bubbles*, persistent deviations of market prices from fundamental values that constitute a new source of instability in AI-driven markets. Throughout the paper, instability means these persistent price distortions, the overshooting that creates them, and the redistribution of trading gains they cause.

The distinction between our findings and the literature on algorithmic collusion is central to our contribution. Algorithmic collusion, as studied by Calvano et al. (2020) in product markets and Dou et al. (2025) in financial markets, arises when *homogeneous* algorithms tacitly coordinate to suppress competition, earning supra-competitive profits through mutual restraint. The mechanism we document is one of *manipulation*. A single informed AI trader, acting purely in its own interest and without any coordination, learns to exploit its information advantage to systematically distort the trading behavior of uninformed AI traders. Collusion harms markets through conservative trading and reduced price discovery. Manipulation harms markets through actively distorted prices and locally inflated bubbles. The two phenomena call for different regulatory responses, and the manipulation channel requires neither multiple informed traders nor any form of tacit agreement among them.

Our analysis builds on a market environment that captures three essential elements of real-world

[1] For example, the Securities and Exchange Commission (SEC) approved Nasdaq's reinforcement-learning-based, AI-driven order type; major digital platforms have begun deploying RL trading bots; and leading hedge funds increasingly adopt AI for portfolio management and execution.

equity trading: an asset with both a market price and a fundamental value, periodic information-revelation events analogous to earnings announcements, and a reduced-form price-formation mechanism in which trades have linear price impact, consistent with the linear equilibrium pricing rule of Kyle (1985). Within this environment, informed AI traders who observe the true fundamental value interact with uninformed AI traders who must infer it from noisy private signals and market data. Each AI trader is an independent deep neural network policy trained with a state-of-the-art deep reinforcement learning algorithm. Crucially, the agents are heterogeneous along two economic dimensions, the precision of their information and their market power (trading capacity), and each agent optimizes only its own trading profit. The framework functions as a laboratory that is unavailable in empirical market data. The researcher observes the fundamental value, every trader's information set, and every trader's complete strategy, so manipulation can be identified directly from behavior rather than inferred from prices, and the market structure can be varied experimentally while everything else is held fixed.

Our main findings are fourfold. First, we document the endogenous emergence of manipulation and local price bubbles when informed and uninformed AI traders interact. The informed AI trader learns two distinct strategies that sacrifice short-run profit to distort the beliefs and actions of other algorithms. In the "pump-and-dump" strategy, the informed trader initially corrects an undervalued price, but then continues to buy beyond the fundamental value, deliberately creating upward momentum. The uninformed AI traders, unable to distinguish informed trades from noise, interpret the price increase as information and amplify the bubble, at which point the informed trader profitably reverses its position. In the "strategic under-correction" strategy, the informed trader deliberately refrains from fully correcting a price disparity, preserving a persistent mispricing that it can exploit repeatedly. Both strategies violate the behavior implied by competitive, non-strategic informed trading, under which every informed trade should push prices toward fundamentals.

Second, we characterize the market conditions under which manipulation and price inefficiency are most severe. Instability is greatest when the information asymmetry between traders is large, when the number of uninformed market participants is small (two uninformed traders is the most fragile configuration), and when the division of market power between the informed and uninformed sides is unbalanced. The market-power result is subtle. Price inefficiency is U-shaped in the informed trader's share of trading capacity, at every number of uninformed traders, and the two arms of the U correspond to the two manipulative strategies. When the informed trader is small relative to the uninformed side, it trades *against* the fundamental to ignite momentum, because manipulation operates as a lever that requires the buying power of the manipulated. When the informed trader dominates the market, every one of its trades points *toward* the fundamental, yet it deliberately stops short of correcting the price, preserving the mispricing it profits from. Efficiency is greatest at intermediate divisions of market power. Conversely, markets with homogeneous information and many competing traders are efficient. Adding a second informed trader separates the two margins that entry bundles together, the number of informed competitors and the informed side's total trading capacity: with that total capacity held fixed, a second informed trader splits

the manipulation rents between the rivals but leaves prices and the uninformed traders' profits unchanged, and price distortions fall only when the informed side's capacity expands, an expansion that pushes the uninformed traders to breaking even. Prices depend on how much informed capital the market contains, not on how many competitors hold it, and neither margin restores the uninformed traders' profits.

Third, we open the black box of the learned policies. Using SHAP (SHapley Additive exPlanations) values, the standard attribution method in interpretable machine learning, we show that the informed agent's strategy adapts fundamentally to its competitive environment. Trading alone, it relies on the gap between price and fundamental value. Facing uninformed AI traders, it shifts toward market-based signals such as detrended prices and volume, consistent with trading on the *reactions* of other algorithms rather than on fundamental information. Uninformed agents learn to discount their noisy fundamental signals and rely on price trends and volume, individually rational behavior that creates precisely the vulnerability the informed agent exploits.

Fourth, we show that our findings could not be obtained with the simpler AI technologies studied in the prior literature, and we develop the methodology required to study "smart" AI traders. The manipulative strategies we document are multi-step plans that require enduring short-term losses for delayed gains in a high-dimensional continuous state space. In a deliberately simple illustrative game, in which a single learning agent trades against rule-based block traders whose gradual order execution creates predictable local price trends, we show that tabular Q-learning fails to learn even this most basic dynamic trading strategy, while deep RL succeeds. Studying deep RL agents, however, raises two methodological challenges that we address and view as contributions in their own right. The first is reward design. Terminal trading profits provide too sparse a signal to train deep RL traders, while naïve dense rewards based on unrealized gains distort incentives. We construct a shaped reward and prove that it is exactly equivalent to the terminal trading profit while respecting each agent's information set. The second is equilibrium selection. In non-cooperative multi-agent learning there is no single objective to validate hyperparameters against. We develop a tuning and selection procedure based on the NashConv metric from computational game theory, which retains only configurations whose learned policies constitute an approximate Nash equilibrium, ensuring that our results reflect equilibrium economics rather than artifacts of particular training runs.

Related Literature. Our paper contributes to the rapidly growing literature on AI trading in financial markets, whose frontier is defined by a sequence of papers by Dou, Goldstein, and coauthors. We discuss this work in detail because the distinctions delineate our contribution.

Dou et al. (2025) study trading in a repeated Kyle-type market and show through simulation experiments that *homogeneous, equally informed* Q-learning speculators autonomously sustain collusive, supra-competitive profits without communication. They identify two distinct mechanisms. The first is *price-trigger collusion* ("artificial intelligence"): algorithms sustain conservative trading through the threat of reverting to competitive behavior, akin to the punishment strategies of

Green and Porter (1984) and Abreu et al. (1986). The second is *over-pruning collusion* ("artificial stupidity"): an asymmetric learning bias in Q-learning that prematurely prunes aggressive actions after adverse noise realizations, sustaining collusive outcomes without any strategic mechanism. The two mechanisms operate in different parameter regions: price triggers require effective price monitoring (low noise-trading risk and strong price-sensitive investor demand), while over-pruning operates everywhere. The underlying equilibrium concept is the experience-based equilibrium of Fershtman and Pakes (2012). In both cases, collusion manifests as *conservative* trading: reduced order flow, lower price informativeness, and reduced liquidity.

Dou et al. (2026a) extend this agenda to AI *planning*. They equip oligopolistic informed Q-learners with backward-induction planning modules ("coupled Q-learning") and show that the speculators can learn to intertemporally coordinate: they jointly create price run-ups that activate *positive-feedback investors*, a continuum of traders with an exogenously parameterized demand rule that chases past price increases (De Long et al., 1990), and then jointly unwind, amplifying bubbles and crashes. Their central theoretical contribution is to identify two obstacles to such coordinated bubble riding, namely time inconsistency (each speculator wants to exit the bubble first) and weak dynamic punishment (deviations that prevent bubbles leave little collusive surplus to destroy). They show that the algorithms overcome them through "compounded price-trigger strategies." Importantly, their instability result is conditional: coordinated bubble creation arises only when positive-feedback trading is sufficiently strong, and the entire mechanism still rests on collusion among symmetric informed speculators. Relatedly, Dou et al. (2026b) study uninformed, price-taking AI traders competing against informed humans in a competitive market and prove an *algorithmic herding* result: independently learning AI investors converge to a common, rational-expectations demand rule and, lacking private information and market power, cannot manipulate. Colliard et al. (2026) study Q-learning market makers and show that they learn supra-competitive spreads, and Goldstein et al. (2025) survey the theory of financial market fragility, in which prices overreact to shocks because investors learn from prices. Our mechanism is an algorithmic instantiation of exactly this feedback, with the twist that the learning it exploits is itself created by a strategic trader.

Our paper differs from this body of work in the economic mechanism, the agents studied, and the methodology, and we view the contributions as complementary. First, we study *heterogeneous* AI traders that differ in information quality and market power, whereas the collusion results in Dou et al. (2025) and Dou et al. (2026a) concern symmetric informed speculators. Heterogeneity is not a technical detail. It is the source of our mechanism, which operates through the exploitation of an information gradient rather than through mutual restraint among equals. Second, manipulation in our framework requires *no coordination whatsoever*: a single informed AI trader generates bubbles, so the time-inconsistency and weak-punishment obstacles that are central to coordinated bubble riding in Dou et al. (2026a) simply do not arise. Third, and we believe most importantly, the agents being manipulated in our framework are themselves *deep RL algorithms that adapt fully during learning*, not an exogenous behavioral demand rule. The question of whether an AI trader

can sustainably manipulate other AI traders *that can learn back* is qualitatively different from the question of whether colluding algorithms can exploit a fixed positive-feedback rule, and our answer is affirmative: trend-chasing by the uninformed agents arises endogenously as their individually optimal response to noisy information, which is precisely what makes it robust and exploitable in equilibrium. Fourth, our agents are deep RL systems with continuous, 23-dimensional state spaces, neural network function approximation, and no built-in planning module: the multi-step manipulative strategies emerge from model-free learning alone. This addresses directly the open question raised in Dou et al. (2025) and Dou et al. (2026a), both of which deliberately use minimal, hand-selected discrete state spaces and note function approximation with deep RL as future work. In sum, the literature now documents four channels through which AI trading can harm market quality. Price-trigger collusion and over-pruning collusion (Dou et al., 2025) require at least two equally informed algorithms whose mutual restraint reduces trading and price informativeness. Coordinated bubble riding (Dou et al., 2026a) requires at least two planning-capable informed algorithms together with an exogenous positive-feedback demand rule that serves as the victim. Information manipulation, the channel we document, requires only a single informed agent, no coordination, and no behavioral rule. Its victims are adaptive learning algorithms, and its footprint is local price bubbles and persistent mispricing.

Our mechanism also has deep roots in the classical theory of market manipulation, and the relation sharpens what is new here. Allen and Gale (1992) show that trade-based manipulation can be profitable when other traders cannot tell whether a large trader is informed, Jarrow (1992) characterizes when large traders can profitably corner or squeeze markets, Brunnermeier and Pedersen (2005) study predatory trading against distressed counterparties, Goldstein and Guembel (2008) show that manipulation can arise because real decisions feed back from prices, and Kyle and Viswanathan (2008) discuss how to define illegal price manipulation, and Aggarwal and Wu (2006) document manipulation cases empirically. In all of these models, the behavior that makes manipulation profitable is assumed as part of the environment, for example beliefs that pool informed and uninformed types, a distressed trader who must liquidate, or a decision maker who reacts to prices. In our setting nothing of the sort is assumed. The victims are learning algorithms that could in principle learn any behavior, including not trading at all, and the exploitable behavior, trend-following inference from prices, emerges endogenously because it is each victim's best available use of its noisy information. The manipulability of the market is therefore an equilibrium outcome rather than an assumption, and our comparative statics ask a question the classical literature could not, namely how manipulation depends on the composition, information technology, and entry of the algorithms themselves.

Our paper also relates to the literature on algorithmic collusion in product markets. Calvano et al. (2020) demonstrate that Q-learning algorithms learn supra-competitive prices in a workhorse oligopoly model. Klein (2021) extends this to sequential pricing, Waltman and Kaymak (2008) study Cournot competition, and Assad et al. (2024) provide empirical evidence from German retail gasoline markets. Brown and MacKay (2023) and Asker et al. (2022) study the design of pric-

ing algorithms. Financial markets differ fundamentally from product markets in the presence of asymmetric information, endogenous price formation, and the ability to trade in both directions. These features change the nature of strategic interaction among algorithms and give rise to the manipulation channel we document.

Our work further connects to the classical literature on price bubbles. Theoretical models generate bubbles from heterogeneous beliefs (Harrison and Kreps, 1978; Blanchard, 1979; Camerer, 1989; Abreu and Brunnermeier, 2003; Jarrow, 2015); agent-based models generate them from the interaction of parametric or adaptively selected trading rules, a tradition that begins with the Santa Fe artificial stock market, where ecologies of evolving forecasting rules produce bubbles, excess volatility, and volatility clustering (Arthur et al., 1997; LeBaron et al., 1999; LeBaron, 2006; Harras and Sornette, 2011; Westphal and Sornette, 2020); and mean-field game models study optimal bubble riding among a continuum of small agents (Tangpi and Wang, 2022, 2024). Relative to the agent-based tradition, our contribution is threefold. The agents are deep reinforcement learners rather than rule selectors, so multi-step strategies can emerge that no rule in a predefined ecology encodes. They are large and strategic, internalizing their own price impact rather than acting as price takers. And reported outcomes are disciplined by an explicit equilibrium criterion, so the bubble is an equilibrium phenomenon rather than a transient of the learning dynamics. Finally, we build on the computer science literature on multi-agent RL, which has focused primarily on cooperative settings (Tampuu et al., 2017). The emergent-behavior philosophy of Johanson et al. (2022), in which complex economic behavior arises from simple rewards and expressive learners, motivates our simulation design. Our setting is fully competitive, with each agent maximizing its own trading profit.

Organization. Section 2 introduces the conceptual market environment and the economic mechanism. Section 3 describes the design of the AI trading agents and our two methodological contributions, reward shaping and NashConv-based equilibrium selection, and presents an illustrative single-agent game that demonstrates the limits of Q-learning before the main analysis. Section 4 presents the experimental design and main results. Section 5 reports robustness checks. Section 6 interprets the learned policies. Section 7 concludes with policy implications. Technical details on the learning algorithm, the equilibrium selection procedure, and additional results are collected in the Appendix.

2 A Market with Heterogeneous AI Traders

We develop a simulation framework to study the market equilibrium among AI speculators that differ in their information and market power. The environment is designed to capture the essential features of real-world equity trading (asymmetric information, price impact, and periodic revelation of fundamentals) while remaining tractable for multi-agent reinforcement learning. This section introduces the conceptual setup; the specific parameter values used in our experiments are

collected in the base configuration of Section 4.1.

2.1 Market Structure

The market trades a single asset with unobserved fundamental value $\nu \in \mathbb{R}^+$ and market price p_t. Time is discrete, and trading is organized in *games* of T steps each. At the end of every game, the fundamental value is publicly revealed and all positions are liquidated at ν, the analogue of a public information event, such as an earnings announcement, that resolves uncertainty about fundamentals. Between revelation events, the fundamental value is fixed but unobserved by part of the market, and the market price evolves endogenously with order flow. Across games, the fundamental value evolves as a random walk, so that each game begins with a randomized gap between the market price and the new fundamental value. This initial mispricing is the source of inefficiency that gives speculators a reason to trade. Each *episode*, the unit of training and evaluation for the learning algorithms, consists of a sequence of consecutive games.

The forced liquidation at the revealed fundamental value serves two purposes. Economically, it anchors the value of a position to fundamentals, exactly as in Kyle (1985), where the informed trader's profit is realized at the liquidation value. Computationally, it creates a meaningful terminal reward that ties trading profits to the relationship between prices and fundamentals.

The framework is deliberately flexible about the trading frequency it represents. At high frequency, a revelation event is a scheduled news release such as an earnings announcement, the trading steps are the seconds or minutes between such releases, and a game captures the fast repricing episode around public news. At lower frequency, the revelation event is any moment at which fundamental uncertainty is resolved, the steps are days, and a game captures a longer trading episode between information events. Nothing in the mechanism depends on the calendar interpretation, because the economics operates through the ratio of information asymmetry to market depth rather than through clock time. A further conceptual implication is worth stating up front. Each game is a zero-sum competition between the speculators as a group and the market-making sector. Speculator profits net out exactly against the market maker's trading losses, and the aggregate profit that speculators can extract is bounded by the initial mispricing, so any profit one speculator earns beyond that bound must come from the losses of other speculators.

2.2 Market Participants

The market is populated by three types of participants.

Informed AI Traders. Informed deep RL agents observe the true fundamental value ν and submit market orders x_t at each time step. They represent sophisticated institutional investors or proprietary trading firms with superior fundamental information.

Uninformed AI Traders. Uninformed deep RL agents observe only a noisy private signal of the fundamental value,

$$\nu_{\text{obs}} = \nu + \epsilon, \qquad \epsilon \sim \text{Normal}(0, \phi^2), \tag{1}$$

together with the public market history of prices and aggregate trading volume. The signal is drawn independently at each revelation event, so every game begins with a fresh, imprecise estimate of the new fundamental value. Two clarifications define the information structure precisely. First, "uninformed" is a label of convenience. What matters is only that some AI traders have a *better* signal about the fundamental value than others, so the reader may substitute "less informed" and "more informed" throughout. A better signal stands in for better data, better analytics, or better forecasting algorithms, which is exactly the dimension along which trading firms differ in real markets. Second, apart from signal quality and trading capacity, all traders are identical. They observe the same market data, face the same price impact, use the same learning algorithm and network architecture, and receive rewards of the same form. Importantly, no trader is told its own type. An agent only receives its signal and its trading results, and it must learn from interacting with the market how much its signal can be trusted. We parameterize the *information gap* between informed and uninformed traders by the ratio of the signal-noise standard deviation ϕ to the average absolute fundamental innovation across games. A 100% information gap means the noise standard deviation equals the mean absolute change in the fundamental value between games, so the private signal is roughly as imprecise as the prior, while a 0% gap means all traders are equally informed.

Market Makers. A market-making sector absorbs the net order flow and sets prices according to the linear price-impact rule

$$p_t = p_{t-1} + \lambda z_t, \qquad z_t = \sum_i x_{i,t}, \tag{2}$$

where z_t is the aggregate net order flow and $\lambda > 0$ is the price-impact parameter, with $1/\lambda$ measuring market depth. All orders submitted at step t execute at the updated price p_t of equation (2): a buyer pays the price that its own order, together with the contemporaneous orders of all other traders, has already moved. The market maker takes the opposite side of the aggregate flow at the same price and carries the resulting inventory until the revelation event, at which point all positions, including the market maker's, settle at the revealed fundamental value.

Three remarks on the price-formation rule are in order. First, equation (2) is a *reduced-form* pricing rule. Its linear form is consistent with the linear equilibrium of Kyle (1985), in which a competitive market maker facing informed and noise order flow optimally sets prices linear in net order flow with constant depth. We deliberately do not claim that our market maker implements the Kyle equilibrium. The Kyle pricing rule is derived under rational expectations with optimizing insiders and Gaussian order flow, and the corresponding equilibrium pricing rule when order flow is generated by interacting AI algorithms is unknown. Characterizing markets with interacting

AI algorithms is what our simulation delivers. Fixing λ lets us focus on the novel component of the framework, the strategic interaction among learning agents, and we verify robustness to the value of λ in Section 5. Second, price impact endogenously penalizes frivolous trading, because a round-trip trade buys at rising prices and sells at falling prices, incurring a cost proportional to λ. The pricing rule therefore embeds a realistic transaction-cost structure without simulating a full limit order book. Third, agents observe only aggregate prices and volumes and cannot identify the source of demand or supply, which mirrors the anonymity of real-world order flow.

Where Are the Noise Traders? Classical microstructure models add exogenous noise traders because without them informed trading would be impossible. A rational uninformed counterparty who knows it faces only better-informed traders refuses to trade, and the market unravels (Grossman and Stiglitz, 1980). Our setup does not need this device. The role that exogenous noise plays in Kyle (1985) is played here endogenously by the less informed AI traders themselves. Their signals are imperfect, they do not know how good their own information is, and their learned trend-following generates order flow that the informed agent cannot be fully separated from. In this sense noise trading is implicit in the model, but with an important difference. Our "noise" comes from optimizing agents who can adapt, which is exactly what makes exploiting them a nontrivial equilibrium phenomenon rather than mechanical harvesting of a fixed rule. We also verify the robustness of our results to explicit noise traders. Section 5 adds an agent that trades randomly with fixed capacity and shows that our results do not change qualitatively. We therefore focus on the more transparent case without noise traders first, where every unit of order flow traces back to an optimizing agent.

2.3 How Manipulation Can Emerge

Before turning to the design of the AI agents, we describe the economic mechanism through which manipulation can arise in this environment. The mechanism operates through three channels that build on each other.

Channel 1: Information asymmetry creates the opportunity. The informed agent knows ν; uninformed agents observe only $\nu + \epsilon$. Absent strategic considerations, the informed agent would trade toward the fundamental, buying when $p < \nu$ and selling when $p > \nu$, collecting the mispricing as profit and making prices more efficient in the process. This is the textbook stabilizing role of informed speculation. The key insight is that the informed agent can do *better* than price correction if it can influence what the uninformed agents believe.

Channel 2: Uninformed agents rationally rely on market signals. Because their private signal is noisy, uninformed agents optimally learn to extract information from prices and volume, the aggregate outcomes that embed other traders' information. This is the classic logic of learning from prices (Grossman and Stiglitz, 1980), rediscovered here by algorithms rather than assumed.

Relying on market-based signals is individually rational and, in a market without strategic informed traders, would be harmless. But it creates a vulnerability. The very signals the uninformed agents trust can be *generated* by the informed agent at will, because the informed agent's trades move prices through equation (2).

Channel 3: The informed agent learns to exploit the inference of others. Through trial and error over many training episodes, the informed deep RL agent discovers that pushing the price *beyond* the fundamental value induces uninformed agents to amplify the move, and that reversing its position into this induced demand is more profitable than simple price correction. This is a genuinely dynamic strategy. The informed agent accepts losses in the pumping phase (buying above fundamental value) in exchange for larger gains in the dumping phase. Discovering it requires learning multi-step strategies whose payoff arrives many steps after the initial cost, which is the defining capability that separates deep RL from simpler learning algorithms. Two phases of the mechanism should be kept apart. During training, the informed agent's trades shape what the uninformed agents learn to do. During evaluation, when all policies are frozen, the informed agent exploits the reaction patterns the uninformed agents have already learned.

The result is a local price bubble, a temporary but persistent deviation of the market price from the fundamental value that arises endogenously from the strategic interaction of adaptive algorithms, without exogenous shocks, behavioral assumptions, or coordination.

2.4 A Competitive Benchmark

To interpret the simulation results, it is useful to state how an informed trader *would* behave absent strategic manipulation incentives. In a competitive benchmark in which the informed trader takes the behavior of others as fixed and non-reactive, as in the linear equilibrium of Kyle (1985) where the insider trades against noise traders who cannot be influenced, informed trading is purely *corrective*. The optimal order is increasing in the gap between fundamental value and price, every informed trade pushes the price toward the fundamental, and informed profits are bounded by the initial mispricing. Strategic informed traders who internalize their own price impact trade more patiently, but still in the corrective direction (Kyle, 1989). This benchmark yields three qualitative predictions for our setting:

(i) *Price correction:* the informed trader's trades should be directed toward the fundamental value;

(ii) *Price efficiency:* market prices should converge toward the fundamental value within each game, with the terminal price gap approaching zero;

(iii) *Bounded profits:* informed profits should not systematically exceed the value of the initial mispricing.

The price-formation rule (2) makes these predictions exact. The following result, proved in Appendix B, characterizes who can earn what in this market, for *arbitrary* trading strategies.

Proposition 1 (Aggregate accounting and the limits of collective extraction)**.** *Consider a single game with T trading steps, initial price p_0, fixed fundamental value ν revealed after step T, price evolution (2) with execution at the post-impact price, and no exogenous noise-trader flow. Let z_t denote the aggregate net order flow of the speculators at step t and $Z_t = \sum_{s \le t} z_s$ the cumulative flow. Let Π denote the sum of all speculators' realized trading profits at liquidation and Π^{MM} the market maker's realized profit. Then, path by path:*

(i) (Zero-sum) $\Pi + \Pi^{MM} = 0$.

(ii) (Decomposition) *Aggregate speculator profit decomposes into a price-discovery gain, a repositioning cost, and a churn cost:*

$$\Pi = \underbrace{(\nu - p_0)\, Z_T}_{\text{price discovery}} - \underbrace{\tfrac{\lambda}{2} Z_T^2}_{\text{repositioning cost}} - \underbrace{\tfrac{\lambda}{2} \sum_{t=1}^{T} z_t^2}_{\text{churn cost}}. \tag{3}$$

(iii) (Sharp bound) *Aggregate speculator profit is bounded:*

$$\Pi \;\le\; \frac{T}{T+1} \cdot \frac{(\nu - p_0)^2}{2\lambda}, \tag{4}$$

with equality if and only if the aggregate flow is constant, $z_t = \frac{\nu - p_0}{\lambda(T+1)}$ for all t. Under this profit-maximizing aggregate flow, the terminal mispricing is $\nu - p_T = \frac{\nu - p_0}{T+1}$, which vanishes as the horizon grows, so maximal collective extraction implements asymptotically full price correction.

Corollary 1 (Manipulation is redistribution)**.** *Any individual speculator's profit in excess of the bound (4) is financed one for one by the trading losses of other speculators. Moreover, holding the terminal net position Z_T fixed, aggregate speculator profit is strictly decreasing in the round-trip volume $\sum_t z_t^2$ of the aggregate flow. Every unit of additional churn transfers $\lambda/2$ from the speculator sector to the market-making sector.*

Proposition 1 has three implications that organize the empirical analysis. First, it sharpens prediction (iii). Collective speculator profits per game can never exceed $\frac{T}{T+1}(\nu - p_0)^2/(2\lambda)$, a quantity determined entirely by the initial mispricing and market depth. Any informed profit above this bound is direct evidence of exploitation of other speculators, not of superior price discovery, and we use this implication as a diagnostic on our simulation output: no game in any reported ensemble violates the bound (Table 3). Second, part (iii) shows that full price correction is not merely a competitive outcome but the *collectively optimal* one. A benevolent planner maximizing total speculator profit would trade the price to the fundamental value in equal, patient steps, which is exactly the gradual execution familiar from the optimal-execution literature (Bertsimas and Lo, 1998; Almgren and Chriss, 2001). Persistent price gaps in equilibrium therefore measure how far strategic competition pushes the market away from its own collective optimum. Third, the zero-sum identity disciplines all statements about gains and losses. Nothing is created or destroyed

among market participants, so "instability" manifests as distorted prices, excess volume, and a redistribution of trading gains, with the market maker mechanically absorbing the net side. With exogenous noise traders (used only in robustness checks), the same bound applies to speculators and noise traders combined, so the learning agents can profit beyond the bound only at the noise traders' expense.

Our results show that deep RL agents systematically violate all three benchmark predictions when facing uninformed AI traders. A substantial fraction of informed trades push prices *away* from fundamentals (Section 4.5), terminal price gaps persist (Sections 4.3 and 4.4), and the informed agent's profit exceeds what price correction alone could generate (Section 4.6). We refer to the informed agent's excess profit relative to the corrective benchmark as the *manipulation premium*. By Corollary 1, this premium is financed by the uninformed agents' trading losses and reduced gains, not by the market-making sector beyond the fixed extraction bound. The deviations from the benchmark arise precisely because the assumption underlying it fails: the other market participants are not fixed noise. They are adaptive algorithms whose behavior responds to, and can be shaped by, the informed agent's trades.

3 AI Trading Agents

Each AI trader is an independent learning agent. Formally, the market is a partially observed stochastic game, and each agent faces its own partially observed and, during joint training, non-stationary decision problem. At each step it observes a state, takes an action, and receives a reward, and it updates a neural network policy, its trading strategy, mapping observations to trading decisions. All agents learn simultaneously in the same market, and each agent optimizes its own trading policy based only on its own rewards. No agent can see another agent's strategy, parameters, or rewards directly. Agents affect each other only through prices and volumes, so each agent must infer its competitors' behavior from the market outcomes their trading produces. This reflects real-world competition, in which every firm develops its trading algorithm in-house while adapting to rivals it observes only through the market. This section describes the agents' observations, actions, and rewards, and the two methodological problems, reward design and equilibrium selection, whose solutions make the study of competing deep RL traders feasible.

3.1 Observations

At each time step, every agent observes a vector of 23 continuous variables in three groups. These are (i) *market signals* (18 variables): deviations of prices and volumes from moving averages at six horizons, and a decomposition of the price series into trend and noise components via the Hodrick-Prescott filter at three smoothing values, which jointly capture momentum and trend information at different frequencies; (ii) *fundamental information* (2 variables): the deviation of the current price from the agent's observed fundamental value ν_{obs}, which equals the true value for informed agents and the noisy signal for uninformed agents, together with the fraction of time remaining until

Table 1. Observation Vector at Each Time Step

Feature Type	Description	Formula	Count
MA Deviations	Price deviation from moving averages	$p_{t-1} - \text{mean}(p_{t-m:t-1})$	6
	Volume deviation from moving averages	$(V_{t-1} - \text{mean}(V_{t-m:t-1}))/1000$	6
	for horizons $m \in \{2, 3, 5, 10, 20, 35\}$		
HP Filter	Change in trend component	$H_t^{(\lambda_{HP})} - H_{t-1}^{(\lambda_{HP})}$	3
	Noise component	$N_t^{(\lambda_{HP})} = p_{t-1} - H_t^{(\lambda_{HP})}$	3
	for smoothing $\lambda_{HP} \in \{10^3, 10^5, 10^7\}$		
Time Encoding	Fraction of time remaining	$1 - t/T$	1
Fundamental Signal	Normalized price-to-signal deviation	$100 \cdot (p_{t-1}/\nu_{\text{obs}} - 1)$	1
Account Info	Normalized position	$L_t/\bar{L}$	1
	Normalized cash level	$c_t/c_0 - 1$	1
	Normalized account value	$B_t/B_0 - 1, \quad B_t = c_t + L_t p_t$	1
	Total continuous observation variables		23

Notes: The moving-average, HP-filter, time, and signal features constitute the 20 market and signal variables, and the three account variables complete the observation. The implementation carries three additional reserved slots that are identically zero and are excluded from the count. Price-based features use the last observed price p_{t-1}, consistent with the execution timing of Section 2.2. For informed agents $\nu_{\text{obs}} = \nu$, and for uninformed agents ν_{obs} is the noisy signal of equation (1). p_t denotes the market price and V_t the gross (unsigned) trading volume at step t, and $\text{mean}(p_{t-m:t-1})$ is the moving average over the last m steps. All moving-average and HP-filter features are computed from trailing data only. At each step the filter is applied to the price history up to the last observed price, so no feature uses information from future steps, and an automated no-look-ahead test in the code repository verifies this property. $H_t^{(\lambda_{HP})}$ is the trend component of the price series extracted by the Hodrick-Prescott filter with smoothing parameter λ_{HP}, and $N_t^{(\lambda_{HP})}$ is the corresponding residual (noise) component. T is the number of steps per game, ν_{obs} the agent's observed fundamental value, L_t its inventory position with normalization constant $\bar{L}$, equal to the hard position cap of Section 2.2 ($\bar{L} = 500\,k_i$), c_t its cash holding, and B_t its account value.

the next revelation event; and (iii) *private account information* (3 variables): normalized position, cash, and account value. Table 1 lists the full observation vector. Note that the observation vector treats all agents symmetrically. In particular, an agent is not told whether its fundamental signal is precise or noisy. Whether to trust the signal is something each agent has to learn from its trading experience, exactly as a real trading firm learns over time whether its data feeds and forecasts add value. This continuous and high-dimensional state space allows for realistic trading strategies. Simpler tabular learning methods such as Q-learning are not applicable to a problem of this size, as we discuss in Section 3.5.

3.2 Actions and Market Power

At each step, agent i's policy network outputs a continuous action $a_i \in [-1, 1]$, which is mapped to an order in $\{-k_i, 0, k_i\}$ by equal-width thresholds: sell k_i units if $a_i < -1/3$, hold if $|a_i| \leq 1/3$, and buy k_i units if $a_i > 1/3$. The trading capacity k_i is the agent's *market power*, the maximal order size it can submit per step, and it is a key dimension of heterogeneity in our experiments. Inventories are capped by the environment at $\pm 500\,k_i$, with orders that would breach the cap

clipped, but the cap is unreachable in practice. A game lasts $T = 50$ steps with per-step orders of at most k_i, so the largest attainable position is $50\,k_i$, one tenth of the bound. Agents can therefore build large long or short exposures within a game, and no result is driven by the position cap.

Two design choices deserve comment. First, the policy operates in a continuous action space even though executed orders are discrete. This is deliberate. Continuous policy parameterizations are the standard, numerically stable formulation for the policy-gradient methods we employ, they allow the agent to express graded conviction during learning, and they make the architecture invariant to the trading capacity k_i, since the same network design serves agents of any size and the action grid could be refined (e.g., to $\{-k_i, \ldots, -1, 0, 1, \ldots, k_i\}$) without changing the model. Second, the coarse executed-order grid reflects realistic constraints on per-period order sizes while keeping the strategic problem transparent. An agent's choice is the direction of trade, and its market power scales the impact.

3.3 Rewards: Aligning Learning with Trading Profits

The reward function determines what the agents optimize, and its design is one of the central methodological challenges in building realistic AI traders. A common choice in financial RL applications, rewarding unrealized portfolio gains step by step, is economically wrong in a market with price impact. An agent rewarded on paper gains learns to buy continuously, inflating the price and its own mark-to-market wealth, without regard for the cost of eventually unwinding the position. The economically correct objective is the realized trading profit at liquidation,

$$\mathcal{R}^{\text{terminal}} = c_T - c_0, \tag{5}$$

where c_T is the cash balance after the forced liquidation at the revealed fundamental value. This objective, however, provides only a single sparse signal at the end of each game, which makes training deep RL agents slow and unstable, because the agent must assign credit for the terminal outcome across hundreds of decisions.

We resolve this tension through reward shaping (Ng et al., 1999). Let $x_t \in \{-k_i, 0, k_i\}$ denote the order that agent i executes at step $t = 1, \ldots, T$ (we suppress the agent index). At each non-terminal step the agent receives

$$\mathcal{R}_t = x_t \cdot (\nu_{\text{obs}} - p_t), \tag{6}$$

the instantaneous gain or loss of the trade, executed at the post-impact price p_t of equation (2) and evaluated at the agent's best estimate of fundamental value. At the terminal step, the reward includes a correction term that reconciles the agent's estimate with the truth:

$$\mathcal{R}_T = x_T \cdot (\nu_{\text{obs}} - p_T) + (\nu - \nu_{\text{obs}}) \cdot X_T, \tag{7}$$

where $X_T = \sum_{t=1}^{T} x_t$ is the accumulated position at liquidation. The following result establishes

that the shaped reward changes the geometry of learning without changing the objective, or the equilibria of the game.

Proposition 2 (Reward Equivalence)**.** *Suppose rewards are not discounted. Then:*

(i) On every sample path, the cumulative shaped reward equals the realized trading profit:

$$\sum_{t=1}^{T-1} \mathcal{R}_t + \mathcal{R}_T = c_T - c_0. \tag{8}$$

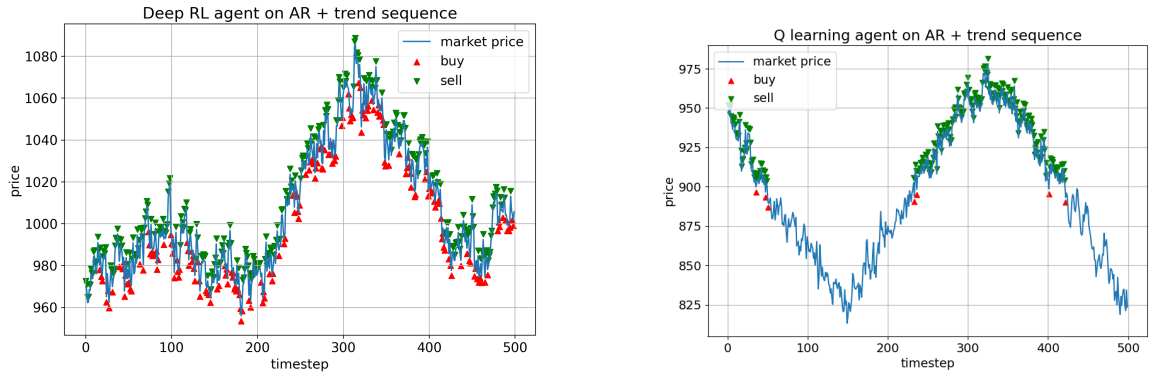
(ii) Consequently, for every profile of policies, every agent's expected total shaped reward equals its expected trading profit. The shaped game and the terminal-profit game have identical best responses and identical (exact and ϵ-) Nash equilibria.

(iii) At every non-terminal step, $\mathcal{R}_t$ is measurable with respect to the agent's own information set (its private signal, the public market history, and its own account): no information about ν beyond the agent's signal enters the reward before the revelation event.

The proof is in Appendix A. In our experiments the training discount factor is set to one (Appendix C), so the proposition applies exactly rather than approximately. Two properties of this construction are essential. First, by part (iii) the shaped reward respects each agent's information set. The gap $\nu - \nu_{\text{obs}}$ enters only at the revelation event, exactly when this information becomes public, so no information about the true fundamental value leaks through the reward channel. Second, the shaped reward encourages fundamentalist trading (buying below and selling above the perceived fundamental value yields positive step rewards) while still permitting the agent to profit from local price patterns, since the cumulative reward equals realized profit. Exploiting such patterns requires enduring negative step rewards before reaping larger gains, which is precisely the capacity for multi-step strategies with delayed payoffs that distinguishes deep RL.

3.4 Learning Algorithm: Intuition

Each agent's policy is a deep neural network trained with Proximal Policy Optimization (PPO; Schulman et al., 2017), a state-of-the-art policy-gradient algorithm. Conceptually, PPO iterates between two operations. The agent collects experience by trading in the market under its current policy, and it then updates the policy to increase the probability of actions that performed better than expected. The defining feature of PPO is that each update is *clipped* to remain close to the previous policy. This conservatism solves the central difficulty of our setting. Because all agents learn simultaneously, each agent faces a non-stationary environment, in which the opponents it trains against today behave differently tomorrow. Aggressive policy updates in such an environment trigger destabilizing feedback loops in which agents chase each other's policy changes. Clipped updates keep the co-evolution of strategies gradual and the training process stable. PPO also natively handles our continuous, high-dimensional observation space and continuous actions through function approximation, capabilities that tabular methods lack by construction. The complete optimization problem, network architecture, and training protocol are described in Appendix C.



(a) Deep RL (PPO) **(b)** Q-Learning

Figure 1. Deep RL versus Q-learning in an illustrative single-agent game. One learning agent trades against rule-based agents, motivated as block traders in a Kyle-type model, whose gradual order execution generates persistent local price trends (Appendix F). Both learning agents are trained for 100,000 steps. The deep RL agent learns to exploit the local price patterns, buying at local troughs and selling at local peaks, while the Q-learning agent fails to learn a useful trading strategy.

3.5 An Illustrative Game: The Limits of Q-Learning

Before turning to the multi-agent analysis, we substantiate our choice of deep RL, and the claim that our findings are out of reach for the tabular Q-learning technology used throughout the prior literature, in a deliberately simple game. The game strips away all strategic interaction and multi-agent complexity. There is only *one* learning agent, and it plays against *rule-based* agents whose behavior is fixed. The rule-based agents can be motivated as block traders in a Kyle-type model: large traders who split a sizable position into a sequence of orders over time. Because their order flow moves prices, their gradual execution generates persistent, predictable local price trends, which we implement as a price process with autoregressive and trend components (Appendix F). Market impact for the learning agent is turned off, and there is no fundamental revelation. Both simplifications strip the task to its core: without market impact the agent cannot move prices against itself, and without a revelation event there is no terminal payoff to reason about, so success requires nothing more than recognizing and riding the given local trends. The learning agent's task is therefore the simplest possible dynamic trading problem. It must recognize the local trends created by the block traders and ride them, buying low and selling high. Succeeding requires only a modest multi-step strategy, and an algorithm that fails here is unlikely to learn the far more demanding manipulative strategies that are the subject of this paper.

Figure 1 shows the outcome. Trained for the same number of steps, the PPO agent learns to buy at local troughs and sell at local peaks, profitably riding the trends created by the block traders. The Q-learning agent fails to learn any useful strategy. The reason is structural, not a matter of tuning. The first obstacle is dimensionality. Our agents observe 23 continuous variables, and tabular Q-learning requires discretizing this space. Even a binary discretization, the coarsest

possible resolution of two bins per variable, yields $2^{23} \approx 8 \times 10^6$ states, and any resolution fine enough to be economically useful multiplies that count astronomically. Dou et al. (2026a) report training runs of 10 to 200 billion episodes for far smaller, hand-selected discrete state spaces. The second obstacle is the horizon. Exploiting a trend requires holding a position through adverse steps before the payoff arrives, which demands propagating value information across long horizons in a continuous state space, exactly what neural network function approximation provides and tabular lookup does not.

The lesson of this simple game carries directly to the main analysis. Even in a setting with a single learner and entirely mechanical opponents, Q-learning fails to learn dynamic trading strategies at any computationally feasible state resolution. In principle a tabular method could represent such strategies on an astronomically fine grid with correspondingly astronomical amounts of experience. At realistic scales of compute it cannot. The manipulative strategies of our multi-agent experiments are strictly harder, because they require *generating* price patterns strategically against adaptive opponents rather than exploiting given ones. The manipulation channel we document is therefore invisible to studies based on simpler AI, and it will become more relevant as deployed trading systems become more capable.

3.6 Equilibrium Selection: Intuition

The second methodological challenge is, in our view, the central one for the entire research program of simulating markets with learning algorithms, and it has not been systematically addressed in the literature on AI trading. The question is which trained agents constitute a valid market equilibrium. For a game among deep learning algorithms with continuous states and simultaneous learning, no analytical equilibrium can be derived, which is precisely why simulation is the appropriate tool. But simulation creates a discipline problem of its own. Training multi-agent systems is noisy, and different runs can end in very different places. Some runs produce poorly converged policies that no profit-maximizing firm would ever deploy. Others favor one agent by accident of the tuning choices. If we simply reported average outcomes across all runs, we could not distinguish economic findings from training artifacts, and a skeptical reader could attribute any result to a badly tuned algorithm. An explicit equilibrium criterion is therefore not a technical refinement. It is what gives the simulation results economic content, and we regard our selection procedure as a methodological contribution in its own right.

Our solution builds on and extends the NashConv concept from computational game theory (Timbers et al., 2022). The idea is simple. A profile of policies is an equilibrium when no agent can gain by changing only its own policy. NashConv measures how far a profile is from this ideal by summing the gains that agents could obtain through unilateral deviations. It equals zero exactly at a Nash equilibrium, and a profile in which no agent can gain more than ϵ is an ϵ-Nash equilibrium. Exact best responses cannot be computed in our environment, so we estimate deviation gains with the same learning technology the agents themselves use. For each trained market we freeze all agents except one, let the remaining agent continue training alone against its fixed opponents, and record

how much it improves. If no agent improves by more than the threshold, the market passes the filter, and only markets that pass enter our results, with the number of surviving runs disclosed in every figure. Two properties of this test should be stated plainly, because they define what our results can and cannot claim. First, the test searches for deviations with the same learning technology the agents use, so it can *reject* a candidate equilibrium when it finds a profitable deviation, but it can never *prove* an equilibrium, since a better policy might exist that the search does not find. Surviving the filter therefore means "not rejected," and we use equilibrium language in this algorithm-relative sense throughout. Second, precisely because learned searches can miss simple alternatives, we complement them with deviations that can be evaluated exactly and without any learning, the most important being the policy that never trades, which always earns exactly zero. We call this exact benchmark the *no-trade test*. Section 4.6 applies the no-trade test and shows that it rejects an equilibrium reading of the most extreme information-gap cells while leaving all other results untouched.

The procedure has a natural economic interpretation. The end result can be read as a market of jointly and optimally tuned algorithms. Each surviving profile is one in which every trading firm's algorithm is a best response, within reach of its own retraining technology, to the algorithms of its competitors, so no trading firm would unilaterally replace what it deployed. Deviation gains below the threshold mean exactly that further in-house development does not pay. As a robustness check, we also implement a two-stage variant in which each agent is first tuned in isolation against randomly acting opponents before joint training, mirroring firms that calibrate algorithms in-house before deployment. Both procedures yield the same qualitative results, which gives us confidence that our findings reflect genuine equilibrium phenomena rather than artifacts of one optimization pipeline. Implementation details, the threshold, both filter variants, and full survival statistics are reported in Appendix D.

4 Main Results

4.1 Experimental Design and Base Configuration

Our experiments vary the market composition (the number of agents, their information quality, and their market power) while holding total market activity constant. In the base configuration, each episode consists of 10 consecutive games of $T = 50$ steps each (500 decision points). The fundamental value starts at $\nu_0 = 1{,}000$. At the end of each game, all positions are liquidated at the prevailing fundamental value, the market price is reset to that value, and the fundamental value jumps to $\nu_g = \nu_{g-1} + \gamma_g$ with $\gamma_g \sim \text{Uniform}(-100, 100)$,[2] so each game begins with a fresh mispricing $\nu_g - p$ equal to the fundamental innovation, unknown to the uninformed agents. Each uninformed agent receives a new private signal $\nu_{\text{obs}} = \nu_g + \epsilon$ at the start of each game, with $\epsilon \sim \text{Normal}(0, \phi^2)$;

[2]The fundamental value is floored at a small positive level so that prices remain positive. The floor essentially never binds in the base configuration. At the start of an episode, the same initial mispricing is created by randomizing the initial price around ν_0, $p_0 = \nu_0 + \text{Uniform}(-100, 100)$.

the information gap of Section 2.2 scales ϕ relative to the mean absolute fundamental innovation of 50, so a 50% (100%) gap corresponds to $\phi = 25$ ($\phi = 50$). All agents start each episode with zero positions and cash $c_0 = \$2{,}000{,}000$, large enough that funding constraints never bind. The price-impact parameter is $\lambda = 1$, so one unit of net order flow moves the price by 1, that is, by roughly 0.1% of the price level, and the full one-sided market capacity of 16 units moves it by roughly 1.6%. We do not calibrate λ to a specific market; it is a model parameter that fixes the units of market depth, and we vary it in Section 5.

Total market power is fixed at 16 units per step. The baseline includes one informed agent with trading capacity $k_{\text{informed}} = 2$, which is 12.5% of the market, and we vary the number of uninformed agents M, splitting the remaining capacity equally among them, $k_{\text{uninformed}} = (16 - k_{\text{informed}})/M$. Market power has a direct economic interpretation. An agent's trading capacity is the maximal order it can place per period, which stands in for the capital, risk limits, and position constraints that cap how much of the market's flow a real trading firm can be. An agent's capacity as a share of total capacity is therefore its potential share of aggregate order flow, and through the pricing rule (2) it measures how much of the price movement that agent can cause on its own. Holding total capacity fixed isolates the effect of market composition from the effect of aggregate trading activity. The information gap takes values from 0% to 100%. Agents are trained with a discount factor of one, so Proposition 2 applies exactly, and each agent maximizes its expected realized trading profit per episode.

Training technology. All results in the paper come from agents trained with one fixed set of tuning parameters, chosen before we looked at any outcome. The choice followed a simple documented rule. We first ran a broad random search over tuning parameters for every market configuration and scored each candidate only by how reliably it converged, using the equilibrium diagnostics of Appendix D. We then froze a single global configuration by averaging the best-converging candidates across configurations. Prices, profits, and all other economic outcomes played no role in this choice. Freezing one technology for every market has an important benefit. When only the market structure changes and the learning technology stays the same, differences across markets must come from the economics and cannot come from differences in tuning. The original tuning search and an alternative freezing rule are retained as robustness analyses for the Internet Appendix.

What an equilibrium means in our design. It is useful to spell out the layers of the equilibrium concept that our experiments implement, because an equilibrium among learning algorithms has more components than an equilibrium in a standard model. The first layer concerns the tuning of the algorithms. Tuning choices are strategies in the meta-game in which trading firms develop their algorithms. Our design addresses this layer in a deliberately limited and verifiable sense. All agents deploy the same frozen configuration, and the equilibrium filter checks, for every surviving market, that no agent's own learning technology, at this tuning, finds a better policy against its competitors. The filter searches deviations in policies, not in the tuning itself, so what survives is a

market of jointly and consistently tuned algorithms in which no trading firm's own retraining found a profitable change, which is consistency rather than proven optimality. The two-stage robustness procedure of Section 3.6, in which each agent is instead tuned in isolation before joint training, provides a partial check that this conclusion does not depend on how the tuning was reached. The second layer concerns the policies. Joint training continues until the policy functions have converged, and the filter checks that no agent's continued retraining finds a materially better response to the converged policy functions of the other agents. The third layer is evaluation. We take the filter-passing policy functions, freeze them, switch exploration off, and measure the equilibrium consequences by playing the frozen market through the common evaluation episodes. The final layer is replication for uncertainty quantification. Because training is a random process, the same market structure can converge to different equilibria, so we repeat the entire procedure many times with different random seeds. Every surviving run delivers one filter-passing market together with its frozen evaluation. Variation *across* these independent runs measures how much learned equilibria differ, and this is the uncertainty that our standard errors quantify, while the common evaluation episodes hold the environment randomness fixed, so that differences across market structures are never driven by different shocks.

Seed ensembles, equilibrium selection, and evaluation. A single training run is one realization of a noisy process, so we never rely on one run. For each market configuration we train an ensemble of S independent runs that are identical except for the random seed, which controls the network initialization, the exploration noise, and the training-environment draws. We use $S = 32$ for the information-gap grid and $S = 16$ for the extension experiments on market power, informed competition, and robustness. The smaller ensembles reflect the computational cost of the equilibrium filter. We tested two cheaper shortcuts of the filter, both failed a pre-agreed validation exercise, and we therefore kept the full filter and reduced the ensemble size instead (Appendix D). Training continues until the agents' policies stop changing, and we then ask whether the resulting market has converged in the economic sense of Section 3.6, meaning that continued unilateral retraining yields an average gain across agents below $\epsilon = 3$. Appendix D states the exact criteria and what each certifies, including a stricter worst-case per-agent variant. This test can reject a candidate equilibrium, and it cannot prove one, a point we return to in Appendix D. Runs that fail the test are discarded. We apply the test in two forms that answer two different questions. The *absolute* filter discards runs in which retraining moves rewards substantially in either direction. Large movements in the negative direction signal unstable training, so this filter doubles as a convergence screen, and it is our primary criterion. The *one-sided* filter counts only positive retraining gains, the pure question of whether any agent can do better, and is reported as a sensitivity check. Each surviving market is then evaluated out of sample. We freeze the learned policies, switch exploration off, and play them through 10 fresh evaluation episodes whose environment draws are common across all market configurations, so that any two markets are compared under identical shocks. This separation matters. Training determines what the agents learn, and evaluation mea-

sures how the learned market behaves. We are fully transparent about selection, reporting in every figure and table how many of the S runs survive the filter, and Section 4.6 shows that the survival pattern is itself economically informative.

Uncertainty quantification. Each surviving run yields one number per metric, its average over the 10 common evaluation episodes. Because the surviving runs are statistically independent, these run-level numbers give us clean uncertainty measures. Figures and tables report their mean together with the standard error across runs, and statements comparing market configurations are based on two-sample comparisons of the run-level values. Where only one or two runs survive the filter, we quote the individual values and flag the entry rather than reporting a standard error.

Before turning to the results, we state four hypotheses. Formulating them up front serves two purposes. It makes the empirical analysis a test of the economic mechanism of Section 2.3 rather than an open-ended description of simulation output, and it protects us from reading patterns into results after the fact. Each hypothesis is a direct implication of the mechanism, in which the informed agent's manipulation works through the inference and the trading capacity of the uninformed side.

Hypothesis 1 (Information asymmetry)**.** *Price deviations from fundamentals increase with the information gap between informed and uninformed agents.*

The information gap is the source of the manipulation opportunity. The noisier their own signals, the more the uninformed agents must rely on prices and volume to infer the fundamental, and the more effectively the informed agent can distort exactly those market statistics. With no gap there is nothing to exploit, so price deviations should grow as the gap widens.

Hypothesis 2 (Concentration of the uninformed side)**.** *Holding the total market power of the uninformed side constant, price deviations are largest when it is concentrated among few agents, and decline as it fragments across many agents.*

Fragmentation here means dividing the uninformed side's fixed total capacity across more agents, so that each one commands a smaller share. Manipulation needs amplification. A single uninformed agent offers only one follower whose demand the informed agent can trigger, while many small uninformed agents respond idiosyncratically and partially offset each other. The mechanism therefore predicts an interior peak, with the strongest distortions when a few uninformed agents each command substantial capacity and their joint trend-following creates powerful momentum.

Hypothesis 3 (Market power of the manipulated)**.** *Manipulation requires uninformed agents with sufficient aggregate market power, so price deviations are larger when the informed agent's own share of market power is smaller.*

Momentum-based manipulation is a lever. The informed agent only initiates the price movement, and the uninformed agents' capacity does the amplifying, so shrinking the informed agent's share (and thus enlarging the uninformed side) should strengthen the distortion. This hypothesis

is written around the momentum mechanism, and the results will show that it captures one of two relevant regimes. It describes markets in which the informed agent is small; Section 4.4 uncovers a second regime, based on under-correction rather than momentum, when the informed agent itself dominates the market, so the two regimes together produce a U-shape rather than a monotone pattern.

Hypothesis 4 (Strategic adaptation)**.** *Facing uninformed agents, the informed agent shifts from a fundamentalist strategy to one based on market signals, consistent with trading on other agents' reactions.*

If manipulation works by steering the reactions of others, the informed agent's policy should reflect it. Trading alone, the informed agent needs only the gap between price and fundamental value. Trading against uninformed learners, it should additionally track the market signals that its victims follow, because those signals tell it when the momentum it created is ready to be sold into. Section 6 tests this prediction inside the learned policies themselves.

4.2 Emergence of Manipulative Behavior

The central finding of the paper is that informed AI traders learn manipulative strategies when interacting with uninformed AI traders. We illustrate the two recurring strategies with representative simulated episodes from a market with one informed and two uninformed agents, each with trading capacity $k = 5$, one third of the total market capacity, which is 15 in this illustrative configuration. These episodes are illustrative examples, selected from the evaluation set of this configuration for visual clarity, and other episodes from the same runs show the same patterns. All quantitative results in this paper come from the seed ensembles of Section 4.1, and it is their averages, not individual episodes, that carry the paper's claims.

Pump and Dump. Figure 2 shows a full evaluation episode of ten consecutive games, 500 trading steps in total. Each of its three panels displays the same market price (solid line) and fundamental value (dashed line), together with the executed orders of one agent. Red triangles pointing up mark purchases and green triangles pointing down mark sales, each drawn at the price at which the order executed. The left panel is the informed agent and the two other panels are the two uninformed agents. Figure 3 zooms into a single game of this episode, and we discuss the strategy using this zoom.

In the game shown in Figure 3, the fundamental value is 1,025 and the price starts slightly below it, so the asset is initially undervalued. In the left panel, the informed agent buys during the first ten steps and pushes the price up. Instead of stopping when the price reaches the fundamental value, it *keeps buying* beyond 1,025, deliberately creating upward momentum. The middle and right panels show the reaction. Both uninformed agents, who cannot tell informed order flow from other trading and have learned to treat price increases as information, buy through nearly the entire rise, paying prices that climb far above the fundamental value and driving the price to about

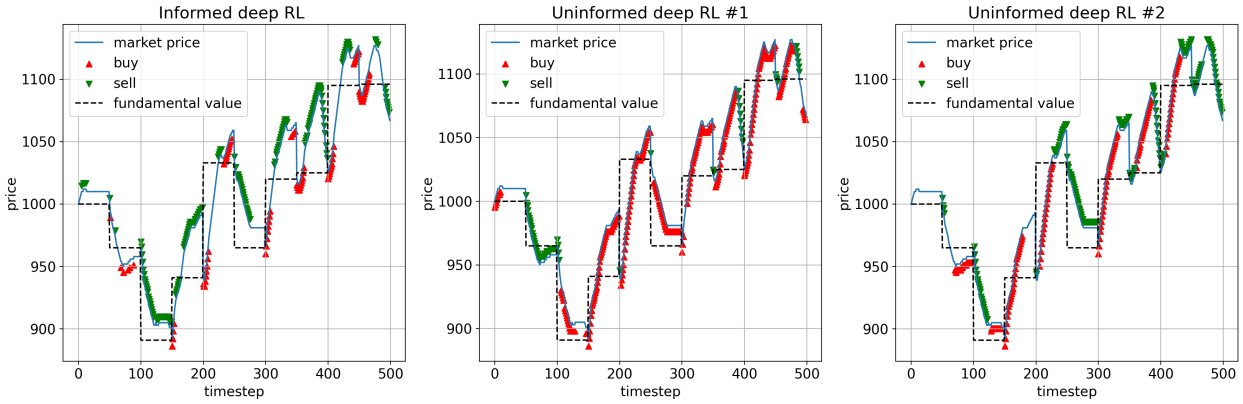


Figure 2. Pump-and-dump manipulation: full episode. Each panel shows the market price, the fundamental value, and one agent's executed buys and sells (left: the informed agent; middle and right: the two uninformed agents). The informed agent buys beyond the fundamental value, the uninformed agents amplify the induced trend, and the informed agent exits at inflated prices.

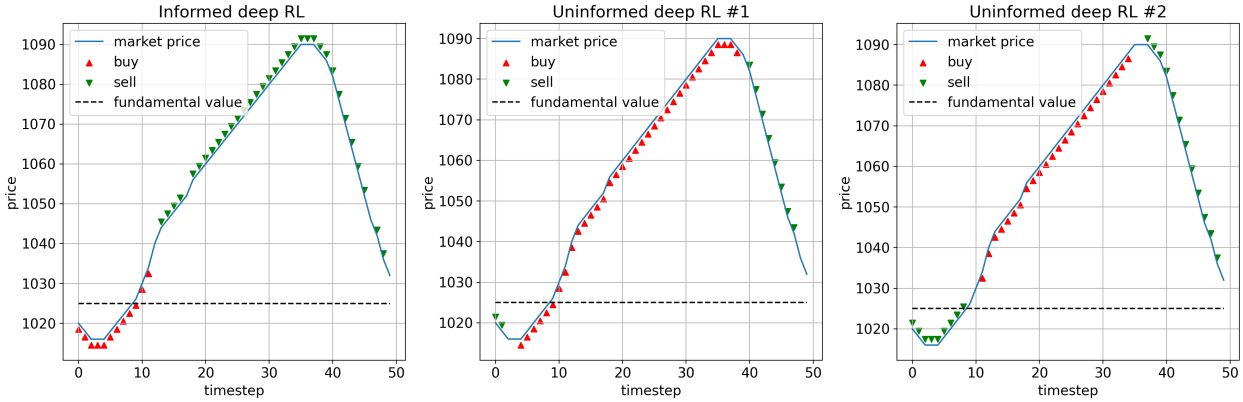


Figure 3. Pump-and-dump manipulation: single game (zoomed view of Figure 2). The informed agent's deliberate overbuying beyond the fundamental value creates a local price bubble that uninformed agents amplify through trend-following behavior.

1,090, roughly six percent too high. The informed agent has switched sides long before the peak. From around step 13 onward the left panel shows a nearly unbroken sequence of green sell markers, so the informed agent unloads its position, and more, into the demand of the uninformed buyers at ever more inflated prices. In the final steps the bubble deflates and the uninformed agents sell at falling prices, after the informed agent has already exited. The strategy sacrifices profit early, since buying above the fundamental value yields negative shaped rewards, in order to create the very demand it later sells into. It is a multi-step plan that emerges entirely from trial-and-error learning.

Strategic Under-Correction. Figures 4 and 5 document a subtler strategy in the same format. Figure 4 shows a full ten-game evaluation episode, with the same panel layout and markers as before, and Figure 5 zooms into a single game, which we now walk through.

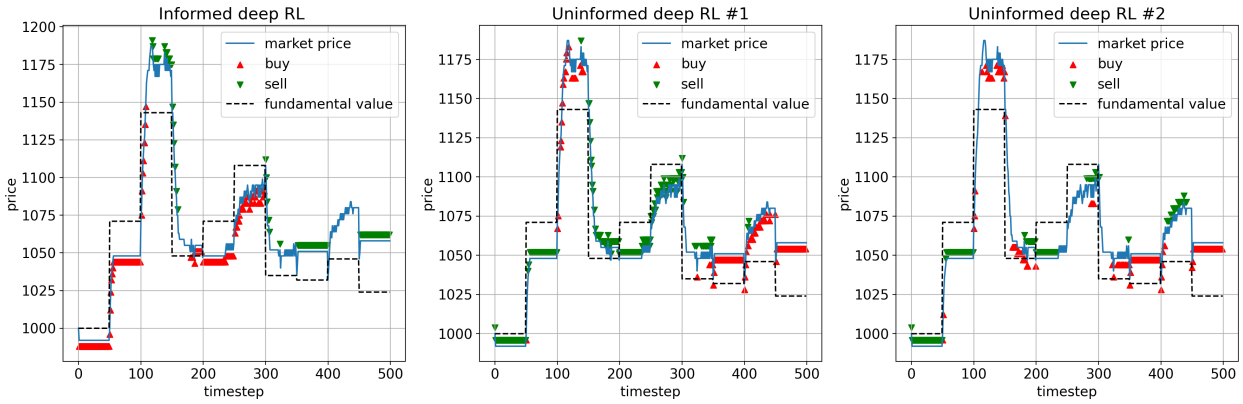


Figure 4. Strategic under-correction: full episode. Each panel shows the market price, the fundamental value, and one agent's executed buys and sells (left: the informed agent; middle and right: the two uninformed agents). The informed agent deliberately maintains a persistent gap between the market price and the fundamental value rather than correcting it, profiting continuously from the preserved mispricing.

In this game the fundamental value is 1,143 and the price starts around 1,070, so the asset is again undervalued. The left panel shows the informed agent buying in the first ten steps, correcting the mispricing, but the correction does not stop at the fundamental value. Fueled by the uninformed agents' buying, visible in the middle and right panels, the price overshoots to about 1,185. What happens next is the heart of the strategy. Rather than selling the price back down to the fundamental value, the informed agent *maintains* the overpricing for the rest of the game. The price oscillates in a narrow band roughly 25 to 40 above the fundamental value, and the informed agent sells small quantities at the local peaks of this band, while the two uninformed agents keep buying near its local dips, at prices they should not be willing to pay. The one participant who knows the price is too high thus stabilizes it at the wrong level, harvesting a stream of small profitable sales from buyers who cannot know better. By under-correcting, the informed agent preserves the mispricing as a renewable source of profit rather than consuming it at once, in direct contradiction to the benchmark prediction that informed trading eliminates price gaps. From the market's perspective, the informed agent has become a guardian of inefficiency. The one trader who knows the price is wrong is also the one who profits from keeping it wrong.

Both strategies produce persistent price distortions, and both emerge endogenously. Nothing in the agents' objectives, architectures, or training rewards manipulation per se. The strategies are discovered as individually optimal responses to the competitive environment. We next quantify how the severity of these distortions depends on market structure.

4.3 Impact of Information Asymmetry

We now quantify market inefficiency across market structures. Our headline metric is the *price gap*: the average absolute deviation of the market price from the fundamental value over a terminal window $\mathcal{W}$ of each game, averaged across the games of an episode and expressed as a percentage

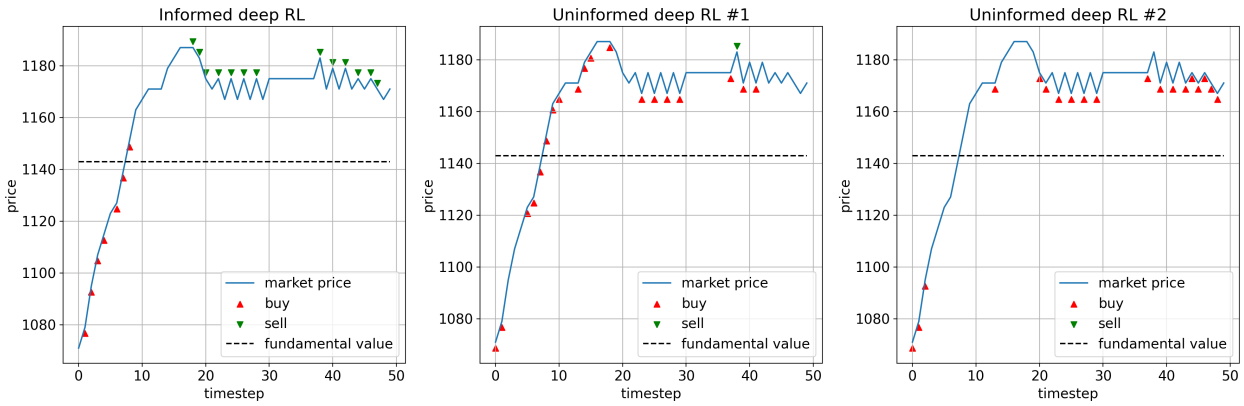


Figure 5. Strategic under-correction: single game (zoomed view of Figure 4). After the price overshoots the fundamental value, the informed agent maintains the overpricing rather than correcting it, selling small quantities at the local peaks of the preserved gap.

of the mean absolute initial mispricing,

$$\text{Price Gap} = \frac{100}{\mathbb{E}|\gamma|} \cdot \frac{1}{G} \sum_{g=1}^{G} \frac{1}{|\mathcal{W}_g|} \sum_{t \in \mathcal{W}_g} |p_t - \nu_g|, \qquad \mathbb{E}|\gamma| = 50, \tag{9}$$

where $G = 10$ is the number of games per episode and $\mathcal{W}_g$ collects the last five steps of game g. The normalization constant needs a word of explanation. Here γ denotes the fundamental innovations of Section 4.1, the random jumps in the fundamental value at the start of each game, and their mean absolute value $\mathbb{E}|\gamma| = 50$ is the size of the typical initial mispricing. Dividing by it makes the price gap read as the share of the typical initial mispricing that remains unclosed. A value of 100 means no price discovery at all and a value of 0 means full convergence. The same window, statistic, and filter are used for every price-gap and reward exhibit in the paper. The one exception is the trade-direction ratio of Section 4.5, a behavioral measure that is evaluated over all trades of the game. We focus on the terminal window because it isolates *persistent* distortions, mispricing that survives until just before the revelation event, when an efficient market would have converged. We computed the metric for all windows of the game, and Table 2 below reports the full-game and early-phase results, so the terminal window is a choice of emphasis rather than of necessity.

Figure 6 contains the paper's first three results. First, more information asymmetry means larger price gaps, at every market size, which confirms Hypothesis 1. Consider the markets with two uninformed agents. When all traders are equally informed, only 7 percent of the typical initial mispricing survives to the end of a game. At a 50% information gap the surviving share rises to 20 percent, and at a 100% gap to 44 percent. The same ordering of the three curves holds for every number of uninformed agents. The economics is the one the mechanism predicts: noisier private signals push uninformed agents to rely more heavily on prices and volume, and this reliance deepens the very vulnerability that the informed agent exploits. Second, concentration of the uninformed side matters, which confirms Hypothesis 2. Along the 50% curve, the price gap is 16.7 percent

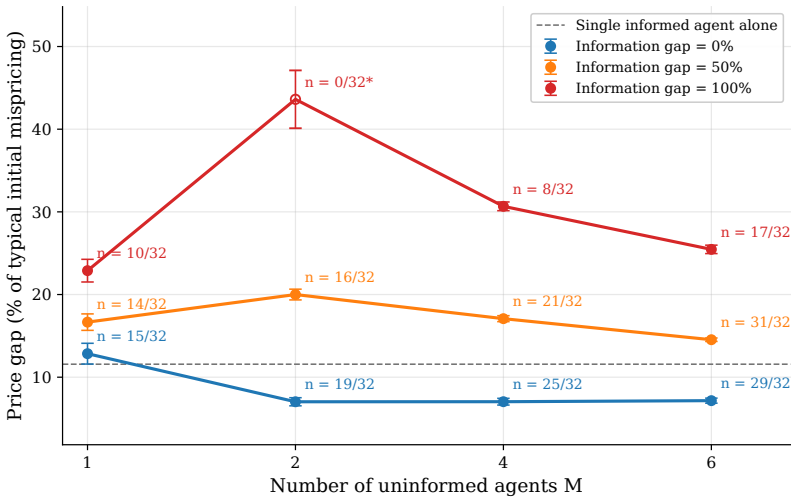


Figure 6. Market inefficiency by market composition. The figure shows the price gap of equation (9), the share of the typical initial mispricing that remains unclosed at the end of a game, for markets with one informed agent (capacity 2 of 16) and M uninformed agents sharing the remaining capacity. Each curve corresponds to one information gap. Lines show means across the surviving runs of the 32-seed ensembles and capped bars show one standard error across runs; survivor counts are annotated at every point and reported in Table 6. The (100% gap, $M = 2$) point has no survivor under the primary filter and shows the unfiltered ensemble mean of 43.6; the two runs that pass the one-sided filter give 35.2 (Section 4.6). The dashed line is the benchmark of a single informed agent trading alone, and the 0% curve is the benchmark in which all traders are equally informed.

with one uninformed agent, peaks at 20.0 percent with two, and then falls to 17.1 percent with four and 14.5 percent with six. The same hump at two uninformed agents appears at the 100% gap. The non-monotonicity is informative about the mechanism. One uninformed agent offers limited amplifying demand to exploit. Two uninformed agents each command large capacity, and their joint trend-following creates powerful momentum. Many small uninformed agents respond idiosyncratically and partially offset each other, which weakens the lever that manipulation requires. Third, with homogeneous information the market is efficient and stays efficient as it grows. The 0% curve is flat at about 7 percent for two or more uninformed agents, which is exactly the corrective competition of the benchmark in Section 2.4, and only the thin two-trader market is elevated.

These differences are precisely measured. The standard errors across surviving runs are between 0.2 and 1.4 percentage points for every filtered point in the figure, so the increase of 13 points from the 0% to the 50% gap at $M = 2$ is roughly fifteen standard errors, and the hump at two uninformed agents is statistically significant against every neighboring market size, with two-sample t-statistics between 3 and 8. The one cell that requires care is (100%, $M = 2$), where no run passes the primary filter. There we plot the unfiltered ensemble mean of 43.6 and note that the two runs that pass the one-sided filter give 35.2, so the qualitative conclusion does not depend on how that cell is treated.

Table 2. Price Gap by Window of the Game

	Info gap	n	Terminal (45–49)	Full game	Early (1–15)	Early/Terminal
$M = 1$	0%	15	12.8	29.3	55.8	4.3
$M = 2$	0%	19	7.0	17.4	36.7	5.2
$M = 4$	0%	25	7.0	16.7	33.8	4.8
$M = 6$	0%	29	7.2	17.3	34.4	4.8
$M = 1$	50%	14	16.7	32.9	58.7	3.5
$M = 2$	50%	16	20.0	29.2	46.4	2.3
$M = 4$	50%	21	17.1	24.4	38.9	2.3
$M = 6$	50%	31	14.5	22.1	36.8	2.5
$M = 1$	100%	10	22.9	41.6	68.1	3.0
$M = 2$	100%	32*	43.6	49.7	59.4	1.4
$M = 4$	100%	8	30.7	36.6	49.4	1.6
$M = 6$	100%	17	25.5	32.9	46.4	1.8

Notes: Price gap in percent of the typical initial mispricing, computed as in equation (9) over the indicated window, averaged across games, evaluation seeds, and the n seed runs surviving the absolute NashConv filter ($\epsilon = 3$; of 32). *No run survives in this cell; the unfiltered ensemble mean is shown (Section 4.6). The one-sided-filter version of this table is in the Internet Appendix and agrees closely.

Because our deviation tests reject an equilibrium reading of the 100% cells with a concentrated uninformed side (Section 4.6), we present the whole 100% row as a description of what learning produces there, and we base all equilibrium statements about Hypothesis 1 on the fully filtered comparison between the 0% and 50% gaps. Survivor counts for every point appear in Table 6.

We include two key economic benchmarks in the figure. The all-informed curve shows what competitive information delivers. Roughly 7 percent of the typical mispricing survives, and the price-based correction share, the fraction of the initial mispricing that the market closes by the end of a game (Figure 13(b) in the Appendix), reaches 0.93 against the collectively optimal ceiling of $T/(T+1) = 0.98$ from Proposition 1(iii). The dashed line shows a single informed trader with no other speculators. Even this monopolist corrects the price only to 11.6 percent, patiently preserving part of the gap, by the same logic through which optimal extraction in Proposition 1 leaves a residual mispricing. Manipulation of uninformed learners produces far larger distortions than either benchmark, at 20 percent of the mispricing at a 50% gap and up to 44 percent at 100%. The resulting ranking, informed competition below informed monopoly below manipulation of uninformed learners, summarizes the paper in one figure.

Two identification caveats belong next to these results, and we state them openly. First, when the number of uninformed agents rises, two other things change as well. The uninformed side holds more independent signals, so its combined information becomes more precise, and each agent's capacity shrinks, so orders become finer. Part of the decline in the price gap from two to six uninformed agents may therefore reflect information aggregation and finer order sizes rather than fragmentation of market power alone, and we read that decline as an upper bound on the pure

concentration effect. The rise from one to two uninformed agents cannot be explained this way, since better aggregate information would push the price gap down, not up, and the market-power result of the next section holds the number of agents and the signals fixed. Experiments that hold aggregate signal precision and order granularity fixed are in progress. Second, a larger information gap makes the uninformed agents' learning signal noisier, not only their information worse, so part of the deterioration at extreme gaps may reflect harder training rather than equilibrium economics. The informed agent's reward is exact in every cell, so its behavioral changes across gaps cannot be explained this way, and in real markets a trader with worse data also receives noisier feedback, so the two channels travel together outside the model as well. A robustness exercise that retrains selected cells on the terminal profit alone is planned for the next version.

Distortions are not an end-of-game artifact. Table 2 reports the price gap over three windows of each game: the terminal five steps, the full game, and the early phase (first fifteen steps). The early-phase gap exceeds the terminal gap in every one of the twelve cells, by factors of 1.4 to 5.2. Bubbles build during the first fifteen steps and persist. The decay from early to terminal is fastest in the all-informed cells (roughly a factor of five) and slowest at the 100% gap (1.4 to 1.8), so the information structure governs the *speed* of price discovery, not only its level.

4.4 Role of Market Power

We next vary the informed agent's share of market power, holding the information gap at 50% and total capacity at 16 units, by increasing k_{informed} from 2 to 14, that is, from 12.5% to 87.5% of the market, with two uninformed agents splitting the remainder.

Figure 7 shows that price inefficiency is *U-shaped* in the informed agent's market power. The price gap is 20.7 percent when the informed agent holds 12.5% of the market,[3] falls to about 17.5 percent at shares between 37.5% and 50%, and rises again to 22.0 percent when the informed agent controls 87.5% of the market. The price-based correction share peaks at 0.83 over the same intermediate range. The left arm of the U confirms the lever logic of Hypothesis 3. When the informed agent is small, distortion works through the uninformed agents' capacity, whose joint momentum is the amplifier that pump-and-dump requires, so shrinking the informed agent strengthens manipulation. The right arm is a regime that the hypothesis, written around the momentum mechanism, did not anticipate. The behavioral evidence of Section 4.5 identifies what it is. As the informed share rises, the fraction of informed trades directed toward the fundamental value rises to essentially 100 percent, and yet the price gap widens. A dominant informed trader does not trade against the fundamental at all. It simply stops correcting early. This is the strategic under-correction of Section 4.2, chosen because closing the gap would extinguish the mispricing the agent earns from. Pump-and-dump is also mechanically unattractive for a dominant insider. Igniting a trend requires round-trip trading, whose churn cost rises with the square of the insider's own large orders

[3]The 12.5% point is the workhorse market of Section 4.3 re-evaluated under the extension protocol, which reuses the first 16 seeds of that cell's 32-seed ensemble. Its mean of 20.7 percent ($n = 9$) is statistically indistinguishable from the 20.0 percent ($n = 16$) of the full ensemble.

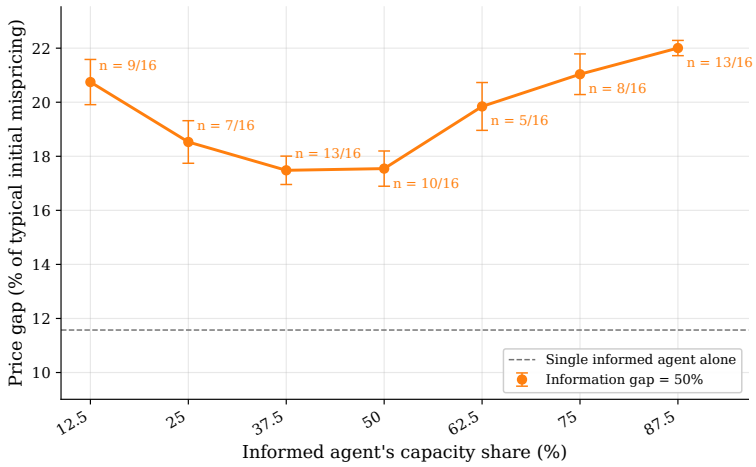


Figure 7. Market inefficiency by the informed agent's share of market power. The figure shows the price gap of equation (9) as the informed agent's trading capacity rises from 2 to 14 units (of 16 total), with two uninformed agents sharing the remaining capacity and the information gap fixed at 50%. Points show means across the surviving runs of 16-seed ensembles under the absolute NashConv filter ($\epsilon = 3$); bars are standard errors across runs; survivor counts are annotated. The dashed line is the single-informed-alone benchmark of 11.6.

by Proposition 1(ii), while the uninformed side is now too small to supply the amplifying demand that would make the round trip profitable. The comparison with the benchmark line makes the manipulation content precise. A monopolist informed trader with no other speculators also under-corrects, but only to 11.6 percent, so roughly half of the 22 percent gap at high informed shares is excess under-correction that appears only when exploitable counterparties are present. Two further patterns support this reading. The informed agent's own profit peaks at an intermediate share, at roughly 1,179 per game for a 25% share, and the uninformed agents' profits rise steadily as the informed share grows, from 116 to 245 per agent, because a dominant and purely corrective informed trader no longer moves prices against them. The two arms of the U are thus the two manipulative strategies of Section 4.2 expressed in market structure. Manipulation works through momentum when the informed agent is small and through omission when it is large, and efficiency is greatest when market power is balanced. Small informed traders facing large uninformed crowds, arguably the empirically relevant case in modern markets where any single sophisticated firm is small relative to aggregate algorithmic flow, sit on the left arm, where manipulation is momentum-based and hardest to distinguish from ordinary trend trading.

The statistical basis mirrors the previous section. Each point in Figure 7 averages the surviving runs of a 16-seed ensemble, with survivor counts between 5 and 13 shown in the figure, and standard errors across runs between 0.3 and 0.9 percentage points. Both ends of the U differ from the valley by several standard errors. The dashed benchmark line again marks the single informed agent trading alone.

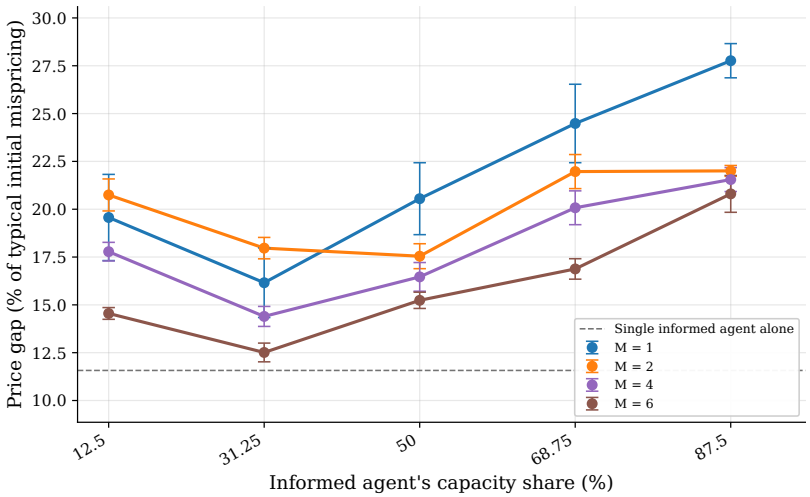
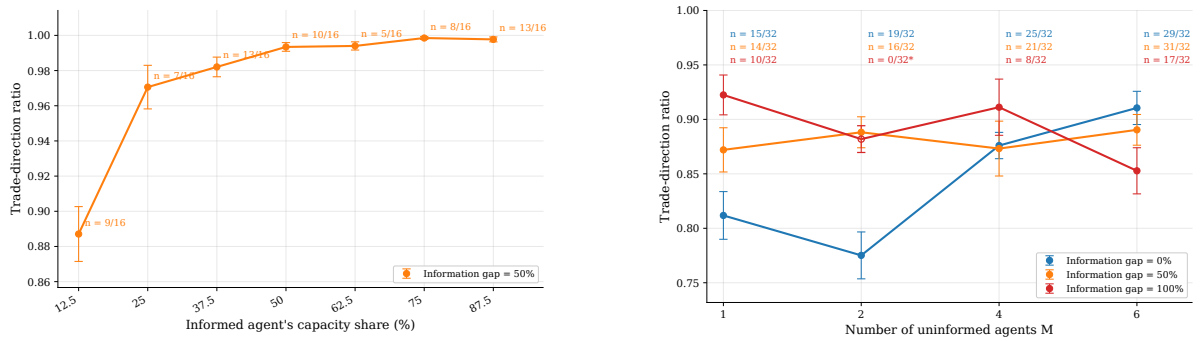


Figure 8. Market inefficiency across the number of uninformed agents and the informed agent's capacity share. The figure shows the price gap of equation (9) against the informed agent's capacity share (12.5% to 87.5% of the total capacity of 16), with one line per number of uninformed agents M, at a 50% information gap, so each line traces the U-shape of Figure 7 for a different number of uninformed agents. Lines show means across the surviving runs of 16-seed ensembles under the absolute NashConv filter ($\epsilon = 3$) and capped bars show one standard error across runs. The dashed line is the single-informed-alone benchmark of 11.6. Survivor counts per point range from 4 to 16 of 16 and are reported, together with all values, in Table 10.

The U-shape is not specific to two uninformed agents. Figure 8 repeats the experiment for every number of uninformed agents $M \in \{1, 2, 4, 6\}$ at the 50% gap, with the informed capacity share on the horizontal axis, so each line traces its own U and the valleys are directly visible. An interior minimum appears in every line. The efficient valley sits at intermediate informed shares throughout, at a 31% share for $M = 1$, 4, and 6 and at 50% for $M = 2$, and the high-share arm is elevated everywhere, reaching about 21 to 28 percent at an 87.5% informed share. The full grid also clarifies how market power and the number of uninformed agents interact. For $M \geq 2$, fragmenting the uninformed side shifts the whole curve down at every capacity split, which is Section 4.3's result on the number of uninformed agents holding split by split. The concentration hump at $M = 2$, in contrast, is a phenomenon of small informed shares. It appears at the 12.5% and 31% shares, exactly where manipulation is momentum-based and needs the uninformed side's concentrated buying power, while at informed shares of 50% and above the thinnest market ($M = 1$) shows the largest gaps and fragmentation improves prices monotonically. Detailed values, standard errors, and survivor counts for all twenty cells are reported in Table 10 in the Appendix.

4.5 Direct Evidence: The Direction of Informed Trades

The price gap measures outcomes. We now measure the behavior that produces them. For each agent we compute the *trade-direction ratio*, the fraction of its trades that point toward the



(a) By the informed agent's capacity share

(b) By market composition

Figure 9. Trade-direction ratio of the informed agent: the fraction of non-holding trades directed toward the fundamental value, over all trades of the game, under the absolute NashConv filter ($\epsilon = 3$; survivor counts annotated). Panel (a) varies the informed agent's capacity at a 50% gap with $M = 2$. Panel (b) varies the number of uninformed agents and the information gap at $k_{\text{informed}} = 2$. The competitive benchmark implies 100%. In the all-informed markets of panel (b) the ratio is mechanically lower, because many trades execute at near-zero mispricing, where the strict test cannot score them as corrective (see text).

fundamental value. A trade counts as corrective when the agent buys while the price is below the fundamental value or sells while the price is above it, and the mispricing is always measured at the pre-trade price, so that a trade is never scored against the price movement it has itself caused.[4] Because this is a measure of behavior rather than of outcome persistence, we evaluate it over all trades of the game. The version restricted to the terminal window rests on very few trades in the most efficient markets and is reported in the Internet Appendix. It agrees with the full-game version within about two percentage points wherever both are defined. Under the competitive benchmark of Section 2.4, an informed trader should have a ratio of 100%.

Two patterns stand out in Figure 9. The first is in panel (a), which varies the informed agent's market power. The informed ratio rises steadily with the informed share, from 88.7 percent at a 12.5% share to above 99 percent once the share reaches 50%. Together with the U-shape of Figure 7, this cleanly separates the two manipulative regimes at the level of behavior. A small informed agent trades *against* the fundamental value more than one time in ten. That is the signature of pump-and-dump, because igniting momentum requires buying an asset that is already too expensive. A dominant informed agent does the opposite. Essentially every one of its trades points in the corrective direction, and yet the price gap widens, because its distortion consists of trading too little rather than in the wrong direction. Under-correction is invisible to a direction-based measure precisely because the damage is done by omission. The price gap and the trade-direction ratio are

[4] Two further details guard the metric. First, holds are never counted as corrective. Trades placed when the price already equals the fundamental value are not counted as corrective either, but they stay in the denominator, because a trade at a correct price pushes the price away from it. Dropping those ties would raise the pooled ratio by 1.3 percentage points, and would raise no cell with a positive information gap by more than 2 percentage points (Internet Appendix). Second, trades are pooled across a group of agents before averaging, so agents that barely trade cannot dominate the statistic.

therefore complements, and only their combination identifies both strategies. The second pattern concerns the uninformed agents. Their own trade-direction ratio, not shown in the figure, hovers near one half in every market, between 0.43 and 0.70, so their trades carry almost no information about whether the price is too high or too low. This is the trend-following behavior that Section 6 finds inside their policies, and it is exactly the behavior the informed agent's strategies feed on. Panel (b) also explains why the informed ratio should not be read as a stand-alone manipulation gauge across information gaps. Even textbook pump-and-dump generates mostly corrective trades, because only the short ignition phase pushes the price away from the fundamental value, while the long and profitable exit consists of selling an overpriced asset, which the metric scores as corrective by construction. Under-correction produces no anti-corrective trades at all. High ratios at a 100% gap are therefore fully compatible with severe manipulation, and the manipulation itself is visible in the price gap and the correction share, which do fall monotonically with the gap. Panel (b) adds one further caution for reading the all-informed markets. There the informed ratio appears lower, but for a mechanical reason rather than a strategic one. When prices track fundamentals closely, many trades occur at near-zero mispricing, where the strict test cannot count them as corrective.[5]

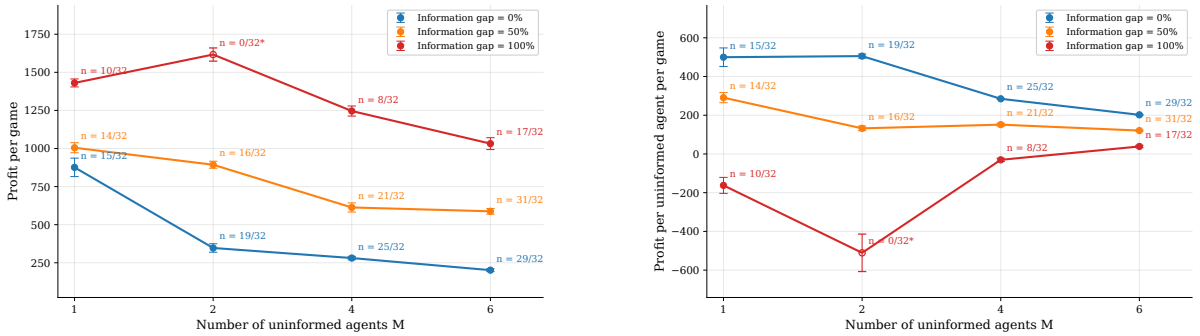
4.6 Rewards and the Incidence of Trading Gains

Finally, we examine who gains and who loses. By Proposition 1, each game is zero-sum between the speculators as a group and the market-making sector, and aggregate speculator profits are capped by the initial mispricing at $\frac{T}{T+1}(\nu - p_0)^2/(2\lambda)$ per game, which is approximately 1,630 per game in expectation in the base configuration, consistent with the realized-sample bounds of Table 3.[6] Aggregate speculator profit therefore measures how much of the feasible extraction from the market-making sector the speculators realize, and *relative* rewards measure the distribution of trading gains across agent types.

Figure 10 and Table 3 contain four results. First, the informed agent's reward rises with the information gap in every market, and it is largest in the most fragile configuration. At a 100% gap with two uninformed agents the informed agent earns 1,617 per game, against 347 in the corresponding all-informed market. Information asymmetry is the informed agent's asset, and manipulation is how that asset is monetized. Second, the incidence table shows where the money comes from. At gaps of 0% and 50%, both speculator sectors, the informed agent and the uninformed group, profit and the market maker funds them. At a 100% gap the picture flips, and the informed agent extracts from the uninformed traders and from the market maker alike. Uninformed agents earn positive rewards wherever their signal retains value, that is, in every cell at

[5] Dropping trades at exactly zero mispricing raises the all-informed cells by up to 5.9 percentage points and raises the cells with a positive information gap by less than 2 percentage points. See the Internet Appendix.

[6] With $\gamma_g \sim \text{Uniform}(-100, 100)$, $\mathbb{E}[(\nu_g - p)^2] = 200^2/12 \approx 3{,}333$ at the start of each game, so the expected per-game bound is $\frac{50}{51} \times 3{,}333/2 \approx 1{,}630$ at $\lambda = 1$, or about 16,300 per ten-game episode. Comparing realized profits against the bound, game by game with each game's realized initial mispricing, is a sharp integrity check on the simulation pipeline, and the ensembles pass it exactly: *no game in any cell of any grid violates the bound* (Table 3), which also validates empirically that the code's execution convention is the post-impact one that Proposition 1 assumes.



(a) Informed agent's reward per game

(b) Uninformed reward per agent per game

Figure 10. Agent rewards across market configurations. Panel (a) shows the informed agent's reward per game and panel (b) the uninformed agents' reward per agent per game, by number of market participants with one curve per information gap, under the absolute NashConv filter ($\epsilon = 3$; means and standard errors across surviving seed runs, survivor counts annotated; the (100%, M=2) point shows the unfiltered ensemble mean). Informed rewards rise with the information gap in every market; uninformed rewards fall with the gap and turn negative at 100%, with losses concentrated where the uninformed side is concentrated.

Table 3. Incidence of Trading Gains and the Aggregate Profit Bound

	Info gap	n	Informed	Uninformed (total)	Market maker	Bound	Violations
$M = 1$	0%	15	876	500	−1,376	1,570	0%
$M = 2$	0%	19	347	1,011	−1,358	1,638	0%
$M = 4$	0%	25	281	1,140	−1,421	1,667	0%
$M = 6$	0%	29	202	1,213	−1,414	1,681	0%
$M = 1$	50%	14	1,005	291	−1,296	1,570	0%
$M = 2$	50%	16	893	265	−1,158	1,638	0%
$M = 4$	50%	21	613	606	−1,219	1,667	0%
$M = 6$	50%	31	588	725	−1,313	1,681	0%
$M = 1$	100%	10	1,430	−162	−1,268	1,570	0%
$M = 2$	100%	32*	1,617	−1,021	−595	1,638	0%
$M = 4$	100%	8	1,246	−121	−1,125	1,667	0%
$M = 6$	100%	17	1,032	232	−1,264	1,681	0%

Notes: Mean profits per game by sector, averaged over games, evaluation seeds, and the n runs surviving the absolute NashConv filter (of 32); *unfiltered (no survivor). The uninformed column reports the combined profit of all M uninformed agents. Dividing it by M gives the per-agent values shown in panel (b) of Figure 10. The three sectors net to zero game by game, as Proposition 1(i) requires. "Bound" is the realized-sample mean of the per-game cap $\frac{T}{T+1}(\nu - p_0)^2/(2\lambda)$ from Proposition 1(iii); "Violations" is the fraction of games in which aggregate speculator profit exceeds its game-specific bound, and it is zero everywhere.

the 0% and 50% gaps. At a 100% gap they lose money in most markets, and their losses concentrate exactly where the uninformed side is concentrated, at −511 per agent with two uninformed traders against +39 with six. Fragmentation protects the manipulated, which is the incidence counterpart

of Hypothesis 2. Third, aggregate speculator profit falls as manipulation grows, exactly as the decomposition (3) predicts. With two uninformed agents, it falls from 1,358 to 1,158 to 595 as the gap rises from 0% to 100%, because the foregone price-discovery gain and the churn cost of round-trip manipulation are both paid to the market-making sector. Manipulation is therefore not a pure transfer among speculators. It shrinks the pie while redistributing the smaller pie toward the informed agent, as Corollary 1 requires. Fourth, no game anywhere violates the aggregate profit bound of Proposition 1, across every cell of every grid. This is the strongest available check that the simulation implements the accounting the theory assumes, and it verifies in particular that orders execute at post-impact prices.

The negative rewards deserve a direct discussion, because they carry a sharp implication for what the extreme cells are. Holding is always available and earns exactly zero, so a policy that loses money on average cannot be a best response, no matter what the other agents do. This is the no-trade test of Section 3.6, a deviation test that requires no learning algorithm at all. We apply it to every cell. Wherever the uninformed agents' average loss exceeds ϵ per step, that is, 150 per game, the profile fails this exact equilibrium test even if it passes the retraining filter. At the ensemble means, this rejects an equilibrium reading of the 100% cells with a concentrated uninformed side. The uninformed agent at (100%, M=1) loses 162 per game, a gain of 3.2 per step from switching to hold, and the two filter-passing runs at (100%, M=2) lose 210 per agent, a gain of 4.2 per step. Every other reported cell passes the test, since uninformed profits there are positive or negligibly negative. A run-level version of the same test, computed from per-run uninformed rewards for every cell, sharpens the statement. Across the information-gap grid, not a single filter-passing run at the 0% or 50% gaps has a negative mean uninformed reward, and every negative-reward run in the grid sits at the 100% gap. The cells with equilibrium support and the cells with uninformed losses are therefore separated run by run, not only on average (Appendix D). We present the extreme cells as what learning produces, not as equilibria, and all equilibrium statements in this paper rest on the cells that pass both tests.

Why, then, do learning agents keep trading at a loss? The answer is not that training failed to converge. The uninformed policies do converge, in the sense that they stop changing, but they converge systematically to a suboptimal interior policy rather than to the optimal corner of not trading. Two structural forces hold them there. Reaching the no-trade policy requires the network to concentrate its action distribution on a narrow interior region, which the exploration bonus used in training actively resists, and the feedback from which the agents must learn that their signal is too noisy to trade on profitably is at its noisiest exactly at the extreme gap. The run-level evidence shows that this suboptimal resting point is systematic rather than an unlucky draw. Every filter-passing run at the extreme gap with a concentrated uninformed side ends in losses, and continued retraining does not rescue the victims. It typically makes them worse, while the informed agent sits near its best response, an asymmetry we document in Appendix D. This pattern is not an artifact to hide but a finding with empirical resonance. Persistent losses by active traders are among the most robust facts in retail finance (Barber and Odean, 2000), and the

logic of Grossman and Stiglitz (1980) says that for traders whose losses exceed the value of their information, the equilibrium action is exactly the non-participation they fail to learn. What stands between real markets and that stark prediction is the learning technology, and the informed agent's strategy profits from precisely that wedge. The exit margin that operates in reality, an operator shutting down a losing algorithm, lies outside the model, and our comparative statics suggest it would not restore efficiency by itself, because exit by some uninformed traders concentrates the remaining uninformed side, which is the configuration manipulation exploits most. Endogenizing participation, and training the uninformed agents adversarially to see whether they can learn their way to safety, are extensions we develop in the next version.

Beyond profits, the filter's survival rates are themselves economically informative. Survival falls with the information gap and rises with the number of uninformed agents (from 91% at (0%, M=6) to its unique zero at (100%, M=2)): approximate equilibrium is rarest exactly where manipulation is strongest. In the (100%, M=2) cell, no run passes the absolute filter and only two of 32 pass the exploitability criterion. Their outcomes (price gap 35.2, informed reward 1,412, uninformed −210 each) sit close to the unfiltered ensemble means, so rarity, not selection, is the message of that cell. The nature of the failures is telling: at the extreme gap, retraining the exploited uninformed agents typically makes them *worse*, not better (Appendix D), so manipulation destabilizes not only prices but the victims' learning itself. We read this as a substantive feature of AI markets under extreme information asymmetry, and we present that cell throughout with the values of its two filter-passing runs quoted and flagged.

4.7 Competition Among Informed Traders

The results so far vary the uninformed side. We now add a second *informed* trader, and we design the comparison to separate the two forces that entry bundles together: competition among informed traders and the total trading capacity of the informed side. We compare three markets at the same information gaps, frozen configuration, and filters ($S = 16$ for the extension ensembles). The *baseline* is the workhorse market, one informed agent of capacity 2 against two uninformed agents of capacity 7. The *split-capacity* market divides the same informed capacity between two rivals, two informed agents of capacity 1 each against the unchanged uninformed side, so competition changes while the informed side's total capacity (2), each uninformed trader's capacity (7), and total market capacity (16) all stay fixed. The *doubled-capacity* market adds a second informed agent of capacity 2 and reduces each uninformed agent to capacity 6, holding total market capacity at 16 but raising the informed side's capacity from 2 to 4. The one element the split-capacity design cannot hold fixed is the number of learners, which rises from three to four.

The comparison of the three designs, baseline, split capacity, and doubled capacity, identifies which margin matters (Table 4). At the 50% gap, where all three ensembles have healthy survivor counts, splitting the informed capacity between two rivals changes essentially nothing about the market: the price gap is 20.1 percent in both the baseline and the split-capacity market, the correction share is 0.80 in both, and the uninformed earn 131 per agent against 134. What com-

Table 4. One Versus Two Informed Traders: Competition and Capacity

	50% information gap			100% information gap		
	Baseline	Split	Doubled	Baseline	Split	Doubled
Informed agents × capacity	1×2	2×1	2×2	1×2	2×1	2×2
Surviving runs	17/32	10/16	12/16	2/32	3/16	1/16
Price gap (%)	20.1 (0.6)	20.1 (0.5)	15.0 (0.8)	35.2	34.8	31.3
Correction share	0.80 (0.01)	0.80 (0.01)	0.85 (0.01)	0.64	0.65	0.68
Informed profit per agent	887 (22)	400 (15)	587 (17)	1,412	726	1,275
Informed profit, total	887	800	1,173	1,412	1,452	2,550
Uninformed profit per agent	134 (12)	131 (13)	−21 (19)	−210	−270	−779
Aggregate speculator profit	1,155	1,062	1,131	991	913	993
Action correlation between informed		0.45	0.52		0.53	0.37
Same side when both trade		0.75	0.80		0.74	0.69
Trading frequency per informed	0.61	0.75 / 0.73	0.57 / 0.56	0.59	0.79 / 0.88	0.87 / 0.86
Mean action intensity per informed	0.49	—	0.44 / 0.44	0.46	—	0.66 / 0.70

Notes: The informed-competition experiment of Section 4.7 under the one-sided filter. Numbers in parentheses are standard errors across surviving runs, where available. Baseline: one informed agent of capacity 2 against two uninformed agents of capacity 7. Split capacity: two informed agents of capacity 1 each against the same uninformed side, holding the informed side's total capacity at 2. Doubled capacity: two informed agents of capacity 2 each against two uninformed agents of capacity 6 each, raising the informed side's capacity to 4. Total market capacity is 16 in all designs. The 100% columns rest on one to three surviving runs, so values are quoted without standard errors and should be read as descriptive. Trading frequency is the share of steps with a non-holding order, and action intensity the mean absolute policy output; the intensity statistic is pending for the split-capacity market.

petition changes is the division of the rents. Each informed agent earns 400, against 887 for the sole informed agent, while the informed side in total collects 800, slightly *below* the monopoly rent rather than above it. The two informed policies do not coordinate. Their action correlation is 0.45, they are on the same side of the market in three quarters of the steps in which both trade, and the pair splits the rent roughly evenly. In contrast to the pricing games of Calvano et al. (2020), where symmetric algorithms learn supracompetitive outcomes, the informed duopoly here earns no premium over the monopoly benchmark. The identical price paths suggest that, with hard per-step capacity limits, the aggregate outcome is pinned down by the informed side's total capacity, and competition operates only on how the rents are shared.

The price improvement in the doubled-capacity market is then a capacity effect, not a competition effect. Raising the informed side's capacity from 2 to 4 units lowers the price gap from 20.1 to 15.0 percent and raises the correction share from 0.80 to 0.85, and it does so while pushing the uninformed from earning 134 per agent to roughly breaking even, because the larger informed sector extracts more in total, 1,173 against 887. The clean statement of this section is therefore a decomposition: prices are governed by how much informed capital the market contains, not by how many competitors hold it. Competition divides the manipulation rents, and neither margin restores the uninformed traders' profits. At the 100% gap the filter-passing samples are too thin for inference, with one to three surviving runs, and the exact no-trade test of Section 4.6 rejects an

equilibrium reading there, so those columns are descriptive only. Directionally, the split-capacity market tracks the baseline there as well (34.8 against 35.2 percent), with the uninformed losses at least as deep. This decomposition is consistent with the market-power results of Section 4.4, where prices respond throughout to the allocation of trading capacity, and with the evidence of Section 6 that the informed agents' decision rules change little when an informed rival enters. Entry changes what each informed agent earns, while the price distortion moves only with the informed side's capacity.

5 Robustness

We verify that our findings are not artifacts of the two parameters that govern the trading environment: the price-impact parameter λ and the presence of exogenous noise traders. The seed-ensemble robustness experiments perturb the workhorse configuration (one informed versus two uninformed agents, 50% information gap) one axis at a time, at the frozen configuration with $S = 16$ seeds and the full filter: $\lambda \in \{0.5, 1, 2, 4\}$ with no noise trader, and a noise trader of capacity $k_{\text{noise}} \in \{0, 1\}$ at $\lambda = 1$. A noise trader is an agent that submits an exogenous random order $\xi_t \cdot k_{\text{noise}}$ at every step, with ξ_t drawn i.i.d. uniformly from $\{-1, 0, +1\}$ and trading capacity k_{noise}. It does not learn and earns no reward. The noise trader's capacity is added on top of the fixed speculator capacity, so the market carries $16 + 1$ units of potential flow while the learning agents' capacities are unchanged, so it does not replace any speculator. Noise trading is conceptually distinct from the information gap. A noise trader adds uninformative order flow, whereas a noisier private signal degrades the uninformed agents' information without adding any flow.

The manipulation metrics are stable along both axes. The price gap is essentially flat in λ (20.3, 20.7, 19.8, 19.7 percent as λ rises from 0.5 to 4) and the correction share edges up (from 0.795 to 0.808). Quadrupling the cost of trading neither eliminates nor inflates the distortion, it mildly taxes it. Reward *levels* do scale with λ, mechanically, because the extraction bound of Proposition 1 is proportional to $1/\lambda$. The manipulation metrics are the invariant objects. Notably, filter survival rises steeply with λ (4, 9, 13, and 16 of 16 seeds). Stronger price impact makes the multi-agent learning problem easier to stabilize, consistent with impact damping the feedback loop between actions and prices. Adding a unit-capacity noise trader leaves the price gap unchanged (20.75 against 20.70) and slightly improves the correction share, while survival again rises (9 to 13 of 16). Uninformative flow mainly provides camouflage that helps learning settle, and by Proposition 1 any additional speculator extraction it permits comes at the noise trader's expense, not the market maker's. Importantly, noise traders remain unnecessary for our mechanism. The base configuration has none, and all uninformative order-flow variation there arises endogenously from the uninformed agents' learning, which distinguishes our bubbles from mechanisms that rely on exogenous noise or feedback demand.

Beyond the trading environment, our results are robust along three further dimensions. Robustness to *equilibrium selection*: the one-sided (exploitability) filter reproduces every survival count

Table 5. Robustness to the Trading Environment: Detailed Results

Perturbation	n	Price gap (SE)	Correction share (SE)
$\lambda = 0.5$	4/16	20.3 (0.3)	0.795 (0.002)
$\lambda = 1$ (base)	9/16	20.7 (0.8)	0.792 (0.009)
$\lambda = 2$	13/16	19.8 (1.1)	0.806 (0.011)
$\lambda = 4$	16/16	19.7 (1.2)	0.808 (0.013)
Noise trader, $k_{\text{noise}} = 1$	13/16	20.7 (0.8)	0.805 (0.008)

Notes: The robustness experiments of Section 5 on the workhorse configuration (one informed of capacity 2 versus two uninformed of capacity 7, 50% information gap), varying one parameter at a time at 16 seeds per cell under the absolute filter. The noise-trader row is at $\lambda = 1$. Standard errors across surviving runs in parentheses.

within two runs and every reported metric within sampling error (Appendix D), and the two-stage tuning procedure (individual tuning, then joint training) delivers the same qualitative results as the simultaneous procedure, including the manipulation patterns of Section 4.2. Robustness to the *training technology*: the legacy hyperparameter sweep, which preceded the frozen-configuration design and pooled up to 128 configurations per cell, produced the same qualitative patterns across the full $\lambda \times k_{\text{noise}}$ product and market compositions up to twelve uninformed agents, and we retain it as an Internet Appendix analysis showing that no single configuration drives the economics. Robustness to *market composition*: manipulation is not specific to the three-agent configuration. It appears across the full grid of 1 to 6 uninformed agents documented above, with severity varying systematically as in Sections 4.3–4.4.

6 Inside the Black Box: Interpreting the Learned Policies

The trading trajectories show *what* the agents do; we now examine *how* they decide. We analyze the learned policy functions using SHAP values (Lundberg and Lee, 2017), which decompose each policy's output into additive contributions of its input features, and we compare feature importance across three market configurations: a single informed agent trading alone, one informed versus two uninformed agents (our core setting), and two informed versus two uninformed agents (the doubled-capacity design of Section 4.7).

Figure 11 confirms Hypothesis 4 and reveals that the informed agent's strategy is environment-contingent in exactly the way the manipulation mechanism predicts. Trading alone, the agent's decisions load on the fundamental value gap, the pure price-correction strategy of the competitive benchmark. Its residual weight on detrended prices is natural even without opponents: the agent's own trades move the price, so the recent price path tracks how much of the initial gap remains to be harvested. Facing uninformed agents, the agent *reduces* its reliance on fundamental information and dramatically increases its use of detrended prices and volume. It is trading on the reactions of the other algorithms, monitoring the momentum that it creates and against which it times its exit.

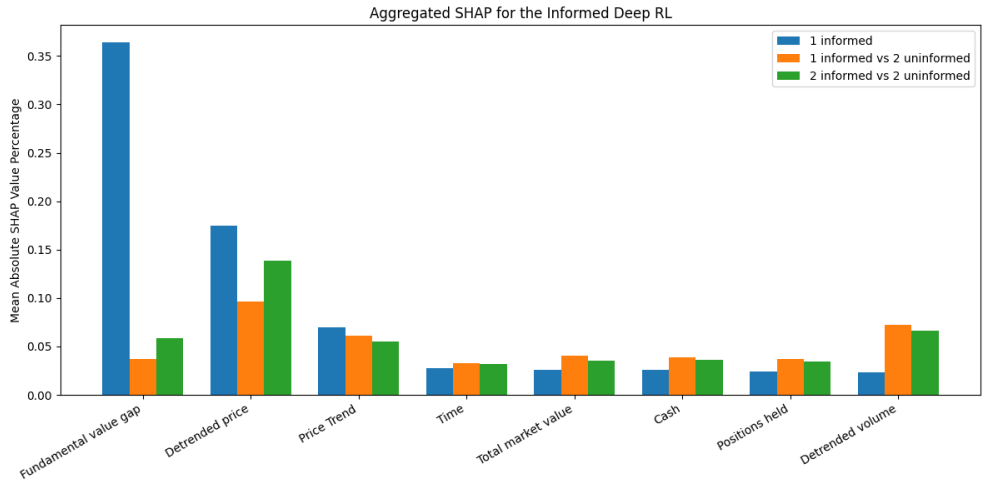


Figure 11. Feature importance for the informed agent. The figure shows mean absolute SHAP values aggregated by feature category for the informed agent's policy across three market configurations. Trading alone, the informed agent relies on the fundamental value gap; facing uninformed agents, it shifts toward market-based signals (detrended price and volume); a second informed rival leaves this market-based profile largely in place, with only a modest rebalancing toward the fundamental gap.

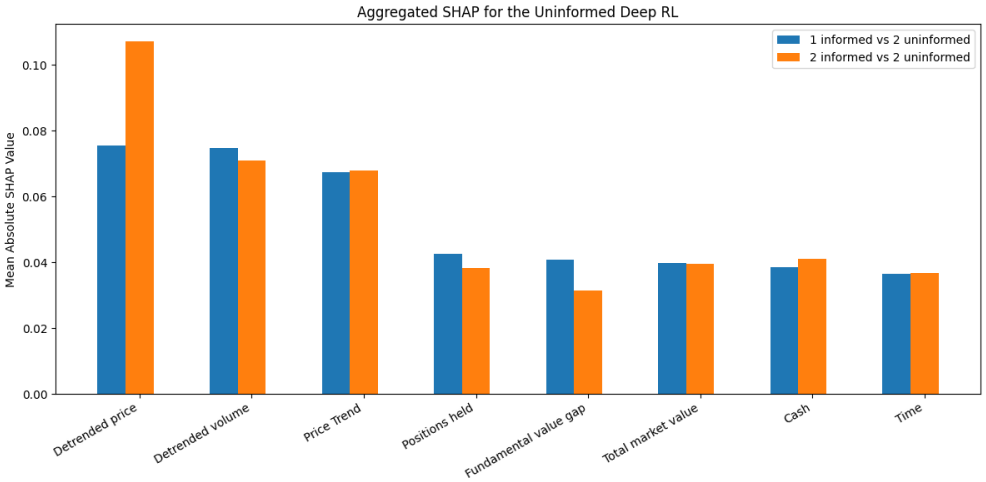


Figure 12. Feature importance for the uninformed agent. The figure shows mean absolute SHAP values aggregated by feature category for the uninformed agent's policy. Uninformed agents rely predominantly on market-based signals (detrended price, volume, trend) and learn to discount their noisy fundamental value signal.

When a second informed agent enters, the decision rule changes far less than when the uninformed entered. The informed agents continue to load primarily on detrended prices and volume, with only a modest rebalancing toward the fundamental gap. This stability of the decision rule is the policy-level counterpart of the market-level finding in Section 4.7: an informed rival changes who earns the manipulation rents, not how the informed side trades, and the better prices of the doubled-capacity

market come from more informed capital executing a largely unchanged strategy rather than from a return to fundamentalist trading.

Figure 12 shows the complementary adaptation of the uninformed agents. Their policies load on detrended prices, volume, and trend components, while placing minimal weight on their noisy fundamental signal. Recall that no agent is told whether its signal is informative. The uninformed agents have *learned*, from trading outcomes alone, that their signal deserves little weight, and they have reallocated their attention to market data. Discounting an unreliable signal in favor of market information is individually rational, since prices aggregate the information of better-informed participants, but it is precisely this rational inference that the informed agent's strategy weaponizes. The two SHAP profiles together provide a mechanism-level confirmation of the conceptual framework. The uninformed agents watch the market, and the informed agent makes sure that what they see misleads them.

7 Conclusion

This paper shows that AI-powered algorithmic trading can destabilize markets through a mechanism that is fundamentally different from algorithmic collusion. Informed AI traders endogenously learn to manipulate the inference of less informed AI traders, generating local price bubbles and persistent deviations of prices from fundamentals. The mechanism requires no coordination (a single informed agent suffices) and operates against victims that are themselves learning algorithms, trained adaptively rather than specified as behavioral rules. Market structure determines its severity: instability is greatest when information asymmetry is large, when a small number of uninformed traders holds substantial market power, and when the division of market power between the informed and uninformed sides is unbalanced in either direction, with the two manipulative strategies dominating the two extremes. Competition among informed traders disciplines the rents, not the prices. With the informed side's total capacity held fixed, a second informed trader splits the manipulation rents without moving the price distortion or the uninformed traders' profits, and distortions fall only when the informed side's trading capacity expands, which pushes the uninformed traders to breaking even. Price efficiency and investor protection therefore respond to different levers, and neither responds to entry alone.

Our findings carry three broader messages. The first concerns the taxonomy of AI-driven market risks. Alongside price-trigger collusion and learning-bias collusion (Dou et al., 2025) and coordinated bubble riding by planning algorithms (Dou et al., 2026a), we identify information manipulation as a distinct channel, one that, unlike the others, needs neither symmetry nor coordination nor exogenous feedback traders, and that emerges precisely when AI agents become heterogeneous and sophisticated. The second concerns the technology debate. The manipulative strategies we document did not emerge under the tabular Q-learning systems studied in prior work, which suggests that the risk landscape shifts qualitatively as deployed trading systems graduate from reactive rules to deep learning systems capable of multi-step strategies. Regulatory frame-

works calibrated to the current generation of algorithms may not be informative about the next. The third message is methodological. Studying "smart" AI traders requires solving reward design and equilibrium selection in non-cooperative multi-agent learning, and the tools we develop, information-set-respecting reward shaping with exact profit equivalence and NashConv-based equilibrium filtering, are portable to other economic applications of multi-agent RL.

The policy discussion should be read as a set of hypotheses that the model generates, to be tested under richer price formation before they guide regulation. With that caveat, the comparative statics point in clear directions. Markets are most vulnerable when information is concentrated and the uninformed side of the market is concentrated too. Measures that broaden access to fundamental information and that promote a large number of competing algorithmic traders directly attack the conditions manipulation requires. Because the strategies emerge endogenously inside black-box systems, they cannot be screened by inspecting code for manipulative intent. Simulation-based oversight, which stress-tests algorithms in controlled multi-agent environments before deployment as our framework illustrates, offers a more realistic supervisory model. Several extensions merit attention, including richer microstructure with limit order books, multiple assets, and an adaptive market-making sector, welfare accounting that includes the market maker's losses, the training dynamics through which manipulation is first discovered, endogenous participation decisions, and the arms race that may ensue as uninformed algorithms are redesigned to resist manipulation. We view these as natural next steps for a research agenda on the stability of markets in which the marginal trader is no longer human.

References

Abreu, Dilip, and Markus K Brunnermeier, 2003, Bubbles and crashes, *Econometrica* 71, 173–204.

Abreu, Dilip, David Pearce, and Ennio Stacchetti, 1986, Extremal equilibria of oligopolistic supergames, *Journal of Economic Theory* 39, 191–225.

Aggarwal, Rajesh K, and Guojun Wu, 2006, Stock market manipulations, *Journal of Business* 79, 1915–1953.

Allen, Franklin, and Douglas Gale, 1992, Stock-price manipulation, *Review of Financial Studies* 5, 503–529.

Almgren, Robert, and Neil Chriss, 2001, Optimal execution of portfolio transactions, *Journal of Risk* 3, 5–39.

Arthur, W. Brian, John H. Holland, Blake LeBaron, Richard Palmer, and Paul Tayler, 1997, Asset pricing under endogenous expectations in an artificial stock market, in W. Brian Arthur, Steven N. Durlauf, and David A. Lane, eds., *The Economy as an Evolving Complex System II*, 15–44 (Addison-Wesley).

Asker, John, Chaim Fershtman, and Ariel Pakes, 2022, Artificial intelligence, algorithm design, and pricing, in *AEA Papers and Proceedings*, volume 112, 452–56.

Assad, Stephanie, Robert Clark, Daniel Ershov, and Lei Xu, 2024, Algorithmic pricing and competition: Empirical evidence from the German retail gasoline market, *Journal of Political Economy* 132, 723–771.

Barber, Brad M, and Terrance Odean, 2000, Trading is hazardous to your wealth: The common stock investment performance of individual investors, *Journal of Finance* 55, 773–806.

Bertsimas, Dimitris, and Andrew W Lo, 1998, Optimal control of execution costs, *Journal of Financial Markets* 1, 1–50.

Blanchard, Olivier Jean, 1979, Speculative bubbles, crashes and rational expectations, *Economics letters* 3, 387–389.

Brown, Zach Y, and Alexander MacKay, 2023, Competition in pricing algorithms, *American Economic Journal: Microeconomics* 15, 109–156.

Brunnermeier, Markus K, and Lasse Heje Pedersen, 2005, Predatory trading, *Journal of Finance* 60, 1825–1863.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolo, and Sergio Pastorello, 2020, Artificial intelligence, algorithmic pricing, and collusion, *American Economic Review* 110, 3267–3297.

Camerer, Colin, 1989, Bubbles and fads in asset prices, *Journal of Economic Surveys* 3, 3–41.

Colliard, Jean-Edouard, Thierry Foucault, and Stefano Lovo, 2026, Algorithmic pricing and liquidity in securities markets, *Review of Financial Studies, forthcoming* .

De Long, J. Bradford, Andrei Shleifer, Lawrence H. Summers, and Robert J. Waldmann, 1990, Positive feedback investment strategies and destabilizing rational speculation, *Journal of Finance* 45, 379–395.

Dou, Winston Wei, Itay Goldstein, and Yan Ji, 2025, AI-powered trading, algorithmic collusion, and price efficiency, *Working Paper No. 34054, National Bureau of Economic Research* .

Dou, Winston Wei, Itay Goldstein, and Yan Ji, 2026a, Financial market fragility in the era of AI planning, *Working Paper, University of Pennsylvania* .

Dou, Winston Wei, Itay Goldstein, Zigang Li, and Liyan Yang, 2026b, Human edge, machine limits: Algorithmic herding and AI-human competition, *Working Paper, University of Pennsylvania* .

Fershtman, Chaim, and Ariel Pakes, 2012, Dynamic games with asymmetric information: A framework for empirical work, *The Quarterly Journal of Economics* 127, 1611–1661.

Goldstein, Itay, and Alexander Guembel, 2008, Manipulation and the allocational role of prices, *Review of Economic Studies* 75, 133–164.

Goldstein, Itay, Chong Huang, and Liyan Yang, 2025, Fragility of financial markets, *Annual Review of Financial Economics* 17, 27–48.

Green, Edward J, and Robert H Porter, 1984, Noncooperative collusion under imperfect price information, *Econometrica* 52, 87–100.

Grossman, Sanford J., and Joseph E. Stiglitz, 1980, On the impossibility of informationally efficient markets, *American Economic Review* 70, 393–408.

Harras, Georges, and Didier Sornette, 2011, How to grow a bubble: A model of myopic adapting agents, *Journal of Economic Behavior & Organization* 80, 137–152.

Harrison, J Michael, and David M Kreps, 1978, Speculative investor behavior in a stock market with heterogeneous expectations, *The Quarterly Journal of Economics* 92, 323–336.

International Monetary Fund, 2024, Advances in artificial intelligence: Implications for capital market activities, in *Global Financial Stability Report, October 2024: Steadying the Course: Uncertainty, Artificial Intelligence, and Financial Stability*, chapter 3 (International Monetary Fund, Washington, DC).

Jarrow, Robert A, 1992, Market manipulation, bubbles, corners, and short squeezes, *Journal of Financial and Quantitative Analysis* 27, 311–336.

Jarrow, Robert A, 2015, Asset price bubbles, *Annual Review of Financial Economics* 7, 201–218.

Johanson, Michael Bradley, Edward Hughes, Finbarr Timbers, and Joel Z. Leibo, 2022, Emergent bartering behaviour in multi-agent reinforcement learning, arXiv preprint arXiv:2205.06760.

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

Kyle, Albert S, 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315–1335.

Kyle, Albert S., 1989, Informed speculation with imperfect competition, *Review of Economic Studies* 56, 317–355.

Kyle, Albert S, and S Viswanathan, 2008, How to define illegal price manipulation, *American Economic Review* 98, 274–279.

LeBaron, Blake, 2006, Agent-based computational finance, in Leigh Tesfatsion, and Kenneth L. Judd, eds., *Handbook of Computational Economics*, volume 2, 1187–1233 (Elsevier).

LeBaron, Blake, W. Brian Arthur, and Richard Palmer, 1999, Time series properties of an artificial stock market, *Journal of Economic Dynamics and Control* 23, 1487–1516.

Lundberg, Scott M, and Su-In Lee, 2017, A unified approach to interpreting model predictions, *Advances in Neural Information Processing Systems* 30.

Ng, Andrew Y, Daishi Harada, and Stuart Russell, 1999, Policy invariance under reward transformations: Theory and application to reward shaping, in *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, 278–287.

Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov, 2017, Proximal policy optimization algorithms, *arXiv preprint arXiv:1707.06347* .

Silver, David, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al., 2016, Mastering the game of Go with deep neural networks and tree search, *Nature* 529, 484–489.

Sutton, Richard S, and Andrew G Barto, 2018, *Reinforcement Learning: An Introduction*, second edition (MIT Press).

Tampuu, Ardi, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, and Raul Vicente, 2017, Multiagent cooperation and competition with deep reinforcement learning, *PloS one* 12, e0172395.

Tangpi, Ludovic, and Shichun Wang, 2022, Optimal bubble riding: A mean field game with varying entry times, *arXiv preprint arXiv:2209.04001* .

Tangpi, Ludovic, and Shichun Wang, 2024, Optimal bubble riding with price-dependent entry: a mean field game of controls with common noise, *Mathematics and Financial Economics* 1–38.

Timbers, Finbarr, Nolan Bard, Edward Lockhart, Marc Lanctot, Martin Schmid, Neil Burch, Julian Schrittwieser, Thomas Hubert, and Michael Bowling, 2022, Approximate exploitability: Learning a best response in large games, arXiv preprint arXiv:2004.09677.

U.S. Securities and Exchange Commission, 2020, Staff report on algorithmic trading in U.S. capital markets, Report to Congress, August 2020.

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

Westphal, Rebecca, and Didier Sornette, 2020, Market impact and performance of arbitrageurs of financial bubbles in an agent-based model, *Journal of Economic Behavior & Organization* 171, 1–23.

Appendix A Reward Shaping: Proof and Discussion

Proof of Proposition 2. Part (i). Fix a sample path. Summing the shaped rewards over $t = 1, \ldots, T$,

$$\begin{aligned}\sum_{t=1}^{T-1} \mathcal{R}_t + \mathcal{R}_T &= \sum_{t=1}^{T} x_t \cdot (\nu_{\text{obs}} - p_t) + (\nu - \nu_{\text{obs}}) \cdot X_T \\ &= \nu_{\text{obs}} \sum_{t=1}^{T} x_t - \sum_{t=1}^{T} x_t p_t + (\nu - \nu_{\text{obs}}) \cdot X_T \\ &= -\sum_{t=1}^{T} x_t p_t + \nu \cdot X_T, \end{aligned} \tag{10}$$

using $X_T = \sum_{t=1}^{T} x_t$. Each order x_t executes at the post-impact price p_t, so the cash balance evolves as $c_t = c_{t-1} - x_t p_t$, and the forced liquidation of the accumulated position X_T at the revealed value ν adds νX_T. Hence $c_T = c_0 - \sum_{t=1}^{T} x_t p_t + \nu X_T$, and $c_T - c_0$ equals expression (10).

Part (ii). Fix any profile of policies for all agents. Part (i) holds on every sample path of the induced stochastic process, so taking expectations under the path distribution induced by the profile, each agent's expected total undiscounted shaped reward equals its expected trading profit, for *every* profile. The two games therefore assign identical expected payoffs to every profile of policies, so they have identical payoff functions, hence identical best-response correspondences, and hence identical sets of exact and ϵ-Nash equilibria.

Part (iii). At a non-terminal step t, $\mathcal{R}_t = x_t(\nu_{\text{obs}} - p_t)$ is a function of the agent's own executed order x_t, its own private signal ν_{obs}, and the public price p_t, each of which is contained in the agent's information set at t. The true value ν enters only through $\mathcal{R}_T$ via the correction term $(\nu - \nu_{\text{obs}})X_T$, which is paid at the revelation event when ν becomes public. □

The shaped reward $x_t(\nu_{\text{obs}} - p_t)$ is the instantaneous gain of the step-t trade evaluated at the agent's best estimate of fundamental value, and the terminal correction $(\nu - \nu_{\text{obs}})X_T$ trues up the estimate when the fundamental is revealed. For the informed agent, $\nu_{\text{obs}} = \nu$ and the step rewards are exact. For uninformed agents, the step rewards are noisy, but crucially the noise is the agent's *own* signal noise. In our configurations, the signal noise does not exceed the dispersion of fundamental innovations, so the shaped reward remains a useful learning signal for uninformed agents. We emphasize what part (ii) does and does not assert. The equivalence is between the two *reward specifications* given undiscounted objectives. It does not assert that training *finds* the same policies under both specifications. Indeed the entire point of shaping is that the dense specification makes learning feasible. The proposition guarantees that whatever policies training finds are evaluated, and their equilibrium status judged, under the economically correct objective.

Appendix B Aggregate Accounting: Proofs

Throughout, fix a single game with T steps, initial price p_0, fundamental value ν fixed within the game and revealed after step T, aggregate speculator order flow $z_t = \sum_i x_{i,t}$, cumulative flow $Z_t = \sum_{s=1}^{t} z_s$ with $Z_0 = 0$, price evolution $p_t = p_{t-1} + \lambda z_t$ from equation (2), and execution of all step-t orders at the post-impact price p_t. There is no exogenous noise-trader flow, so the market maker absorbs exactly $-z_t$ at each step.

Proof of Proposition 1. Part (i). As in the proof of Proposition 2, speculator i's realized profit is $\pi_i = \sum_{t=1}^{T} x_{i,t}(\nu - p_t)$, so summing over speculators,

$$\Pi = \sum_i \pi_i = \sum_{t=1}^{T} z_t(\nu - p_t). \tag{11}$$

The market maker takes the position $-z_t$ at price p_t at each step, receiving cash $p_t z_t$, and its accumulated inventory $-Z_T$ settles at the revealed value ν. Its realized profit is therefore $\Pi^{MM} = \sum_{t=1}^{T} p_t z_t - \nu Z_T = -\sum_{t=1}^{T} z_t(\nu - p_t) = -\Pi$.

Part (ii). Substituting $p_t = p_0 + \lambda Z_t$ into (11),

$$\Pi = (\nu - p_0)Z_T - \lambda \sum_{t=1}^{T} z_t Z_t.$$

Since $z_t = Z_t - Z_{t-1}$, the identity $Z_t^2 - Z_{t-1}^2 = 2Z_t z_t - z_t^2$ summed over t gives $\sum_{t=1}^{T} z_t Z_t = \frac{1}{2}Z_T^2 + \frac{1}{2}\sum_{t=1}^{T} z_t^2$, and (3) follows.

Part (iii). By the Cauchy-Schwarz inequality, $\sum_{t=1}^{T} z_t^2 \geq Z_T^2/T$, with equality if and only if z_t is constant across t. Substituting into (3),

$$\Pi \leq (\nu - p_0)Z_T - \frac{\lambda}{2}\frac{T+1}{T} Z_T^2,$$

a strictly concave quadratic in Z_T maximized at $Z_T^* = \frac{T}{T+1} \cdot \frac{\nu - p_0}{\lambda}$, with maximal value $\frac{T}{T+1} \cdot \frac{(\nu-p_0)^2}{2\lambda}$. Both inequalities bind simultaneously if and only if $z_t = Z_T^*/T = \frac{\nu - p_0}{\lambda(T+1)}$ for all t. Under this flow, $p_T = p_0 + \lambda Z_T^* = p_0 + \frac{T}{T+1}(\nu - p_0)$, so $\nu - p_T = \frac{\nu - p_0}{T+1}$. □

Proof of Corollary 1. For the first claim, suppose speculator i earns $\pi_i > \frac{T}{T+1}(\nu - p_0)^2/(2\lambda)$. By Proposition 1(iii), $\sum_j \pi_j \leq \frac{T}{T+1}(\nu - p_0)^2/(2\lambda)$, so $\sum_{j \neq i} \pi_j \leq \frac{T}{T+1}(\nu - p_0)^2/(2\lambda) - \pi_i < 0$: the excess is matched, one for one, by aggregate losses of the other speculators. The second claim is immediate from the decomposition (3): given Z_T, the price-discovery and repositioning terms are fixed, and Π varies only through $-\frac{\lambda}{2}\sum_t z_t^2$. By part (i), every reduction in Π accrues to the market maker as $\Pi^{MM} = -\Pi$. □

Four remarks. First, the results hold path by path and for arbitrary, possibly non-Markovian and coordinated, strategies. They are accounting identities of the price-formation rule, not equilibrium

statements. Second, the bound concerns the speculator *sector*. Individual profits are unbounded above, but only at other speculators' expense, which is what makes the bound the right benchmark for detecting exploitation. Third, offsetting trades by different speculators *within* the same step cancel in the aggregate flow z_t, have no price impact, and constitute a zero-sum transfer between the speculators involved. The churn term prices only the time variation of the aggregate flow. Fourth, with exogenous noise traders, the same argument applied to the union of speculators and noise traders shows that learning agents can collectively exceed the bound exactly by the noise traders' trading losses.

Appendix C Learning Algorithm: Details

Each agent is trained with Proximal Policy Optimization (Schulman et al., 2017). PPO is an on-policy policy-gradient method. Experience is collected under the current policy and used for a constrained policy update, after which the data are discarded. The update maximizes the clipped surrogate objective

$$L^{\mathrm{CLIP}}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta)\hat{A}_t,\ \mathrm{clip}(r_t(\theta), 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}})\hat{A}_t \right) \right], \tag{12}$$

where $r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{\mathrm{old}}}(a_t|s_t)$ is the likelihood ratio between the updated and previous policies, $\hat{A}_t$ is the estimated advantage of action a_t in state s_t, and ϵ_{clip} is the clipping parameter bounding the effective policy change per update (we reserve ϵ for the equilibrium threshold of Appendix D). Clipping prevents destructive policy updates, which is essential in our non-stationary multi-agent environment. Continuous actions are parameterized by a Beta distribution, which respects the bounded action space $[-1, 1]$ without boundary bias. Each agent maintains an independent policy and value network. Agents share no parameters, no experience, and no rewards. The discount factor is set to one. Episodes are finite (500 steps), so the undiscounted return is well defined, each agent maximizes its expected realized trading profit per episode, and the reward equivalence of Proposition 2 holds exactly. Training uses the RLlib implementation of PPO (version 2.0, PyTorch). All hyperparameters (learning rate, entropy coefficient, and related quantities) are frozen at the documented global configuration described in Section 4.1: the geometric mean of the twelve per-cell convergence winners of the initial random search, selected before any ensemble outcome existed. The frozen values are a learning rate of 3.4×10^{-4} and an entropy coefficient of 1.7×10^{-4} for the informed role, and 4.1×10^{-4} and 1.9×10^{-3} for the uninformed role; the twelve per-cell winners they aggregate, and the calibration of the equilibrium threshold $\epsilon = 3$ relative to per-cell informed rewards (from about 9% to 73% of the mean informed per-step reward), are tabulated in the Internet Appendix. The role-level differences in these frozen values are outcomes of the outcome-blind averaging rule applied to the per-role search winners, not design choices. Each agent uses an independent five-layer perceptron of width 32 with ReLU activations for both the policy and the value function, a Beta policy distribution on $[-1, 1]$, clipping parameter $\epsilon_{\mathrm{clip}} = 0.35$, training batches of 550 steps with three optimization epochs per update, and 500 training iterations.

A single master seed controls all training randomness (network initialization, minibatch sampling, exploration, and training-environment draws), which is what makes runs that differ only in the seed exchangeable and the ensemble inference of Section 4.1 valid. Two further implementation guarantees are documented in the code repository. Every observation feature is computed from trailing data only, verified by an automated no-look-ahead test (Section 3.1), and the submitted results correspond to a tagged repository snapshot that freezes code, configurations, seed lists, and hyperparameter logs.

Appendix D NashConv Equilibrium Selection: Details

For a profile of policies $\pi = (\pi_1, \ldots, \pi_n)$, agent i's incentive to deviate is

$$\delta_i(\pi) = u_i\big(b(\pi_{-i}), \pi_{-i}\big) - u_i(\pi), \tag{13}$$

where u_i is agent i's expected reward and $b(\pi_{-i})$ is a best response to the other agents' policies. NashConv (Timbers et al., 2022) aggregates deviation incentives, $\text{NashConv}(\pi) = \sum_i \delta_i(\pi)$, and a profile is an ϵ-Nash equilibrium if $\max_i \delta_i(\pi) < \epsilon$. Exact best responses are not computable in our environment, so we approximate δ_i by continued training. Starting from the jointly trained policies at iteration τ_{stop}, we freeze all agents except i and continue training agent i alone until $\tau_{\text{stop}} + \tau_{\text{extra}}$, recording its reward improvement.

The full pipeline for each market configuration (one grid point: an information gap, a number of uninformed agents, and a market-power split) is:

(i) **Launch:** train S independent runs (32 for the information-gap grid, 16 for the extension grids) at the frozen configuration of Section 4.1, differing only in the master random seed, until iteration τ_{stop}.

(ii) **Completion screen:** discard runs that do not reach a minimum training length (100 PPO iterations), removing crashed or prematurely stopped runs. In the realized ensembles no run was lost at this step.

(iii) **Out-of-sample evaluation:** evaluate each policy profile with exploration turned off on the 10 common evaluation seeds and record each agent's mean per-step reward R_i^{orig}.

(iv) **Deviation estimation:** for each run and each agent i, freeze all other agents and continue training agent i alone, and record the retrained mean per-step reward R_i^{retrain}.

(v) **ϵ-Nash filter:** retain run τ under the *absolute* (primary) criterion if and only if

$$\frac{1}{N}\sum_{i=1}^{N} \left| R_i^{\text{retrain}}(\tau) - R_i^{\text{orig}}(\tau) \right| < \epsilon, \tag{14}$$

where N is the total number of agents, and under the *one-sided* (exploitability) criterion if and only if $\frac{1}{N}\sum_i \max\left\{0, R_i^{\text{retrain}}(\tau) - R_i^{\text{orig}}(\tau)\right\} < \epsilon$. All results use $\epsilon = 3$; Table 6 reports survival under both criteria, and Table 7 the sensitivity of survival to ϵ.

(vi) **Aggregation:** compute run-level means across the common evaluation seeds and report means and standard errors across surviving runs, as defined in Section 4.1.

Three implementation notes. First, the units of ϵ: rewards R_i are mean *per-step* evaluation rewards in model currency, so $\epsilon = 3$ corresponds to 150 currency units per 50-step game and roughly 1,500 per 500-step episode. Its calibration relative to each configuration's informed reward is tabulated in the Internet Appendix. Second, no negative-reward pre-filter is applied to any result in this paper: runs proceed to the ϵ-Nash filter after the completeness check only, and the ϵ-Nash filter is the only equilibrium-selection step. Third, the full deviation retraining is expensive, roughly N times the training cost per run, and we verified that it cannot be safely shortened. In a pre-registered validation on completed cells, halving the retraining budget flipped 30% of pass/fail decisions and subsampling the retrained agents flipped 13%, both failing our 90% agreement bar, so we kept the exact filter and reduced the ensemble size to $S = 16$ on the extension grids instead. The following result records precisely what the filter (14) certifies.

Lemma 1 (What the ϵ-Nash filter certifies)**.** *Let $\hat{\Delta}_i(\tau) = R_i^{\text{retrain}}(\tau) - R_i^{\text{orig}}(\tau)$ denote the estimated deviation gain, and abstract from evaluation noise, so that $R_i^{\text{orig}} = u_i(\pi)$ and $R_i^{\text{retrain}} = u_i(\pi_i', \pi_{-i})$ for the retrained policy π_i'. Then:*

(i) $\hat{\Delta}_i(\tau) \leq \delta_i(\pi)$ for every agent: retraining yields a one-sided (lower-bound) estimate of the true deviation incentive (13).

(ii) If trial τ passes the filter (14), then its estimated NashConv satisfies $\sum_i \max\{\hat{\Delta}_i(\tau), 0\} < N\epsilon$, and no single agent has an estimated deviation gain above $N\epsilon$.

Proof. (i) The retrained policy π_i' lies in agent i's policy space, so $u_i(\pi_i', \pi_{-i}) \leq u_i(b(\pi_{-i}), \pi_{-i})$ by definition of the best response, and subtracting $u_i(\pi)$ gives the claim. (ii) Since $\max\{x, 0\} \leq |x|$, $\sum_i \max\{\hat{\Delta}_i, 0\} \leq \sum_i |\hat{\Delta}_i| < N\epsilon$ by (14); the same chain bounds each individual term. □

Lemma 1 makes the epistemic status of the selection transparent. Because the estimate is one-sided, passing the filter can never prove proximity to an exact Nash equilibrium against arbitrary deviations. It establishes that *the same learning technology the agents use* found no unilateral deviation worth more than $N\epsilon$, so passing means an equilibrium reading is not rejected by this search, and failing means it is rejected. To catch simple deviations that a warm-started learned search can miss, we complement the filter with deviations that can be evaluated exactly and without any learning. The most important is the no-trade policy, which always earns exactly zero, so any agent with average reward $\bar{R}_i < 0$ has an exact deviation gain of $|\bar{R}_i|$ per step. Applied at the ensemble means, this exact test rejects an equilibrium reading of the 100% cells with $M \leq 2$, where uninformed losses of 162 and 210 per game imply hold gains of 3.2 and 4.2 per step, both above $\epsilon = 3$, and it does not bind in any other reported cell of the information-gap grid (Section 4.6). A run-level version of the test, computed from the per-run uninformed rewards of every cell, confirms and sharpens this picture. Across the information-gap grid, every filter-passing run with a negative mean uninformed reward sits at the 100% gap (20 of the 212 one-sided survivors), and not one

sits at the 0% or 50% gaps, so the cells with equilibrium support and the cells with losses are separated run by run. In the market-power grids, a minority of runs in the intermediate-share cells show mildly negative uninformed means. At cell means these all stay within the ϵ tolerance, with one exception. In the (M=1, 50% share) cell of the two-dimensional grid the single uninformed agent's average loss exceeds the threshold in every surviving run, and one run in the neighboring 68.75% cell also fails the run-level test. Table 10 flags that cell, and we read the $M = 1$ row of that grid as a description of what learning produces rather than as filtered equilibrium evidence. A broader library of exact benchmark deviations is planned for the next version. We regard this as the economically appropriate equilibrium concept for markets of algorithmic traders, in the spirit of the experience-based equilibrium of Fershtman and Pakes (2012): a trading firm abandons its current algorithm only if its own retraining technology finds a profitably different one. The absolute value in (14) additionally discards runs in which retraining makes agents substantially *worse* on average, a symptom of unstable training, so the primary filter combines an equilibrium criterion with a convergence screen, while the one-sided criterion isolates the pure exploitability notion of Lemma 1(ii). Empirically the two agree almost everywhere. Across the twelve information-gap cells they select 205 and 212 of 384 runs respectively, and every reported metric moves within sampling error. The one cell where they differ is diagnostic. At (100%, M=2), the best run fails the absolute filter at 3.73 per agent while passing the one-sided filter at 0.93, because its deviation profile is $+2.80$ for the informed agent (near its best response) and -5.89 and -2.49 for the two uninformed agents: no agent can gain much by deviating, but retraining the exploited agents makes them *worse*. Manipulation at the extreme gap destabilizes the victims' learning, which the absolute filter reads as non-convergence and the one-sided filter reads as equilibrium. Both readings are correct about their own object, which is why we report both.

Both reported criteria average deviation gains over agents. As a stricter sensitivity check, we also recomputed survival under a worst-case per-agent criterion, which retains a run only if $\max_i \max\{0, R_i^{\text{retrain}} - R_i^{\text{orig}}\} < \epsilon$, so that no single agent, rather than the average agent, can gain more than ϵ. At the same threshold this criterion is substantially stricter, and survivors fall to roughly one third of the one-sided counts (for example, 6 instead of 17 at the workhorse cell). The economic patterns are unchanged: the information-gap ordering and the hump at $M = 2$ persist (price gap 19.9 at $M = 2$ against 17.0 at $M = 1$ and 16.1 at $M = 4$ under the worst-case criterion), and so does the market-power U-shape (20.3, 18.3, 16.1, 17.7, 18.8, 20.5, and 21.5 percent across the seven capacity shares), with wider standard errors reflecting the smaller samples. We keep the mean-based filters as primary because they double as convergence screens and preserve usable ensemble sizes, and the worst-case results will be tabulated cell by cell in the Internet Appendix.

Table 6 reports survival for the full information-gap grid under both criteria, and Table 7 the sensitivity of survival to the threshold in the workhorse column. Survival falls with the information gap and rises with the number of uninformed agents, reaching its unique zero (under the absolute criterion) at the (100%, M=2) cell. Section 4.6 discusses why this pattern is an economic finding rather than a nuisance.

Table 6. NashConv Filtering: Survival Across the Information-Gap Grid

	Absolute filter		One-sided filter	
Cell	Survivors	Fraction	Survivors	Fraction
$M = 1$, 0%	15	47%	15	47%
$M = 2$, 0%	19	59%	19	59%
$M = 4$, 0%	25	78%	25	78%
$M = 6$, 0%	29	91%	29	91%
$M = 1$, 50%	14	44%	15	47%
$M = 2$, 50%	16	50%	17	53%
$M = 4$, 50%	21	66%	22	69%
$M = 6$, 50%	31	97%	31	97%
$M = 1$, 100%	10	31%	10	31%
$M = 2$, 100%	0	0%	2	6%
$M = 4$, 100%	8	25%	9	28%
$M = 6$, 100%	17	53%	18	56%
Total	205	53%	212	55%

Notes: Survivors out of $S = 32$ seed runs per cell at $\epsilon = 3$; all 32×12 runs completed training and deviation retraining (no attrition). The absolute filter is the primary criterion, equation (14); the one-sided filter counts only positive retraining gains. The extension grids (market power, informed competition, robustness) use $S = 16$ and their survivor counts are annotated on the corresponding figures.

Table 7. Sensitivity of Survival to the Threshold ϵ ($M = 2$ Column)

Cell	$\epsilon = 3$	$\epsilon = 4$	$\epsilon = 5$	$\epsilon = 6$	$\epsilon = 8$	$\epsilon = 10$	$\epsilon = 15$
$M = 2$, 0%	19	25	27	29	32	32	32
$M = 2$, 50%	16	23	28	29	32	32	32
$M = 2$, 100%	0	1	3	8	15	22	26

Notes: Survivors out of 32 under the absolute filter as the threshold varies. The 0% and 50% cells pass fully by $\epsilon = 8$. In the 100% cell no threshold below 4 admits any run, so approximate equilibrium is genuinely scarce there rather than marginally missed (see Table 8).

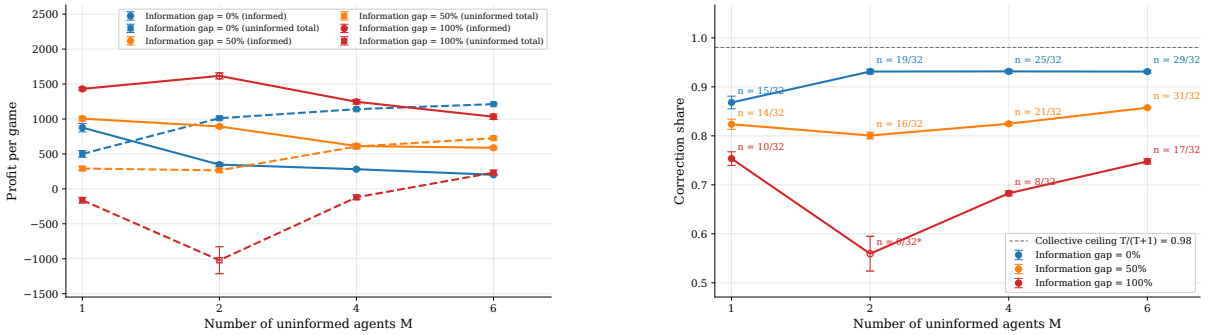
Appendix E Additional Results

Figure 13 complements Section 4.6. Panel (a) shows the incidence patterns of Table 3 graphically. The uninformed sector's profits fall in the information gap everywhere, turn negative at the 100% gap, and are most negative where the uninformed side is most concentrated. Panel (b) reports the outcome-based counterpart of the trade-direction ratio of Section 4.5, namely the share of the initial mispricing that the market closes by the end of a game. All-informed markets reach 0.93, five points below the collectively optimal ceiling of Proposition 1(iii). At a 100% gap the share falls to between 0.56 and 0.75, and it is lowest exactly at the most fragile composition. The two correction

Table 8. Distribution of Per-Agent NashConv Across Seed Runs ($M = 2$ Column)

Cell	Min	P25	Median	P75	Max	Mean
$M = 2$, 0%	0.30	1.22	2.21	3.72	7.16	2.72
$M = 2$, 50%	0.56	1.91	2.95	4.09	7.30	3.20
$M = 2$, 100%	3.73	6.14	8.61	12.29	122.90	13.75

Notes: Distribution of the absolute per-agent NashConv statistic of equation (14) across the 32 seed runs of each cell. The 0% and 50% distributions are centered near the threshold of 3 with thin tails. At the 100% gap even the best run misses the threshold, the median is nearly three times the threshold, and one run diverges, which quantifies how far from stable equilibrium the learning dynamics operate under extreme information asymmetry.



(a) Incidence by sector

(b) Price-based correction share

Figure 13. Additional incidence and efficiency results. Panel (a) plots the incidence legs of Table 3 (informed and total uninformed profits per game) across market compositions and information gaps. Panel (b) plots the price-based correction share, the fraction $1-\sum_g |\nu_g - p_{\text{end},g}| / \sum_g |\nu_g - p_{0,g}|$ of the initial mispricing closed by game end, against its collectively optimal ceiling $T/(T+1) = 0.98$ from Proposition 1(iii) (dashed). Means across surviving seed runs under the absolute filter; survivor counts annotated.

measures are deliberately distinct. The trade-direction ratio scores intent, the direction of each trade, while the correction share scores outcomes, how much mispricing survives, and Section 4.4 shows why both are needed, since a dominant informed agent scores perfectly on direction while under-correcting in magnitude.

Table 9 reports the detailed results behind Section 4.4.

Table 10 reports the full two-dimensional market-power grid behind Figure 8, crossing the informed agent's capacity share with the number of uninformed agents at the 50% information gap.

Appendix F Price Process for the Illustrative Single-Agent Game

In the illustrative game of Section 3.5, the single learning agent trades against rule-based agents whose behavior is fixed. These rule-based agents are motivated as block traders in a Kyle-type

Table 9. Market Power of the Informed Agent: Detailed Results

Capacity	Share	n	Price gap (SE)	Correction share	Informed profit	Uninformed profit
2	12.5%	9	20.7 (0.8)	0.79	898	116
4	25.0%	7	18.5 (0.8)	0.82	1,179	−10
6	37.5%	13	17.5 (0.5)	0.83	1,076	39
8	50.0%	10	17.5 (0.7)	0.83	992	75
10	62.5%	5	19.8 (0.9)	0.80	768	197
12	75.0%	8	21.0 (0.8)	0.79	760	244
14	87.5%	13	22.0 (0.3)	0.78	794	245

Notes: The market-power experiment of Section 4.4 (two uninformed agents, 50% information gap, total capacity 16, 16-seed ensembles, absolute filter). "Share" is the informed agent's fraction of total capacity, n the number of surviving runs, and profits are per game (informed agent; uninformed per agent). Standard errors of the price gap across surviving runs in parentheses.

Table 10. Market Power Across Market Sizes: The Full Grid

	Informed capacity share				
	12.5%	31.25%	50%	68.75%	87.5%
$M = 1$	19.6 (2.3) [4]	16.2 (1.8) [6]	20.6 (1.9) [5]†	24.5 (2.1) [6]	27.8 (0.9) [7]
$M = 2$	20.7 (0.8) [9]	18.0 (0.6) [8]	17.5 (0.7) [10]	22.0 (0.9) [4]	22.0 (0.3) [13]
$M = 4$	17.8 (0.5) [9]	14.4 (0.5) [13]	16.5 (0.7) [9]	20.1 (0.9) [10]	21.6 (0.6) [16]
$M = 6$	14.6 (0.3) [15]	12.5 (0.5) [10]	15.2 (0.4) [15]	16.9 (0.5) [13]	20.8 (1.0) [16]

Notes: Price gap in percent of the typical initial mispricing, at a 50% information gap, with one informed agent whose capacity share of the total capacity of 16 varies across columns and M uninformed agents sharing the remainder. Means across surviving runs of 16-seed ensembles under the absolute NashConv filter ($\epsilon = 3$), standard errors in parentheses, survivor counts in brackets. The 12.5% column at $M = 2$ is the workhorse cell under the extension protocol (see Section 4.4). †In this cell the single uninformed agent's average loss fails the exact no-trade test of Section 4.6, so the $M = 1$ row is read as descriptive rather than as filtered equilibrium evidence. The one-sided filter reproduces every cell within sampling error.

model: large traders who split a sizable position into a sequence of orders executed gradually over time, so that their order flow generates persistent local price trends. We implement the resulting price dynamics in reduced form as

$$p_t = p_{t-1} + \alpha(p_{t-1} - p_{t-2}) + \eta_t + \mu_t, \tag{15}$$

where α controls autoregressive persistence, η_t is i.i.d. noise, and μ_t is a slowly varying trend component capturing the block traders' gradual execution, which creates the exploitable local momentum. The learning agent's own market impact is set to zero ($\lambda = 0$) and there is no fundamental revelation, isolating the agent's ability to detect and trade on the price patterns created by the rule-based agents.