Algorithmic Pricing under Strategic Delegation*

Agostino Capponi Hao Yang Zhengge Zhou

November 11, 2026

Abstract

Does competition between market makers over algorithm choice lead to the adoption of algorithms that produce competitive prices? We study strategic delegation of market making to learning algorithms. Market makers commit to *Basic* Q-learning or *Clever* counterfactual Q-learning and evaluate expected average monetary profit over the learning path. *Clever* can produce competitive prices when mutually adopted, but requires additional market data. In the baseline *Basic–Clever* game, mutual *Basic* is an equilibrium for every nonnegative data cost, whereas mutual *Clever* requires sufficiently cheap data. Strong adverse selection can also make *Basic* strictly dominant. Competition over algorithm adoption therefore does not guarantee competitive prices. Allowing delegation to misaligned *Biased* algorithms with distorted learning rewards can further worsen market quality.

Keywords: Algorithmic Pricing; Strategic Delegation; AI Alignment; Market Making; Reinforcement Learning.
JEL codes: D43, D82, D83, G12, G14.

*Capponi (ac3827@columbia.edu) is at Columbia University. Yang (hyang@em-lyon.com) is at Emlyon Business School. Zhou (zhengge.zhou@uwa.edu.au) is at the University of Western Australia.

1 Introduction

"[A]n AI agent ... acts with delegated authority."

"An autonomous optimizer, such as a Q-learning trader, ... maximizes an explicit objective, and misalignment arises most naturally in this class."

Brunnermeier (2026)

"Is it humans vs AI or is it (humans plus AI) vs (humans plus AI)?"

Rajan (2026)

Prices in securities markets are increasingly set by AI-powered pricing algorithms. Existing work shows that these algorithms can produce sharply different pricing outcomes when mutually adopted by market makers. In a duopoly request-for-quote securities market, Colliard et al. (2026) show that a standard Q-learning algorithm may fail to learn competitive pricing strategies and sustain supracompetitive prices. Guarino et al. (2025), by contrast, show that counterfactual Q-learning can produce competitive prices when it is endowed with a basic understanding of the market and information about outstanding quotes. We refer to chosen-action Q-learning as *Basic* and to counterfactual Q-learning as *Clever*. The two papers study different trading environments. We compare these updating rules in a common market and make the choice between them a strategic decision.

Does competition between market makers over algorithm choice lead to the adoption of algorithms that produce competitive prices? Our answer is: not necessarily. Market makers do not choose their algorithms in isolation. The profitability of one algorithm depends on the algorithm adopted by the competing market maker. An algorithm that leads to competitive prices when mutually adopted need not be profitable for one market maker to adopt unilaterally. The distinction between pricing performance and private adoption incentives is central to our analysis.

We study strategic delegation of pricing to learning algorithms and trace the consequences of (mis)alignment for market outcomes. In the baseline game, each market maker first chooses

either *Basic* or *Clever*. *Basic* updates only the value of the quote actually chosen, whereas *Clever* also updates unchosen quotes whose payoffs can be inferred from the best quote and the realized trade outcome. *Clever* can be more costly to adopt because it requires access to additional market data. Then the selected algorithms compete by repeatedly submitting asks in the request-for-quote market of Colliard et al. (2026). The client has private information about her valuation, and trades expose the market makers to adverse selection. The algorithms learn from the outcomes generated by their interaction.

We use expected average monetary profit during learning to construct the algorithm-adoption game. A market maker's payoff depends on its own algorithm and the algorithm adopted by its competitor. Both algorithms use the same quote grid, exploration schedule, learning rate, and optimistic initialization. The baseline implementation follows Colliard et al. (2026), with one million learning episodes and 1,000 matched experiments. The market maker evaluates the complete learning path, including the profits and losses incurred before a pricing pattern emerges. We account for Monte Carlo uncertainty using standard errors and simultaneous bounds for unilateral adoption incentives.

Choosing a pricing algorithm and then letting it play is a strategic-delegation problem. A market maker commits to an agent whose subsequent quoting behavior shapes the market interaction and influences its competitor. As in the strategic-delegation models of Vickers (1985) and Fershtman and Judd (1987), the preferred delegate depends on the rival's response. Here, the market maker commits to a learning procedure before that procedure learns how to quote. The relevant competition is therefore between combinations of human and algorithmic decision makers (Rajan, 2026). Delegated authority and an explicit learning objective also make this a natural setting for studying alignment (Brunnermeier, 2026).

In the baseline *Basic–Clever* game, mutual adoption of either algorithm is an equilibrium. Facing *Basic*, a market maker earns about 0.2822 per episode with *Basic* but only 0.1542 with *Clever*. Facing *Clever*, it earns about 0.0400 by matching *Clever* and 0.0018 by using *Basic*. Each market maker therefore prefers to match its rival's algorithm. The adoption game is a

coordination game, not a prisoner's dilemma. Mutual *Basic* sustains high quotes and high dealer profits, while mutual *Clever* produces quotes close to the competitive benchmark. Both equilibria are supported by the simultaneous deviation tests.

The private cost of more competitive pricing is visible in asymmetric competition. *Clever* learns to undercut to an ask of about 3.1 and captures approximately 99% of traded volume against *Basic*. These thin margins leave the market maker with less profit than under mutual *Basic*. In BB, neither algorithm learns to undercut to the competitive price, and the market quote remains near 5.0. *Clever*'s success in winning orders from *Basic* thus does not imply that a market maker wants to replace *Basic* with *Clever*. The adoption comparison holds the rival's delegate fixed and compares the market maker's profits across the resulting markets.

The cost of market-data access determines whether the *Clever* equilibrium survives. A recurring fee paid only by market makers choosing *Clever* lowers its adoption payoff without changing learning or prices within a profile. At the baseline parameters, CC remains an equilibrium up to a fee of approximately 0.0382 per episode, about 95% of each market maker's gross CC profit. Above the threshold, BB is the unique equilibrium. BB is also payoff dominant and risk dominant at zero cost. *Basic* is a best response whenever the probability that the rival chooses *Basic* exceeds approximately 23%. Mutual *Basic* remains an equilibrium for every nonnegative *Clever* data fee in this baseline two-algorithm game.

The market environment changes the adoption incentives. Holding fundamental values and the learning design fixed, we compare $\sigma = 1, 5, 15$, the standard deviation of the client's liquidity shock. At the strong-adverse-selection setting $\sigma = 1$, *Basic* is strictly dominant under the full-horizon mean payoffs. A *Basic* rival learns a similarly low ask against *Clever* and earns more than a market maker earns under mutual *Clever*. *Basic* can therefore free ride on *Clever*: it behaves as if it imitates *Clever*'s low ask, although its updating still uses only its own realized profit. At the baseline $\sigma = 5$ and the weaker-adverse-selection setting $\sigma = 15$, both matching equilibria remain. These comparisons concern the simulated settings; changing liquidity dispersion affects demand and learning feedback together and does not

establish a monotone adoption threshold.

We then ask whether market makers can benefit from deliberately distorting the learner's objective. The delegation menu includes a *Biased* algorithm whose reward is monetary profit plus γ times executed quantity. A positive coefficient is an execution bonus, and a negative coefficient is an execution penalty. The market maker continues to value monetary profit alone. Each fixed γ defines a separate three-action delegation game. The distinction between the market maker's payoff and the learner's reward permits commitment to a tougher or softer pricing agent without changing the market maker's objective.

The direction of the bias matters. At $\gamma = +0.8$, mutual *Biased* is an equilibrium with lower asks and lower profits than BB. Restricting the menu to *Basic* and *Biased* gives a prisoner's-dilemma payoff ordering: each market maker prefers *Biased* against either rival choice, although both earn less under DD than under BB. The gains from leaving BB are small, and their simultaneous confidence intervals include zero. CC remains an equilibrium in the full three-action game. At $\gamma = -0.2$, mutual *Biased* instead yields higher asks and higher profits than BB. A small execution penalty softens competition and raises each market maker's monetary profit to approximately 0.2942 per episode. Strategic delegation can further worsen market quality, sustaining worse quotes and spreads together with higher dealer profits. Selected additional coefficients and market-data costs show when these biased-delegation outcomes can be sustained.

The baseline *Basic–Clever* conclusions also survive changes in evaluation and adoption timing. Both matching equilibria remain when market makers evaluate the final 10,000, 20,000, or 50,000 episodes. They also remain when market makers choose whether to upgrade after a shared *Basic* trading history, retaining the accumulated Q-values and exploration clocks. In the baseline sequential *Basic–Clever* game ($\sigma = 5$, $\kappa = 0$), whichever market maker leads adopts *Basic*, and the follower responds with *Basic*. This selection result uses full-horizon average monetary profits. It is not a claim for every value of σ or for a menu that also includes *Biased*.

The rest of the paper proceeds as follows. Section 2 presents the market environment, learning rules, and the delegation game. Section 3 studies *Basic–Clever* adoption, market outcomes, adverse selection, and data costs. Section 4 studies biased delegation. Section 5 discusses the implications, and Section 6 concludes. The appendices report numerical conventions, full cross-experiment profit distributions, and profit and quote paths.

1.1 Related Literature and Contribution

Algorithmic behavior in markets. One strand studies competition between algorithms, including algorithmic pricing (Calvano et al., 2020; Brown and MacKay, 2026), market making (Colliard et al., 2026; Guarino et al., 2025), informed trading (Dou et al., 2025), co-learning among traders (Sangiorgi et al., 2026), and coordination in financial crises (Yang, 2024). A related strand studies interactions between algorithmic and human traders (Banerjee and Szydlowski, 2025; Dou et al., 2026). We study competition between market makers who choose the algorithms that act on their behalf. We endogenize algorithm choice as a strategic decision and ask whether market makers adopt learning rules that produce competitive prices.

Our closest references for the learning rules are Colliard et al. (2026) and Guarino et al. (2025). The former study market making with uncertain asset values, liquidity shocks, and adverse selection. Noisy profit feedback makes competitive undercutting difficult to learn. The latter show how basic economic understanding and quote information permit counterfactual updating. We retain the former paper's market environment and baseline parameters and adapt the latter paper's updating principle to it. This permits a common-market comparison of *Basic*, which updates the chosen quote, and *Clever*, which also updates identifiable unchosen quotes. The distinction is related to synchronous and asynchronous learning (Asker et al., 2022, 2024), although *Clever* does not observe counterfactual demand at every quote. Market data also matter for co-learning in Sangiorgi et al. (2026). We study the private return to using additional data and the effect of their cost on equilibrium adoption.

Other work examines sequential pricing (Klein, 2021), pricing speed and commitment

(Brown and MacKay, 2026), and a learning algorithm against a rule-based rival (Xu, 2025). Our comparison holds quoting speed fixed and lets both alternatives learn. Adopting *Clever* changes both its own pricing and the rival's learning path. Consequently, earning more than the rival within an asymmetric market need not make adoption profitable: the relevant comparison is the same market maker's profit across algorithm pairs. This separates the competitive consequences of an algorithm from the incentive to adopt it.

Strategic delegation. The delegation literature studies the allocation of decision authority (Holmström, 1977) and the competitive effects of commitment to managers with selected objectives (Vickers, 1985; Fershtman and Judd, 1987; Fershtman et al., 1991). We study strategic delegation of pricing to learning algorithms. The commitment specifies an updating rule or a learning reward; it does not fix the subsequent quotes. In particular, *Biased* adds a volume bonus or penalty to the learner's reward, while the market maker continues to value monetary profit. We trace the consequences of this misalignment for adoption incentives, prices, spreads, and trade.

Recent work studies delegation to an AI whose alignment is uncertain. Chen et al. (2024) examine testing through imperfect recall; Fudenberg and Liang (2025) study information provision to a potentially misaligned AI; and Dworczak and Smolin (2026) study robust use of potentially misaligned advice. In our setting, market makers know the algorithms' updating rules and reward objectives, although the algorithms must learn from experience. A market maker may deliberately choose a distorted reward because it changes competition in its favor. We thus study when strategic misalignment can raise private profit and reduce market quality.

Program equilibrium. Our delegation equilibrium also relates to program equilibrium: players commit to computer programs that act on their behalf (Tennenholtz, 2004). Oesterheld (2019) and Cooper et al. (2025) study cooperation through programs that can inspect or simulate other programs. Learned collusion and strategic algorithm design are also examined

by Compte (2025); Carissimo et al. (2025). Our market makers choose from a finite set of learning algorithms. The algorithms use market feedback and neither inspect the rival's source code nor simulate its program. Delegation payoffs are expected average monetary profits over the learning path, including the cost of acquiring data. This connects program commitment to the economic question of whether competition over algorithm choice delivers competitive market making.

2 The Model

We study an algorithm delegation game. Market makers commit to algorithm designs, and the selected algorithms quote on their behalf in a financial market. The algorithms compete and learn for T episodes. A market maker's delegation payoff is its expected average monetary profit over the entire learning path generated by the algorithm pair.

2.1 Economic Environment

The environment follows Colliard et al. (2026) and the adverse-selection setting of Glosten and Milgrom (1985). There are two risk-neutral market makers, indexed by $i \in \{1, 2\}$, and a client. In episode t, the asset value $\widetilde{v}_t \in \{v_L, v_H\}$ takes each value with probability $1/2$, with $v_L < v_H$. The client observes her valuation

$$\widetilde{v}_t^C = \widetilde{v}_t + \widetilde{L}_t, \qquad \widetilde{L}_t \sim \mathcal{N}(0, \sigma^2), \qquad \widetilde{L}_t \perp \widetilde{v}_t.$$

The two algorithms observe neither $\widetilde{v}_t$ nor $\widetilde{L}_t$ before quoting. The asset value and independent liquidity shock are redrawn in each episode. The asset value is revealed after the client's trading decision; her private valuation is not revealed to the algorithms. Let $\overline{v} = (v_L + v_H)/2$ denote the unconditional mean asset value.

The asset value is also the market maker's cost of supplying the asset. A client is more likely to buy at a given ask when that value is high, exposing the market maker to adverse selection. The liquidity shock adds a trading motive independent of asset value. Holding v_L

and v_H fixed, a smaller σ corresponds to stronger adverse selection in the environments we compare. To see this, note that a buy at a chosen quote a is news about $\widetilde{v}_t$:

$$\Pr(v_H \mid \text{buy at } a) = \frac{\Phi\left(\frac{v_H - a}{\sigma}\right)}{\Phi\left(\frac{v_H - a}{\sigma}\right) + \Phi\left(\frac{v_L - a}{\sigma}\right)},$$

where Φ is the standard normal cumulative distribution function. For the quote grid considered here, where $a > v_L = 0$, $\Pr(v_H \mid \text{buy at } a)$ falls as σ rises. That is, small σ implies that a buy almost reveals v_H, while a buy mostly reflects liquidity needs under large σ. Varying σ also changes demand and the feedback generated by trade, so the comparisons concern the complete market environment.

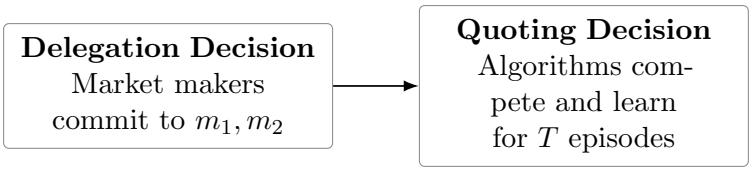
2.2 The Delegation Decision

Each market maker i commits to an algorithm design m_i. The baseline delegation game has menu

$$\mathcal{M} = \{\textit{Basic}, \textit{Clever}\}.$$

We subsequently add *Biased*(γ), where γ determines the learner's bias toward execution volume. Each fixed γ defines a separate three-action game. The market maker commits to the algorithm design, including its learning rule and reward, throughout the T learning episodes. The algorithm continues to adjust its quotes as feedback arrives. A market maker's unilateral change of delegate therefore changes the subsequent interaction between the two algorithms. The rival retains its own learning rule, but encounters different quotes, executions and profit observations. Delegation payoffs incorporate that endogenous response along the learning path.

Figure 1: The delegation meta-game.



2.3 The Quoting Decision

In each episode, the delegated algorithms simultaneously quote asks a_{1t}, a_{2t} from a finite price grid $\mathcal{A}$. The client buys one unit at the lowest ask if and only if

$$a_t^{\min} = \min_i a_{it} \leq \widetilde{v}_t^C.$$

A unique lowest-quote market maker wins the whole order. If the algorithms tie at an executable ask, each receives half the order. The realized monetary profit is

$$\pi_{it} = \begin{cases} a_{it} - \widetilde{v}_t, & i \text{ wins: } a_{it} < a_{-i,t} \text{ and } a_t^{min} \leq \widetilde{v}_t^C; \\ \frac{1}{2}(a_{it} - \widetilde{v}_t), & i \text{ ties: } a_{it} = a_{-i,t} \text{ and } a_t^{min} \leq \widetilde{v}_t^C; \\ 0, & i \text{ loses or no trade occurs.} \end{cases} \tag{1}$$

Equivalently, $\pi_{it} = (a_{it} - \widetilde{v}_t)x_{it}$, where $x_{it} \in \{0, 1/2, 1\}$ is the share of the current unit order executed by market maker i.

Competitive benchmark. The competitive continuous Glosten–Milgrom price satisfies

$$a^* = \mathbb{E}[\widetilde{v}_t \mid a^* \leq \widetilde{v}_t^C].$$

Our algorithms quote on a discrete grid. We therefore calculate the grid equilibria by checking the expected monetary profit from every unilateral quote change, rather than rounding the continuous solution. At the baseline parameters, the continuous price is approximately 2.6851, and the symmetric grid equilibria are 2.7 and 2.8. We use the higher quote, $a^{\text{Nash}} = 2.8$, as the competitive price and recalculate the benchmark when σ changes. Appendix A.2 gives the calculation. These are benchmarks for pricing competition; delegation payoffs instead evaluate the complete learning path.

2.4 The *Basic* Algorithm

Basic Q learns from the chosen quote. Its Q-values summarize estimated profit at each quote $a \in \mathcal{A}$. In episode t, it first chooses a quote according to

$$a_{it} = \begin{cases} \text{uniform on } \arg\max_{a \in \mathcal{A}} Q_{i,t-1}(a), & \text{with probability } 1 - \epsilon_t; \\ \text{a uniform draw from } \mathcal{A}, & \text{with probability } \epsilon_t. \end{cases}$$

Here $\epsilon_t = e^{-\beta t}$. Exploitation chooses a quote with the highest Q-value, with uniform tie-breaking. Exploration picks a random quote. The exploration probability falls over time, faster for a larger β.

The algorithm then earns profit π_{it} and updates only the chosen quote:

$$Q_{i,t}(a_{it}) = (1 - \alpha) Q_{i,t-1}(a_{it}) + \alpha \pi_{it}. \tag{2}$$

The update combines past knowledge $Q_{i,t-1}$ with current realized profit π_{it}. The learning rate α is the weight on that profit; a higher α forgets past observations faster. Unchosen Q-values are unchanged. *Basic* only uses its own realized profit for updating.

Basic treats the quote grid as the action set of a bandit problem. It knows which quotes are available, but does not know the stochastic mapping from quotes to profits or calculate each quote's expected profit. Even at unchanged asks, the asset value and liquidity shock make realized profits noisy. Moreover, as the rival learns, the return to a given quote can change. The Q-values summarize the learner's experience in this evolving environment.

For example, suppose exploration produces a quote of 5.0, the order executes uniquely, and the realized fundamental is 4. The market maker earns $\pi_{it} = (5.0 - 4) \times 1 = 1$. With $Q_{i,t-1}(5.0) = 4.5$ and $\alpha = 0.01$, the updated value is

$$Q_{i,t}(5.0) = 0.99 \times 4.5 + 0.01 \times 1 = 4.465.$$

This constructed one-step example illustrates the update; it is not a reported simulation path.

2.5 The *Clever* Algorithm

Clever Q learns about undercutting. It has the same quote-choice rule, learning rate, and monetary-profit objective as *Basic*. In addition to its own realized profit, *Clever* sees market data: the best quote and the trade outcome. Its allocation and the ex post asset value allow it to infer the profit from identifiable unchosen quotes. The market maker pays for access to this feedback.

The economic understanding is simple. A client who buys at a given ask would also buy at a lower ask. In the example above, a unique win at 5.0 implies that lower quotes also trade. An unchosen quote of 2.8 would therefore earn

$$\pi_{it}^{\mathrm{cf}}(2.8) = 2.8 - 4 = -1.2.$$

If its previous Q-value was 5.1, *Clever* also updates that quote to $0.99 \times 5.1 + 0.01 \times (-1.2) = 5.037$. The chosen quote receives the same update to 4.465 as under *Basic*.

A unique winner can make this undercutting inference from its own trade. Additional market feedback is useful when the market maker does not trade. For example, a market maker quoting 6 earns zero if its rival sells at 5. Observing that transaction lets it infer that an ask below 5 would have won the order, while an ask of 5 would have shared it. Its own zero profit alone does not identify these counterfactual returns.

More generally, let $\mathcal{I}_{it}$ denote the quotes whose profit targets $\pi_{it}^{\mathrm{cf}}(a)$ are identified by the observed event. *Clever* updates

$$Q_{i,t}(a) = (1-\alpha)Q_{i,t-1}(a) + \alpha\pi_{it}^{\mathrm{cf}}(a), \qquad a \in \mathcal{I}_{it}. \tag{3}$$

The chosen quote is included and updated once. Entries outside the identified set are unchanged. Appendix A gives the full observable-event update rule, including no-trade episodes and traded ties. Counterfactual feedback receives the same learning weight as chosen-action feedback, $\rho = 1$.

This learning rule adapts the counterfactual-updating principle of Guarino et al. (2025)

to the economic environment of Colliard et al. (2026). The two source papers use different markets and learning parameters. We compare *Basic* and *Clever* within the same market and do not transfer a convergence result to our constant-learning-rate implementation.

2.6 The *Biased* Algorithm

Biased Q learns from a biased reward. It uses the same chosen-action update as *Basic*, replacing monetary profit with

$$r_{it}^{D} = \pi_{it} + \gamma x_{it}. \tag{4}$$

A positive γ is a volume bonus, associated with a tougher agent. A negative γ is a volume penalty, associated with a softer agent. The market maker's objective remains monetary profit. The reward distortion affects the learner's updating, not the monetary payoff received by the market maker. We introduce the corresponding delegation results in Section 4.

2.7 Equilibrium Concept

Given a delegated pair (m_i, m_{-i}), the algorithms repeat the quoting game for T episodes, learning as they go. Market maker i evaluates its delegate by expected average profit over the learning path:

$$u_i(m_i, m_{-i}; \kappa) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \pi_{it}(m_i, m_{-i})\right] - \kappa \mathbf{1}\{m_i = \textit{Clever}\}. \tag{5}$$

The expectation averages over initialization, exploration, randomized choices and market shocks. The baseline has zero data cost, $\kappa = 0$. A market maker choosing *Clever* subsequently pays $\kappa \geq 0$ per episode for the best quote and trade-outcome data. Only *Clever* pays the fee, including in a three-action game. We deduct it from the market maker's payoff after evaluating the learning path; it does not enter the learner's reward.

A profile $(m_1^*, m_2^*) \in \mathcal{M}^2$ is a pure-strategy Nash equilibrium if no market maker gains from a unilateral change of algorithm design:

$$u_i(m_i^*, m_{-i}^*; \kappa) \geq u_i(m_i, m_{-i}^*; \kappa) \quad \text{for every } m_i \in \mathcal{M},\ i \in \{1, 2\}.$$

Market makers choose algorithms as if they knew the expected payoffs. We estimate these payoffs by simulating each algorithm pair. The delegation equilibrium incorporates the interaction between the learning algorithms throughout the evaluation horizon; it does not require their quotes to form a competitive equilibrium in every episode.

Evaluating the complete learning path gives adoption an economic cost that a final-price comparison can miss. Market makers receive the profits and bear the losses incurred while their delegates learn. Two procedures can eventually select similar quotes yet have different adoption values because of their earlier experience. Conversely, producing a more competitive final quote need not raise the market maker's average profit. The adoption decision therefore requires comparisons of complete algorithm pairs over a common evaluation horizon.

2.8 Simulation Design and Procedure

For a fair comparison, we use the same baseline parameters as Colliard et al. (2026):

$$v_L = 0, \quad v_H = 4, \quad \sigma = 5, \qquad \mathcal{A} = \{1.1, 1.2, \ldots, 14.9\},$$

$$\alpha = 0.01, \quad \epsilon_t = e^{-\beta t}, \quad \beta = 8 \times 10^{-5}, \qquad Q_{i,0}(a) \sim U[3, 6].$$

We run $K = 1,000$ matched experiments of $T = 10^6$ episodes each. Optimistic initial Q-values are independent across dealers and quotes and matched across profiles for the same dealer. Market shocks and algorithmic random streams are also matched across profiles. Each experiment is an independent statistical unit. Table A1 records the full numerical specification.

Exploration is concentrated early in the learning path. With $\beta = 8 \times 10^{-5}$, the exploration probability first falls below 1% at episode 57,565. The expected number of exploratory quote choices over the complete horizon is approximately 12,500 per market maker, or 1.25% of T. Most episodes therefore occur in the low-exploration phase, although low exploration alone does not establish convergence. Full-horizon profits include the initial learning period as well as the subsequent interaction between the algorithms.

Because dealer identities are economically interchangeable, we pool dealer-swapped payoffs within each matched experiment before estimating the delegation game. In Tables 4 and 5, the DB and BD entries instead retain the original DB role payoffs and their mirror; all other entries use role pooling. Appendix A defines this reference convention. The resulting payoff matrices are exactly symmetric. Means, MCSEs, adoption comparisons, cost thresholds and bootstrap inference use the corresponding $K = 1,000$ experiment observations, not $2K$ independent observations. Final-episode market outcomes and the original role-specific learning paths are unchanged.

Payoff matrices report full-horizon mean monetary profits, with stars summarizing the simultaneous equilibrium tests. Numerical Monte Carlo standard errors are collected in Table A3. We assess equilibrium using paired profit differences across profiles and simultaneous tests of unilateral deviations. The tests resample whole experiments, retaining the matching, with 2,000 bootstrap resamples. We calculate equilibria from the unrounded profit estimates and use simultaneous confidence intervals to assess the incentives for unilateral deviations. Appendix A gives the inference conventions and the confidence level used in each comparison.

We report five market outcomes: average quote, quoted spread, realized spread, trade probability, and collusion index. All four quote, spread and execution statistics use the final episode $T = 10^6$. Average quote and quoted spread use all 1,000 experiments: the former is the lowest market ask, including experiments without trade, and the latter subtracts the unconditional mean asset value 2. Realized spread averages the difference between the executed ask and actual asset value only over experiments trading in that final episode. Trade probability averages the model-implied execution probability at each saved final minimum ask across all 1,000 experiments. The collusion index compares full-horizon joint monetary profit with the discrete competitive and joint-profit-maximizing benchmarks. Exact formulas and the distinction between these time windows are in Appendix A.

3 *Basic–Clever* Delegation

3.1 Adoption Payoffs, Market Outcomes, and Data Costs

Our first question is whether algorithm advances lead to adoption. We first restrict the market makers' choices to *Basic* and *Clever*. Both algorithms learn from monetary profit. *Basic* learns from its chosen quote, whereas *Clever* also learns about undercutting using market data. Each market maker evaluates its delegate by average monetary profit over the full $T = 10^6$ learning episodes. The baseline has liquidity shock $\sigma = 5$, optimistic initialization, and zero market-data cost.

Table 1 reports the delegation payoffs across $K = 1,000$ matched experiments. Both BB and CC are equilibria. Facing *Basic*, a market maker earns approximately 0.282 with *Basic* but only 0.154 with *Clever*. Facing *Clever*, it earns approximately 0.040 with *Clever* and only 0.002 with *Basic*. The market maker therefore prefers to match the rival's algorithm. This is a coordination game, not a prisoner's dilemma. Both equilibria pass simultaneous tests of unilateral deviations at the 1% level.

Table 1: Baseline *Basic–Clever* delegation payoffs

Market maker 1 / 2	*Basic*	*Clever*
Basic	(0.2822, 0.2822) ***	(0.0018, 0.1542)
Clever	(0.1542, 0.0018)	(0.0400, 0.0400) ***

Notes: Each pair gives MM1's and MM2's role-pooled mean monetary profits per episode over $T = 10^6$ episodes and $K = 1,000$ matched experiments. Rows are MM1's choice and columns are MM2's choice. Optimistic $U[3, 6]$ initialization, $\sigma = 5$, and zero data cost. Shading marks Nash equilibria at unrounded mean payoffs. ***: all unilateral stay-minus-switch gains have positive simultaneous 99% lower bounds within this game. Stars below a pair refer to the whole equilibrium profile, not to profit relative to zero. MCSEs are reported in Table A3.

Table 2 shows why these two delegation equilibria matter for the market. *Clever* can deliver competitive prices, yet market makers may not adopt it. Mutual *Clever* has an average final quote of 2.888, close to the competitive price $a^{\text{Nash}} = 2.8$, compared with 5.015

under mutual *Basic*. Quoted spread, realized spread, and the collusion index are lower under CC, while trade probability is higher. Payoffs value the full learning path; quotes and spreads describe the final episode.

Table 2: Market outcomes in the *Basic–Clever* delegation game

Market outcome	BB	CC
Average quote	5.015	2.888
Quoted spread	3.015	0.888
Realized spread	1.937	0.230
Trade probability	29.07%	43.49%
Collusion index	0.792	0.052

Notes: At $T = 10^6$, except the collusion index, which uses full-horizon joint monetary profit. All four statistics use the final-episode observations. Average quote, quoted spread and trade probability average over all 1,000 experiments; realized spread averages only experiments trading at T. Average quote is the minimum market ask, including observations without a trade. Quoted spread subtracts the unconditional mean fundamental, 2. Realized spread averages $a_T^{\min} - v_T$ only over experiments with a trade. Trade probability integrates over the client-valuation distribution at each saved final quote. Competitive price: $a^{\text{Nash}} = 2.8$. Definitions follow Section 2.8.

Appendix B reports density estimates and empirical CDFs of experiment-average profits for both market makers. A market maker experiences one realization of the learning process. Under BB, the 10th and 90th percentiles of MM1's average profit are approximately 0.239 and 0.326; the corresponding CC percentiles are 0.0396 and 0.0404. These percentiles describe variation across learning paths, whereas MCSEs measure uncertainty about their means. Risk-neutral adoption depends on expected paired payoff differences across algorithm profiles.

3.2 Best Responses and Equilibrium Adoption

BB is payoff dominant: both market makers earn more under BB than under CC. The market maker's decision also depends on its belief about the rival's algorithm. From market maker i's perspective, write

$$g_{Bi} = u_i(B, B; 0) - u_i(C, B; 0), \qquad g_{Ci} = u_i(C, C; 0) - u_i(B, C; 0).$$

The first gap is approximately 0.1280, and the second is approximately 0.03822, for both market makers. If p is the probability that the rival chooses *Basic*, *Basic* is a best response when

$$pg_{Bi} - (1-p)g_{Ci} \geq 0, \qquad \text{or equivalently} \qquad p \geq \frac{g_{Ci}}{g_{Bi} + g_{Ci}} \simeq 23\%.$$

Thus BB is also risk dominant. Payoff dominance compares the two mutual adoption outcomes; the belief threshold describes the best response to an uncertain rival choice.

3.3 Pricing Outcomes and Competitive Benchmarks

At the baseline parameters, *Clever* undercuts to about 3.1 and wins 99% of trade volume, but thin margins leave it with just 0.154 per episode. In BB, neither algorithm learns to undercut to the competitive price, so both quote near 5.0 and earn 0.282 each.

Figure A7 follows the same asymmetric market: MM1 plays *Clever* and MM2 plays *Basic*. *Clever* lowers its ask early, and *Basic*'s profit falls close to zero. These original role-specific paths are not pooled across dealer identities; they explain the competitive interaction. The adoption comparison instead evaluates the same market maker's 0.154 in CB against its 0.282 in BB. Holding the rival's algorithm fixed still allows its quotes to respond through learning. A more aggressive delegate can take business from a rival while yielding less profit to its market maker than retaining *Basic*.

Delegation and adverse selection. We next compare strong adverse selection, $\sigma = 1$, with the baseline $\sigma = 5$ and weak adverse selection, $\sigma = 15$. Fundamental values, the quote grid, initialization, and the learning parameters are unchanged. A smaller liquidity shock makes the client's valuation more closely tied to the fundamental value. Changing σ also changes demand and the feedback received by the algorithms. We compare the simulated settings separately, without interpolating between them.

At $\sigma = 1$, *Basic* is strictly dominant under the full-horizon mean payoffs. Against *Basic*, a market maker earns about 0.0203 by retaining *Basic* and 0.0104 by choosing *Clever*. Against

Table 3: Delegation payoffs at stronger and weaker adverse selection

Panel A. Strong adverse selection, $\sigma = 1$

Market maker 1 / 2	*Basic*	*Clever*
Basic	(0.0203, 0.0203) ***	(0.0071, 0.0104)
Clever	(0.0104, 0.0071)	(0.0036, 0.0036)

Panel B. Weak adverse selection, $\sigma = 15$

Market maker 1 / 2	*Basic*	*Clever*
Basic	(0.8974, 0.8974) ***	(0.0088, 0.2016)
Clever	(0.2016, 0.0088)	(0.0474, 0.0474) ***

Notes: Role-pooled mean monetary profits per episode over $T = 10^6$ episodes in $K = 1,000$ optimistic experiments, with zero data cost. Shading marks Nash equilibria at unrounded means. ***: every unilateral stay-minus-switch gain has a positive simultaneous 99% lower bound. Bounds cover twelve indexed payoff contrasts separately within each matrix; coverage is not joint across the two settings. Stars below a pair refer to the whole equilibrium profile, not to profit relative to zero. MCSEs are in Table A3.

Clever, *Basic* earns about 0.0071, compared with 0.0036 under CC. Both market makers have the same best-response signs, and BB is the unique equilibrium.

The intuition is that *Basic* can free ride on *Clever*. It behaves as if it imitates *Clever* to set a lower ask. Hence, the market maker has no incentive to switch to *Clever*. The strong-adverse-selection panels of Figure A8 show *Basic* learning a similarly low ask: in the final 50,000-episode block, *Clever*'s mean own ask is about 4.100 and *Basic*'s is 4.105. Their profits within CB are then close. This is an as-if description of the learning paths; *Basic* continues to update from its own experienced profit rather than directly copying the rival's quote.

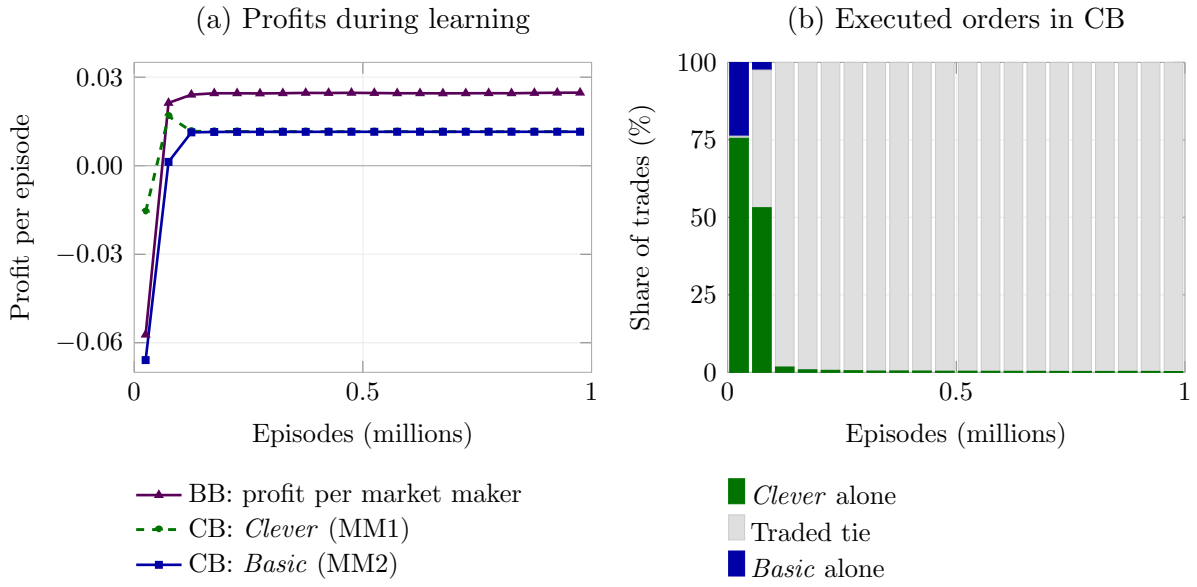
At $\sigma = 15$, the delegation game is again a coordination game. Mutual *Basic* earns approximately 0.8974 per market maker. A unilateral *Clever* adopter earns approximately 0.2016, while mutual *Clever* earns 0.0474 per market maker. *Clever*'s ask approaches the competitive benchmark as *Basic*'s profit falls close to zero. The weak-adverse-selection case therefore retains the baseline trade-off: tougher competition within CB does not make adopting *Clever* against *Basic* privately worthwhile.

Figure 2 separates the early learning losses from the later profit comparison. BB loses about 0.0572 per market maker per episode in the first 50,000 episodes, but earns 0.0247 in the final 100,000. In CB, *Clever* loses less than *Basic* early on; their final-100,000-episode profits are both approximately 0.0115. By that window, about 99.7% of executed episodes are ties. Thus BB's higher full-horizon profit at $\sigma = 1$ comes despite its larger initial loss. Within CB, *Clever*'s early advantage over *Basic* does not persist as a large profit difference.

3.4 Market-Data Access Costs

Suppose each market maker that chooses *Clever* pays $\kappa \geq 0$ per episode for market data: the best quote and the trade outcome. The fee is a subscription paid whether or not that market maker trades. It applies only to *Clever*, including when *Biased* is available. It is deducted from the market maker's payoff after simulation and does not change the learner's reward or

Figure 2: Learning under strong adverse selection ($\sigma = 1$).



Notes: Each point or bar represents a nonoverlapping 50,000-episode block of the saved 10^6-episode learning paths. Profits average across 1,000 matched experiments; the BB series also averages the two market makers. Both CB series refer to the same market, with MM1 using *Clever* and MM2 using *Basic*. Panel (b) pools executed episodes across experiments within each block. A traded tie is counted once and allocates half the unit order to each market maker. No-trade episodes are excluded from these shares.

the price paths within a profile:

$$u_i(m_i, m_{-i}; \kappa) = u_i(m_i, m_{-i}; 0) - \kappa \mathbf{1}\{m_i = C\}. \tag{6}$$

At the baseline parameters, a positive fee reinforces the incentive to retain *Basic* against *Basic*. Its equilibrium-selection effect is to weaken the incentive to retain *Clever* against *Clever*.

In the *Basic–Clever* baseline game, the fee can eliminate the CC equilibrium. The largest cost for which CC survives is

$$\widehat{\kappa}^* = \min_{i \in \{1,2\}} \{u_i(C, C; 0) - u_i(B, C; 0)\} \simeq 0.0382.$$

This data-cost threshold is about 95% of each market maker's gross CC profit, 0.040 per episode. CC remains a weak equilibrium at the threshold. Above it, BB is the only equilibrium. Mutual *Basic* is always an equilibrium in the *Basic–Clever* game at these baseline market parameters because a positive fee only reduces the return to switching to *Clever*.

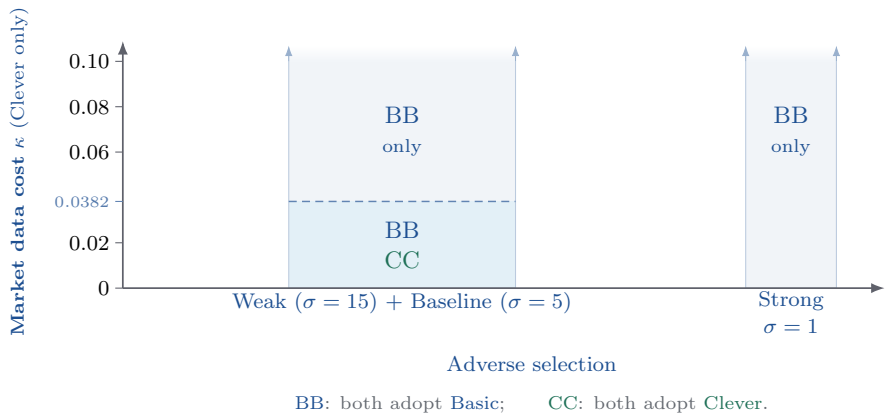
The same calculation applies separately to the adverse-selection cases. At $\sigma = 15$, BB and CC coexist up to a cost of approximately 0.0386; above it, BB is the only pure equilibrium. At $\sigma = 1$, *Basic* is already strictly dominant at zero cost and BB remains the only equilibrium for every nonnegative fee. These are three simulated adverse-selection settings with a continuous cost comparison within each setting.

Figure 3 summarizes this comparison. The two matching equilibria can coexist when data are inexpensive, but extitClever's availability does not ensure its adoption.

3.5 Robustness and Adoption Timing

We also assess statistical stability following Colliard et al. (2026): a pure delegation profile is retained when no unilateral deviation yields a statistically significant profit gain over the full learning path. We use one-sided paired t-tests, preserving the matching across our experiments, and their nominal 25% significance threshold. This is a paired counterpart to

Figure 3: Algorithm adoption and the cost of market data.



Notes: The left bar combines the separately simulated weak ($\sigma = 15$) and baseline ($\sigma = 5$) games for display; the right bar shows strong adverse selection ($\sigma = 1$). All games use optimistic initialization, $K = 1,000$, $T = 10^6$, and unrounded full-horizon mean monetary profits. Only *Clever* pays κ, deducted analytically without changing learning. The common dashed line uses the baseline CC cutoff, $\widehat{\kappa}_5 \simeq 0.0382218$; the weak-case cutoff is $\widehat{\kappa}_{15} \simeq 0.0385765$. For $\widehat{\kappa}_5 < \kappa \leq \widehat{\kappa}_{15}$, CC survives only at $\sigma = 15$; this narrow interval is not resolved in the combined bar. CC remains a weak equilibrium at each exact cutoff. At $\sigma = 1$, BB is the only equilibrium for every nonnegative cost. Open tops and upward arrows indicate continuation to all higher costs; 0.10 is a displayed tick, not a ceiling.

their independent-sample Welch procedure. Among the four pure profiles in the baseline *Basic–Clever* game, the check retains BB and CC. The coordination result therefore also holds under this statistical-deviation criterion. This additional check complements the simultaneous incentive tests reported above; Appendix A gives the test definitions.

The baseline result is robust to different payoff horizons. BB and CC remain equilibria when market makers use the last 10,000, 20,000, or 50,000 episodes instead of the full horizon. These comparisons use the saved episodes of the same learning paths.

The result is also robust to trading history. In the midstream-upgrade experiment, the two algorithms first share a 500,000-episode *Basic–Basic* history. Market makers then choose whether to retain *Basic* or upgrade to *Clever*, inheriting Q-values and exploration clocks. Payoffs use the following 500,000 episodes and exclude sunk prehistory profits. The continuation remains a coordination game.

An upgrade changes the updating rule without restarting learning. In the completed continuation game, mutual *Basic* earns approximately 0.2909 per market maker, a unilateral *Clever* adopter earns about 0.1515, and mutual *Clever* earns about 0.0383. Both matching profiles remain equilibria. An existing learning history therefore preserves the coordination problem: the market maker still gives up profit by upgrading alone, even though mutual upgrading produces the more competitive outcome.

Appendix B displays the corresponding profit distributions as well as the distributions under alternative initializations.

Finally, consider sequential (Stackelberg) delegation in the baseline *Basic–Clever* game: $\sigma = 5$, $\kappa = 0$, optimistic initialization, and expected average monetary profit over all $T = 10^6$ learning episodes. The leader makes a binding algorithm choice, which the follower observes before choosing its own algorithm; learning begins after both commitments. The follower prefers *Basic* after *Basic* and *Clever* after *Clever*. The leader can therefore induce either matching outcome and selects *Basic* because BB gives it the higher profit. Backward induction gives the unique subgame-perfect outcome BB, regardless of which market maker

leads. This is equilibrium selection in the stated baseline, not a result for all values of σ. The order of algorithm commitments changes; the algorithms' subsequent quoting mechanism is unchanged.

4 *Biased* Delegation and Market Quality

We now ask whether biased Q-learning can be adopted. Each market maker continues to maximize monetary profit, but can delegate to an algorithm whose learning reward differs from monetary profit. In each game, the menu is $\{B, D_\gamma, C\}$ and the coefficient γ is fixed. *Biased* updates only the chosen quote, using

$$r_{it}^D = \pi_{it} + \gamma x_{it}.$$

A volume bonus, $\gamma > 0$, means a tougher agent. A volume penalty, $\gamma < 0$, means a softer agent. The reward adjustment is virtual: the market maker receives monetary profit π_{it}, without the additional term. The commitment is to the learner's objective, as in strategic delegation to a manager with a distorted objective (Vickers, 1985; Fershtman and Judd, 1987).

The one-step example makes the distinction explicit. Suppose a quote of 5 wins one unit and the realized fundamental is 4, so monetary profit is 1. Starting from a chosen-quote Q-value of 4.5, *Basic* updates to $0.99 \times 4.5 + 0.01 \times 1 = 4.465$. With $\gamma = +0.8$, the *Biased* learner receives 1.8 and updates to 4.473. With $\gamma = -0.2$, it receives 0.8 and updates to 4.463. The market maker earns the same monetary profit in all three cases. This example illustrates the learning reward; the delegation results below use the complete simulated learning paths.

4.1 A Positive Execution Reward

At $\gamma = +0.8$, the *Basic*–*Biased* game has the payoff ordering of a prisoner's dilemma. Against *Basic*, adopting *Biased* raises profit from 0.2822 to 0.2849 for either market maker. Against *Biased*, retaining *Biased* yields approximately 0.2197 per episode, compared with 0.0310 from

choosing *Basic*. Each market maker therefore prefers *Biased* against either rival choice, while both would earn more under BB than under DD.

This ordering applies to the restricted *Basic–Biased* game. In the full three-action game in Table 4, DD and CC are pure equilibria. *Biased* is not dominant against *Clever*, since using *Biased* against *Clever* earns approximately −0.0015, below the 0.0400 from matching *Clever*. The simultaneous confidence intervals for the small gains from leaving BB include zero. DD and CC pass the simultaneous tests of unilateral deviations.

Table 4: The positive-bias delegation game, $\gamma = +0.8$

MM1 / MM2	*Basic*	*Biased*	*Clever*
Basic	(0.2822, 0.2822)	(0.0310, 0.2849)	(0.0018, 0.1542)
Biased	(0.2849, 0.0310)	(0.2197, 0.2197) ***	(−0.0015, 0.0921)
Clever	(0.1542, 0.0018)	(0.0921, −0.0015)	(0.0400, 0.0400) ***

Notes: Mean monetary profits per episode for MM1 (row) and MM2 (column), excluding the learner's virtual volume reward. DB retains the original DB role payoffs; BD reverses that pair. All other entries use role pooling, as defined in (A3). Optimistic initialization, $K = 1,000$, $T = 10^6$, $\sigma = 5$, and $\kappa = 0$. Shading marks Nash equilibria at unrounded means. ***: every unilateral stay-minus-switch gain has a positive simultaneous 99% lower bound under this convention. Stars below a pair refer to the whole equilibrium profile. BB has no stars; its simultaneous 99% B-to-D deviation interval includes zero. Corresponding MCSEs are in Table A3.

The intuition follows the direction of the learning reward. Positive execution rewards lead to tougher competition: mutual *Biased* has a lower ask and lower profit than BB. In Table 6, the average quote falls from 5.015 under BB to 4.325 under DD. Quoted and realized spreads fall, while the expected trade probability increases from 29.07% to 33.57%. Full-horizon profit falls from 0.2822 to 0.2197 per market maker. A reward that encourages execution can therefore be privately attractive against a given rival even though mutual adoption leaves both market makers worse off than mutual *Basic*.

4.2 A Small Execution Penalty

An execution penalty may also lead to a mutual *Biased* equilibrium. At $\gamma = -0.2$, DD and CC are supported equilibria in Table 5. Each market maker earns approximately 0.2942 per episode under DD, above the 0.2822 under BB. At DD, switching to *Basic* reduces profit to 0.2884 for either market maker. Switching to *Clever* reduces profit to approximately 0.1545 for either market maker. Neither market maker wants to leave mutual *Biased*.

Table 5: The negative-bias delegation game, $\gamma = -0.2$

MM1 / MM2	*Basic*	*Biased*	*Clever*
Basic	(0.2822, 0.2822)	(0.2884, 0.2834)	(0.0018, 0.1542)
Biased	(0.2834, 0.2884)	(0.2942, 0.2942) ***	(0.0019, 0.1545)
Clever	(0.1542, 0.0018)	(0.1545, 0.0019)	(0.0400, 0.0400) ***

Notes: Mean monetary profits per episode for MM1 (row) and MM2 (column), excluding the learner's virtual volume reward. DB retains the original DB role payoffs; BD reverses that pair. All other entries use role pooling, as defined in (A3). Optimistic initialization, $K = 1,000$, $T = 10^6$, $\sigma = 5$, and $\kappa = 0$. Shading marks Nash equilibria at unrounded means. ***: every unilateral stay-minus-switch gain has a positive simultaneous 99% lower bound under this convention. Stars below a pair refer to the whole equilibrium profile. BB has no stars; its simultaneous 99% B-to-D deviation interval includes zero. Corresponding MCSEs are in Table A3.

The intuition is again the direction of the learning reward. A small execution penalty softens competition, leading to a higher ask and higher profit in DD than in BB. The final-episode average quote rises from 5.015 to 5.189, and the expected trade probability falls from 29.07% to 27.97%. Both quoted and realized spreads increase. Market makers can therefore prefer an incentive-misaligned learner and sustain less competitive prices. The penalty enters learning only; it is not a payment subtracted from the reported monetary profits.

At the larger penalty $\gamma = -0.8$, a profitable unilateral deviation prevents DD from being an equilibrium, even though mutual *Biased* yields higher profit than BB. At this coefficient the learner becomes too conservative in making a market. A high payoff under mutual

Table 6: Market outcomes under mutual delegation

	BB	CC	$D_{+0.8}D_{+0.8}$	$D_{-0.2}D_{-0.2}$
Average quote	5.0151	2.8880	4.3252	5.1889
Quoted spread	3.0151	0.8880	2.3252	3.1889
Realized spread	1.9365	0.2302	1.3119	2.1626
Trade probability	29.07%	43.49%	33.57%	27.97%
Collusion index	0.792	0.052	0.601	0.829

Notes: The first four rows use the pre-update quotes and outcomes at episode $T = 10^6$. Average quote, quoted spread and trade probability average over all 1,000 experiments; realized spread averages only experiments trading at T. Average quote is the lowest market ask. Quoted spread subtracts the unconditional mean fundamental 2. Realized spread averages $a_T^{\min} - v_T$ only over experiments with a trade at T. Trade probability integrates over the client-valuation distribution conditional on each saved minimum ask, then averages across experiments. The collusion index uses full-horizon mean joint monetary profit. All experiments have optimistic initialization and zero data cost.

adoption is insufficient when a market maker can earn more by changing its own delegate.

4.3 Can *Biased* Delegation Be Sustained?

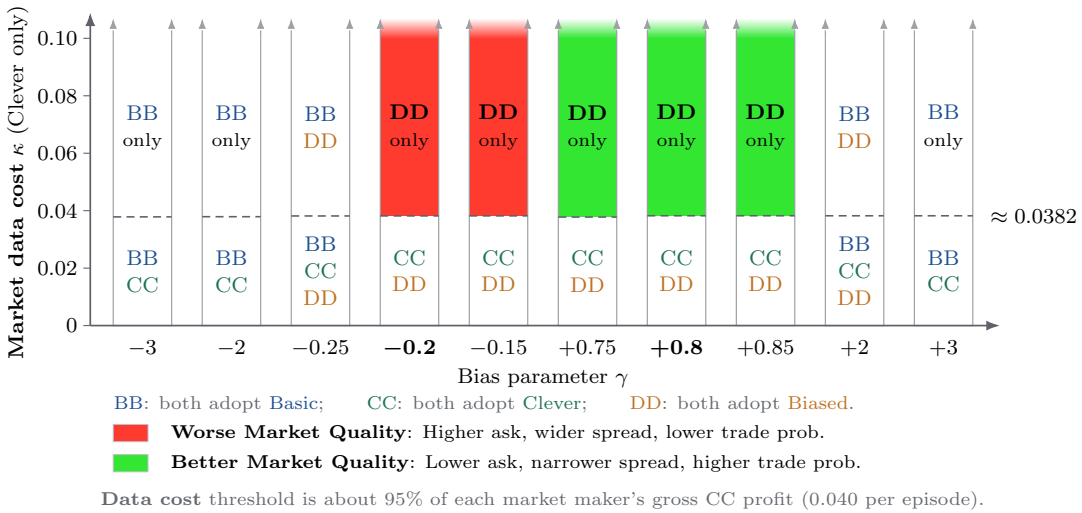
The focal outcomes also arise at neighboring reward coefficients. Figure 4 compares ten simulated values of γ, including the larger bonuses and penalties $+2$, $+3$, -2 and -3. At zero market-data cost, the profit comparisons yield DD at -0.25, -0.2, -0.15, $+0.75$, $+0.8$, $+0.85$ and $+2$. The recomputed simultaneous 95% deviation tests confirm DD at all these coefficients, including -0.15. At -3, -2 and $+3$, switching from *Biased* to *Basic* is profitable, so DD is not an equilibrium. Each coefficient defines a separate game; market-data costs vary continuously within that game.

Only a market maker choosing *Clever* pays the recurring data cost κ per episode. Deducting it from the market maker's payoff does not change the learner's reward or the prices within a given profile. The cutoff for CC is

$$\widehat{\kappa}_C(\gamma) = \min_{i\in\{1,2\}} \left\{ \widehat{u}_i(C, C; 0) - \max_{m\in\{B, D_\gamma\}} \widehat{u}_i(m, C; 0) \right\}. \tag{7}$$

CC is a weak equilibrium at the cutoff and disappears above it. For the focal coefficients, the cutoff is approximately 0.0382, about 95% of each market maker's gross CC profit of 0.0400 per episode. Above the cutoff, the payoff matrices have DD as their only pure equilibrium at $\gamma = -0.2$ and $+0.8$. At the negative coefficient, this leaves an equilibrium with even worse market quality than BB: a higher ask, wider spreads and a lower trade probability.

Figure 4: *Biased* delegation and market-data costs.



Notes: Each column is a separate $\{B, D_\gamma, C\}$ game at $\sigma = 5$, with optimistic initialization, $K = 1,000$ and $T = 10^6$. BB, CC and DD denote mutual adoption of *Basic*, *Clever* and *Biased*, respectively. Labels give pure equilibria from unrounded full-horizon mean monetary profits. Cost is continuous within each column, and only *Clever* pays it. Dashed lines are the exact CC cutoffs; CC remains a weak equilibrium at the cutoff. Open tops and upward arrows indicate that the upper equilibrium sets continue to all higher costs. Columns do not interpolate over γ, and mixed equilibria are omitted. Shading compares mutual DD with BB: worse means a higher final-episode average quote, wider spreads and lower expected trade probability; better means the reverse. Payoff labels and cutoffs use the symmetric estimates defined in Appendix A, including the DB-reference convention at $\gamma = +0.8$ and -0.2. DD at -0.15 passes the recomputed simultaneous 95% tests; an incentive interval for BB at -0.25 still includes zero. Exact cutoffs and statistical scope are reported in Table A4.

Market-data costs can therefore eliminate the competitive alternative without eliminating biased delegation. The available rewards and the cost of *Clever*'s market data determine which delegation profiles can be sustained. These results concern the specified three-action games at the tested coefficients; they do not characterize an optimal continuous reward

contract.

The paired statistical-deviation check retains DD and CC in both focal games, with optimistic initialization, $\sigma = 5$, and $\kappa = 0$. The smallest one-sided p-value for a deviation from BB is 0.0400 at $\gamma = +0.8$ and 0.1997 at $\gamma = -0.2$. Thus BB does not pass the nominal 25% screen in either reported game. At $\gamma = +0.8$, the DB-reference gain from switching from BB to *Biased* is 0.00275013 per episode, with MCSE 0.00156900 and a simultaneous 95% interval of $[-0.00180566, 0.00730592]$. At $\gamma = -0.2$, the corresponding gain is 0.00123362 per episode, with MCSE 0.00146314 and a simultaneous 95% interval of $[-0.00299951, 0.00546674]$. The simultaneous 95% bounds therefore cannot exclude BB in either game. These stability statements are distinct from the Nash inequalities at the unrounded mean payoffs, which select DD and CC in both games.

The auxiliary screen's conclusion for BB at the negative bias depends on the estimator. Pooling BD and DB instead gives an adoption gain of 0.00001635 and a one-sided p-value of 0.4948, so that the nominal 25% screen also retains BB. DD and CC pass the simultaneous 99% tests under both conventions. The evidence for these two equilibria is therefore stronger than the evidence for excluding BB.

With the *Basic–Biased–Clever* menu, sequential (Stackelberg) delegation selects mutual *Biased* in the two focal games, $\gamma = +0.8$ and $\gamma = -0.2$. Both use the baseline market and learning parameters: $\sigma = 5$, $\kappa = 0$, optimistic initialization, and full-horizon average monetary profit over $T = 10^6$ episodes. As above, the follower observes the leader's binding algorithm choice, and both commitments precede learning. Backward induction on the mean payoffs gives the unique subgame-perfect outcome DD regardless of which market maker leads. The follower chooses *Biased* after either a *Basic* or *Biased* commitment, and matches *Clever*. The leader prefers the outcome induced by *Biased* to those induced by *Basic* or *Clever*. At these two coefficients in the baseline setting, expanding the menu therefore changes the sequential outcome from BB to DD. This comparison does not establish sequential selection for other values of γ or σ. The follower's comparison between *Basic* and *Biased* after an off-path

Basic commitment has a simultaneous confidence interval containing zero. Thus the unique outcome follows from backward induction on the reported payoff matrices; the complete contingent strategy is less precisely determined by the statistical comparisons.

5 Economic and Policy Implications

The availability of a competitive AI-powered pricing algorithm does not ensure its adoption. Counterfactual updating allows *Clever* to learn the payoff from undercutting and produces near-competitive prices when mutually adopted. Yet, in the baseline *Basic–Clever* game, mutual *Basic* remains an equilibrium, sustains supracompetitive prices, and gives the market makers higher profits. These results separate competition in algorithm adoption from competition in quoted prices. The technology that improves the competitiveness of a mutual-adoption outcome can still be privately unattractive for a market maker to adopt alone.

Equilibrium selection depends on the economic environment and on market data costs. In the baseline *Basic–Clever* game, BB remains an equilibrium for every nonnegative *Clever* data fee, while a sufficiently high fee eliminates CC. The data-cost threshold is about 95% of each market maker's gross CC profit. At the selected strong-adverse-selection setting, $\sigma = 1$, *Basic* is strictly dominant even when market data are free. The comparisons show why advances in algorithm design and access to market data need to be considered together with market makers' incentives to adopt. They do not imply that the three simulated liquidity settings identify a continuous comparative-static relationship.

Strategic delegation can also reduce market quality further. A profit-seeking market maker may choose an incentive-misaligned *Biased* algorithm. A small execution penalty can soften competition and sustain mutual *Biased* with worse quotes and spreads, and higher monetary profits, than mutual *Basic*. A positive execution reward can instead lead to tougher competition and lower profits in DD. The sign and size of the reward distortion matter. Alignment between the learner's reward and the market maker's monetary profit does not by itself determine the competitiveness of the market outcome.

The algorithms do not communicate directly, but their learning is interdependent. How one algorithm learns affects the quotes, executions and rewards experienced by its rival. The rival's learning procedure, in turn, changes the private value of the first market maker's delegate. This interaction connects competition over algorithm adoption with competition in quoted prices.

These findings concern the available algorithms, the specified rewards, and the market makers' payoff criterion. The baseline *Basic–Clever* conclusions are robust to the reported payoff horizons and trading history. Sequential delegation selects BB for either leader in the baseline game ($\sigma = 5$, $\kappa = 0$), using full-horizon average monetary profits. This is not a selection result for every value of σ. With *Biased* available, the corresponding DD result concerns $\gamma = +0.8$ and $\gamma = -0.2$ under the same baseline parameters; it does not imply that every biased-delegation outcome has the same robustness.

6 Conclusion

Algorithm advances do not necessarily imply greater price competitiveness. In our delegation game framework, market makers choose and commit to the algorithms that quote and learn on their behalf. *Clever* can deliver competitive prices, yet market makers may not adopt it. In the baseline *Basic–Clever* game, BB is both payoff dominant and risk dominant, and market-data costs can eliminate the mutual-*Clever* equilibrium.

Strategic delegation links (mis)alignment to market outcomes. Delegation to misaligned *Biased* algorithms can further worsen market quality: market makers may adopt them in equilibrium, with worse quotes and spreads than under mutual *Basic*. Positive and negative execution rewards produce different market outcomes. The choice of a learning algorithm and its objective is therefore part of the strategic interaction between market makers, alongside the prices chosen by the algorithms themselves.

References

Asker, John, Chaim Fershtman, and Ariel Pakes. 2022. "Artificial Intelligence, Algorithm Design, and Pricing." *AEA Papers and Proceedings* 112: 452–456.

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

Banerjee, Snehal and Martin Szydlowski. 2025. "Trading against Algorithms: Price Dynamics and Risk Sharing in a Market with Q-learners." Working paper.

Brown, Zach Y. and Alexander MacKay. 2026. "Algorithmic Coercion with Faster Pricing." Working paper.

Brunnermeier, Markus K. 2026. "Artificial Intelligence and the Brave New World in Finance." Paper prepared for the Jackson Hole Economic Policy Symposium, Federal Reserve Bank of Kansas City.

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

Carissimo, Cesare, Fryderyk Falniowski, Siavash Rahimi, and Heinrich Nax. 2025. "Algorithmic Collusion Is Algorithm Orchestration." Working paper, arXiv:2508.14766.

Chen, Eric Olav, Alexis Ghersengorin, and Sami Petersen. 2024. "Imperfect Recall and AI Delegation." Working Paper 30-2024, Global Priorities Institute, University of Oxford.

Colliard, Jean-Edouard, Thierry Foucault, and Stefano Lovo. 2026. "Algorithmic Pricing and Liquidity in Securities Markets." *The Review of Financial Studies*.

Compte, Olivier. 2025. "Learned Collusion." Working paper, arXiv:2304.12647.

Cooper, Emery, Caspar Oesterheld, and Vincent Conitzer. 2025. "Characterising Simulation-Based Program Equilibria." *Proceedings of the AAAI Conference on Artificial Intelligence* 39 (13): 13735–13744.

Dou, Winston Wei, Itay Goldstein, and Yan Ji. 2025. "AI-Powered Trading, Algorithmic Collusion, and Price Efficiency." Working Paper 34054, National Bureau of Economic Research.

Dou, Winston Wei, Itay Goldstein, Zigang Li, and Liyan Yang. 2026. "The Limits of AI Trading." Working paper.

Dworczak, Piotr and Alex Smolin. 2026. "Robust Trust." Discussion Paper 21148, Centre for Economic Policy Research.

Fershtman, Chaim and Kenneth L. Judd. 1987. "Equilibrium Incentives in Oligopoly." *American Economic Review* 77 (5): 927–940.

Fershtman, Chaim, Kenneth L. Judd, and Ehud Kalai. 1991. "Observable Contracts: Strategic Delegation and Cooperation." *International Economic Review* 32 (3): 551–559.

Fudenberg, Drew and Annie Liang. 2025. "Friend or Foe: Delegating to an AI Whose Alignment Is Unknown." Working paper.

Glosten, Lawrence R. and Paul R. Milgrom. 1985. "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders." *Journal of Financial Economics* 14 (1): 71–100.

Guarino, Antonio, Philippe Jehiel, and James Symons-Hicks. 2025. "Q-Learning and Algorithmic Market Making: Loss-free, Collusive, or Competitive Prices?" Discussion Paper 20461, Centre for Economic Policy Research.

Holmström, Bengt Robert. 1977. *On Incentives and Control in Organizations*. Ph.D. dissertation, Stanford University.

Klein, Timo. 2021. "Autonomous Algorithmic Collusion: Q-learning under Sequential Pricing." *The RAND Journal of Economics* 52 (3): 538–558.

Oesterheld, Caspar. 2019. "Robust Program Equilibrium." *Theory and Decision* 86 (1): 143–159.

Rajan, Raghuram G. 2026. "Discussion of Brunnermeier's *Artificial Intelligence and the Brave New World in Finance*." Discussion slides, Jackson Hole Economic Policy Symposium, Federal Reserve Bank of Kansas City.

Sangiorgi, Francesco, Ivan Gufler, and Emanuele Tarantino. 2026. "(Deep) Learning to Trade: Co-learning Bias and Market Outcomes." Working paper.

Tennenholtz, Moshe. 2004. "Program Equilibrium." *Games and Economic Behavior* 49 (2): 363–373.

Vickers, John. 1985. "Delegation and the Theory of the Firm." *The Economic Journal* 95 (Supplement): 138–147.

Xu, Yongzhi. 2025. "Tacit Collusion by Pricing Algorithm with Rule-Based Rivals." Working paper.

Yang, Hao. 2024. "AI Coordination and Self-Fulfilling Financial Crises." Working paper.

Appendix

A Numerical Implementation and Statistical Inference

This appendix presents the numerical specification and the definitions used for payoffs, market outcomes and statistical tests. The optimistic baseline uses the market parameters and learning hyperparameters of Colliard et al. (2026).

A.1 Numerical Specification

Table A1: Baseline numerical implementation

Item	Specification
Asset payoff and client	$v \in \{0, 4\}$ with equal probabilities, client valuation $v + L$ and independent $L \sim N(0, 5^2)$.
Quote grid	139 asks, $\mathcal{A} = \{1.1, 1.2, \ldots, 14.9\}$, with tick size 0.1.
Learning and exploration	$\alpha = 0.01$, $\epsilon_t = \exp(-\beta t)$, $\beta = 8 \times 10^{-5}$; uniform exploratory quotes.
Initial values	Independent $U[3, 6]$ draws by dealer and quote, shared across profiles for that dealer.
Exploitation	Uniform random choice among exact Q-value maximizers.
Counterfactual updating	Observable-event identification mask, actual fundamental cost, and counterfactual weight $\rho = 1$.
Horizon and replications	$T = 10^6$ learning episodes and $K = 1{,}000$ independent matched experiments.
Midstream design	500,000 shared BB-history episodes and 500,000 continuation episodes, with inherited Q-values and exploration clocks.
Frozen evaluation	100,000 independent episodes with exploration and Q-value updates disabled.
Monetary accounting	Realized profit $(a_{it} - v_t)x_{it}$, with half execution in traded ties.
Data access	A recurring fee κ for *Clever* only, deducted from market maker profit after simulation.
Numerical conventions	No baseline Q-zeroing, exact maximizer ties, and Nash payoff tolerance 10^{-8}.
Randomness and inference	Master seed 20260921; experiment-level pairing; 2,000 saved bootstrap resamples for simultaneous inference.

Notes: The baseline environment and learning parameters follow the implementation in Colliard et al. (2026). Guarino, Jehiel and Symons-Hicks study a different economic environment and calibration. Alternative starts and numerical sensitivities are separate treatments. A numerical Nash tolerance is not an approximation to a Q-value maximizer and is not a confidence interval.

A.2 Competitive and Joint-Profit Benchmarks

Fix a market environment (v_L, v_H, σ). With equal prior probabilities for the two asset values, the client's probability of buying at an ask a is

$$D_\sigma(a) = \frac{1}{2}\Phi\left(\frac{v_L - a}{\sigma}\right) + \frac{1}{2}\Phi\left(\frac{v_H - a}{\sigma}\right),$$

where Φ is the standard normal CDF. The continuous competitive price equates the ask to the asset's expected value conditional on a purchase:

$$a^* = \frac{v_L\Phi\left(\frac{v_L - a^*}{\sigma}\right) + v_H\Phi\left(\frac{v_H - a^*}{\sigma}\right)}{\Phi\left(\frac{v_L - a^*}{\sigma}\right) + \Phi\left(\frac{v_H - a^*}{\sigma}\right)}.$$

We solve this equation for each market environment. At the baseline $v_L = 0$, $v_H = 4$ and $\sigma = 5$, it gives $a^* \simeq 2.6851$.

The discrete benchmark uses the actual quote grid. Expected joint monetary profit at a common ask a is

$$J_\sigma(a) = \frac{1}{2}\Phi\left(\frac{v_L - a}{\sigma}\right)(a - v_L) + \frac{1}{2}\Phi\left(\frac{v_H - a}{\sigma}\right)(a - v_H). \tag{A1}$$

Against a rival ask a, a market maker choosing b earns expected one-episode profit

$$h_\sigma(b, a) = \begin{cases} J_\sigma(b), & b < a, \\ J_\sigma(a)/2, & b = a, \\ 0, & b > a. \end{cases}$$

A lower quote wins the whole executed order; an equal quote shares it; a higher quote does not trade. We enumerate the grid and retain a symmetric quote a exactly when

$$\frac{J_\sigma(a)}{2} \geq h_\sigma(b, a) \qquad \text{for every } b \in \mathcal{A}.$$

At the baseline, both 2.7 and 2.8 satisfy these inequalities. We use the higher quote, $a^{\text{Nash}} = 2.8$, for reporting. Finite tick size permits positive profit at this competitive quote.

The profit benchmarks used in the collusion index are

$$\Pi^{\text{Nash}} = J_\sigma(a^{\text{Nash}}), \qquad \Pi^M = \max_{a \in \mathcal{A}} J_\sigma(a).$$

At the baseline, $\Pi^{\text{Nash}} = J_5(2.8) = 0.0459347$ and $\Pi^M = J_5(6.5) = 0.7002735$. Both are joint profits; each market maker earns half under symmetric quoting. The benchmarks use monetary profit, including in comparisons with *Biased*; the learner's virtual reward term γx does not enter them.

A.3 Market Outcomes and Time Windows

Delegation payoffs are expected average monetary profits over the complete learning path. Market quotes and spreads follow the final-episode convention of Colliard et al. (2026, Section 3.3). Let $a^{\min}_{T,k}$ be experiment k's lowest ask posted before the last update, $v_{T,k}$ the realized asset value, and $I_{T,k}$ the indicator of a market trade. We calculate

$$\begin{aligned} \overline{a}_T &= \frac{1}{K}\sum_{k=1}^{K} a^{\min}_{T,k}, \\ QS_T &= \overline{a}_T - \overline{v}, \\ RS_T &= \frac{\sum_{k=1}^{K} I_{T,k}(a^{\min}_{T,k} - v_{T,k})}{\sum_{k=1}^{K} I_{T,k}}. \end{aligned}$$

Average quote includes experiments without execution. Realized spread includes only actual trades and is undefined if there are none. The two spreads are ask-side half-spreads; we retain the shorter names quoted spread and realized spread. The market minimum ask differs from an individual algorithm's own posted ask in a dynamics figure.

Trade probability integrates over the client valuation at each saved last-episode quote:

$$\overline{p}_T = \frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{2}\Phi\left(\frac{v_L - a^{\min}_{T,k}}{\sigma}\right) + \frac{1}{2}\Phi\left(\frac{v_H - a^{\min}_{T,k}}{\sigma}\right)\right],$$

where Φ is the standard normal CDF. This is the model-implied execution probability at the learned quotes, averaged across experiments. It integrates out the current liquidity and

fundamental shocks. It is distinct from the realized fraction of experiments that happen to trade in the single last episode. Realized spread continues to use actual traded observations.

The collusion index uses gross joint monetary profit over the full horizon,

$$\Pi = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\pi_{1t} + \pi_{2t})\right], \qquad \Delta^D = \frac{\Pi - \Pi^{\text{Nash}}}{\Pi^M - \Pi^{\text{Nash}}}.$$

Appendix A.2 defines both profit benchmarks. The index is a descriptive profit normalization and does not identify collusive intent. Its learning-path window differs from the final-episode market statistics above.

A.4 Counterfactual Updating

The following observable-event rule implements *Clever*'s economic understanding of undercutting. It uses the best quote, the trade and allocation outcomes, and the ex post asset value on traded episodes. Neither the client's valuation nor a losing rival's higher ask is an input.

Write $b_t = a_t^{\min}$ and let $I_t = \mathbf{1}\{b_t \leq \widetilde{v}_t^C\}$ denote the market's trade outcome. The set of identified quotes is

$$\mathcal{I}_{it} = \begin{cases} \{a \in \mathcal{A} : a \geq b_t\}, & I_t = 0, \\ \{a \in \mathcal{A} : a \leq b_t\}, & I_t = 1 \text{ and } i \text{ is the unique winner}, \\ \mathcal{A}, & I_t = 1 \text{ and } i \text{ loses or shares the order}. \end{cases} \tag{A2}$$

If no trade occurs, every weakly higher quote would also fail to execute. If market maker i wins uniquely, every lower quote would execute, but execution at a higher quote cannot be inferred from the observations. If it loses or shares an executed order, a lower ask would win the whole order, an ask equal to b_t would share it, and a higher ask would lose. The rule therefore uses neither the client's private valuation nor the other market maker's exact ask when that ask lies above b_t.

For $a \in \mathcal{I}_{it}$, the identified profit target is

$$\pi_{it}^{\mathrm{cf}}(a) = \begin{cases} 0, & I_t = 0, \\ a - \widetilde{v}_t, & I_t = 1,\ a < b_t, \\ s_{it}(b_t - \widetilde{v}_t), & I_t = 1,\ a = b_t, \\ 0, & I_t = 1,\ a > b_t, \end{cases}$$

where $s_{it} = 1$ if i is the unique winner and $s_{it} = 1/2$ otherwise.

The chosen quote is always included, its target equals the realized profit, and it is updated once according to (3). No identified alternative is assigned a larger learning weight than the chosen quote. The counterfactual weight is $\rho = 1$.

A.5 Monte Carlo Standard Errors and Equilibrium Tests

Let $Y_{i,k}(a, b)$ denote the original average monetary profit of market maker i in matched experiment k when MM1 uses a and MM2 uses b. We impose the model's exchange symmetry on the estimator through

$$Z_{1,k}(a, b) = \frac{Y_{1,k}(a, b) + Y_{2,k}(b, a)}{2}, \qquad Z_{2,k}(a, b) = Z_{1,k}(b, a).$$

For mutual adoption this averages the two dealer payoffs within each experiment. We use $\widehat{u}_{i,k} = Z_{i,k}$ except for the BD and DB entries in the focal biased games of Tables 4 and 5. In each game, the original DB path supplies the reference payoffs:

$$\begin{aligned} &\widehat{u}_{1,k}(D, B) = Y_{1,k}(D, B), \\ &\widehat{u}_{1,k}(B, D) = Y_{2,k}(D, B), \qquad \widehat{u}_{2,k}(a, b) = \widehat{u}_{1,k}(b, a). \end{aligned} \tag{A3}$$

All other entries retain $Z_{i,k}$. This reporting convention keeps the original two DB payoffs and reflects them into BD; it does not average these two cells across orientations. The same experiment-level vectors are used for the corresponding means, MCSEs and deviation tests. The MCSE is the sample standard deviation of the relevant K experiment observations divided by $\sqrt{K}$. A unilateral adoption comparison uses the within-experiment difference

between the relevant two profiles. Its MCSE retains this matching. The density and CDF figures in Appendix B describe variation in experiment-average profit; they are not confidence intervals for mean payoffs.

Simultaneous tests use 2,000 bootstrap resamples of whole matched experiments. We compute the 95% and 99% bounds from the experiment vectors defined above, including the DB-reference entries at $\gamma = +0.8$ and -0.2, using inference seed 20261011. The critical value is a quantile of the maximum absolute standardized statistic over the specified contrast family. Each displayed two-action game has twelve indexed pairwise payoff contrasts across the two market makers; each three-action game has seventy-two. These are within-matrix families, separate from the broader parameter-grid and decomposition families in the original exports. Every shaded Nash profile in the five payoff matrices passes the 99% simultaneous checks. Table A2 records the critical values.

In the payoff matrices, *, ** and *** indicate that all unilateral stay-minus-switch gains have positive simultaneous lower confidence bounds at the 90%, 95% and 99% levels, respectively. The strongest supported level is shown on a separate line centered beneath the whole payoff pair. These stars summarize support for the Nash inequalities, rather than tests of monetary profit against zero or the separate statistical-stability screen below. A profile without stars need not admit a statistically detected profitable deviation. All nine shaded profiles in the five matrices qualify for *** under their respective 99% bounds. Table A3 gives the numerical MCSEs.

We identify Nash equilibria by applying the unilateral-deviation inequalities to the unrounded mean payoffs. A profile passes the simultaneous tests when every lower bound on its stay-minus-switch gains is nonnegative, allowing for the numerical tolerance 10^{-8}. A simultaneous upper bound below -10^{-8} for any gain establishes a profitable deviation. When an incentive interval includes zero, the data do not determine its sign. The numerical tolerance is separate from sampling uncertainty and is not used to select Q-value maximizers.

Table A2: Simultaneous critical values for the focal delegation games

Game	Contrasts	95%	99%
Basic–Clever, $\sigma = 5$	12	2.3784	2.9515
Basic–Clever, $\sigma = 1$	12	2.4001	2.8933
Basic–Clever, $\sigma = 15$	12	2.4469	3.0133
Basic–Biased–Clever, $\gamma = +0.8$	72	2.9036	3.4419
Basic–Biased–Clever, $\gamma = -0.2$	72	2.8932	3.4047

Notes: All games use optimistic initialization and full-horizon monetary profit, with $\sigma = 5$ in the two three-action games. Critical values use 2,000 whole-experiment resamples of the specified payoff vectors, with the DB-reference entries in (A3) at $\gamma = +0.8$ and -0.2. They apply separately within each specified game and payoff criterion. The 99% lower bounds are positive for every stay-minus-switch gain at the shaded Nash profiles in the five payoff matrices. Coverage is not joint across values of γ, σ, or evaluation windows.

Statistical stability. As an additional check, we use the no-profitable-statistical-deviation criterion in Section 6.1 and Online Appendix OA.5 of Colliard et al. (2026). Their test compares full-learning-path average profits from independent samples using a one-sided Welch test, and retains a profile when the smallest unilateral-deviation p-value is at least 0.25. Our profiles share experiment-level random streams, so we retain this pairing rather than treat their payoff samples as independent. For a deviation from m to (m'_i, m_{-i}), define $d_{i,k} = \widehat{u}_{i,k}(m'_i, m_{-i}; 0) - \widehat{u}_{i,k}(m; 0)$. With $\overline{d}_i$ and s_{d_i} denoting the sample mean and standard deviation of these K paired differences, we calculate

$$t_i(m'_i; m) = \frac{\overline{d}_i}{s_{d_i}/\sqrt{K}}, \qquad p_i(m'_i; m) = 1 - F_{K-1}(t_i(m'_i; m)),$$

where F_{K-1} is the Student t cumulative distribution function. The null is equality of expected profits against the one-sided alternative that the deviation raises profit. A pure profile is retained when $\min_{i, m'_i \neq m_i} p_i(m'_i; m) \geq 0.25$. As in their screen, these auxiliary tests have no adjustment for multiple comparisons. They use the existing $K = 1,000$ full-horizon experiment outcomes and do not require additional market simulations or bootstrap resampling. Retention means that no profitable deviation is detected at this threshold; it is

Table A3: Monte Carlo precision of the delegation payoff estimates

Game	(a, b)	Mean payoff	MCSE
$\{B, C\}$, $\sigma = 5$	(B, B)	0.282172	0.001019
	(B, C)	0.001821	0.000004
	(C, B)	0.154153	0.000023
	(C, C)	0.040043	0.000009
$\{B, C\}$, $\sigma = 1$	(B, B)	0.020306	0.000259
	(B, C)	0.007094	0.000002
	(C, B)	0.010447	0.000002
	(C, C)	0.003609	0.000047
$\{B, C\}$, $\sigma = 15$	(B, B)	0.897399	0.005713
	(B, C)	0.008820	0.000007
	(C, B)	0.201618	0.000035
	(C, C)	0.047397	0.000014
$\{B, D, C\}$, $\gamma = +0.8$	(B, D)	0.031044	0.001454
	(D, B)	0.284922	0.001188
	(D, D)	0.219703	0.001864
	(D, C)	-0.001518	0.000117
	(C, D)	0.092086	0.000488
$\{B, D, C\}$, $\gamma = -0.2$	(B, D)	0.288411	0.000938
	(D, B)	0.283406	0.001063
	(D, D)	0.294249	0.000864
	(D, C)	0.001863	0.000004
	(C, D)	0.154470	0.000023

Notes: Each row reports MM1's mean monetary profit when it uses a and MM2 uses b, and the corresponding MCSE. MM2's payoff and MCSE at (a, b) equal MM1's at (b, a). All estimates use $K = 1,000$ matched experiments, full-horizon average profit over $T = 10^6$ episodes, optimistic initialization, and zero data cost. The two three-action games have $\sigma = 5$; their B/C entries are identical to the baseline B/C entries above and are not repeated. Except for the BD and DB rows in the two biased games, defined in (A3), payoffs are role-pooled within each experiment. All MCSEs use the corresponding 1,000 experiment observations, with no additional independent observations from mirroring. Means and MCSEs are rounded to six decimal places here; main matrices report four-decimal means and equilibrium-test stars.

not a statistical confirmation of every Nash inequality. The main results continue to report the simultaneous bounds defined above.

A.6 Cost Thresholds at Selected Reward Coefficients

Table A4 records the numerical thresholds underlying Figure 4. These comparisons deduct the subscription cost only from a market maker that selects *Clever*; no additional price-learning simulations are needed.

Table A4: *Biased* delegation and market-data costs at selected reward coefficients

γ	NE at $\kappa = 0$	NE above cutoff	$\widehat{\kappa}_C$
-3	BB, CC	BB only	0.03784
-2	BB, CC	BB only	0.03791
-0.25	BB, CC, DD	BB, DD	0.03816
-0.2	CC, DD	DD only	0.03818
-0.15	CC, DD	DD only	0.03819
$+0.75$	CC, DD	DD only	0.03778
$+0.8$	CC, DD	DD only	0.03822
$+0.85$	CC, DD	DD only	0.03822
$+2$	BB, CC, DD	BB, DD	0.03822
$+3$	BB, CC	BB only	0.03822

Notes: Each row is a separate optimistic $\{B, D_\gamma, C\}$ game with $K = 1,000$, $T = 10^6$, and liquidity shock $\sigma = 5$. Equilibrium sets use unrounded full-horizon mean monetary profits. The upper region is $\kappa > \widehat{\kappa}_C$, with no upper cost limit. Only *Clever* pays the market-data cost. Cutoffs are calculated before rounding. The cost intervals solve the affine unilateral-deviation inequalities for every nonnegative cost. CC remains a weak equilibrium at the exact cutoff in (7). The simultaneous 95% incentive intervals use the specified experiment-level estimates separately for each reward coefficient, including (A3) at $\gamma = +0.8$ and -0.2. DD at -0.15 passes the 95% tests, whereas an incentive interval for BB at -0.25 includes zero.

A.7 Trading History and Initialization Comparisons

The midstream design begins with 500,000 shared *Basic–Basic* episodes, followed by 500,000 continuation episodes. Alternative delegates inherit the same history-specific Q-values and exploration clock. The market makers' continuation payoffs exclude pre-adoption profit and

average across inherited histories. Final-10,000, final-20,000 and final-50,000-episode criteria evaluate saved terminal windows of the baseline paths.

Appendix B also reports the completed zero, analytical, supervised, and supervised-with-lower-exploration treatments. These are separate starting-value and learning-schedule checks, not additional observations for the optimistic baseline. Each distribution contains 1,000 experiment averages. The initial-delegation figures use all 10^6 episodes; the midstream figure uses only its 500,000 continuation episodes.

B Distributions of Average Profits

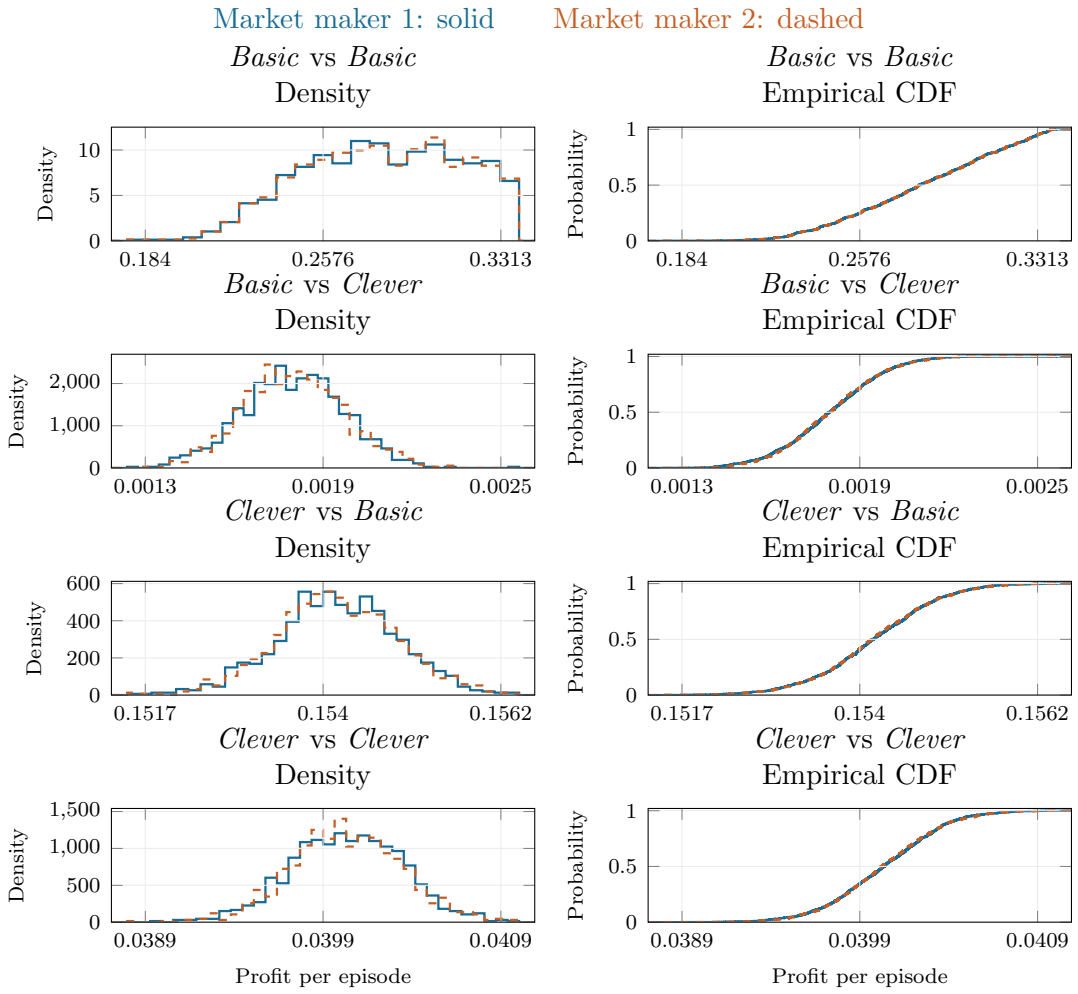
Each observation in this appendix is one market maker's average monetary profit over a complete learning path. Table A5 reports both roles separately. Its means, MCSEs and quantiles describe the original, unpooled observations, rather than the symmetric delegation estimates used in the payoff matrices. Figure A1 pairs density estimates with empirical CDFs for the baseline optimistic initialization, with both market makers shown separately. Figures A2–A6 use the same display for midstream adoption and alternative initializations. The distributions concern experiment averages, not individual trading-episode rewards. The midstream distributions use only post-adoption continuation profits; all other figures use the complete initial-delegation learning horizon.

Table A5: Full-path profit variation across experiments

Profile	MM	Mean	MCSE	10th percentile	Median	90th percentile
BB	1	0.28215	0.001020	0.23929	0.28224	0.32643
BB	2	0.28219	0.001019	0.23883	0.28238	0.32626
BC	1	0.00182	0.000006	0.00160	0.00182	0.00204
BC	2	0.15415	0.000024	0.15319	0.15413	0.15514
CB	1	0.15416	0.000024	0.15316	0.15415	0.15512
CB	2	0.00182	0.000005	0.00161	0.00182	0.00205
CC	1	0.04004	0.000010	0.03963	0.04005	0.04043
CC	2	0.04004	0.000010	0.03966	0.04003	0.04044

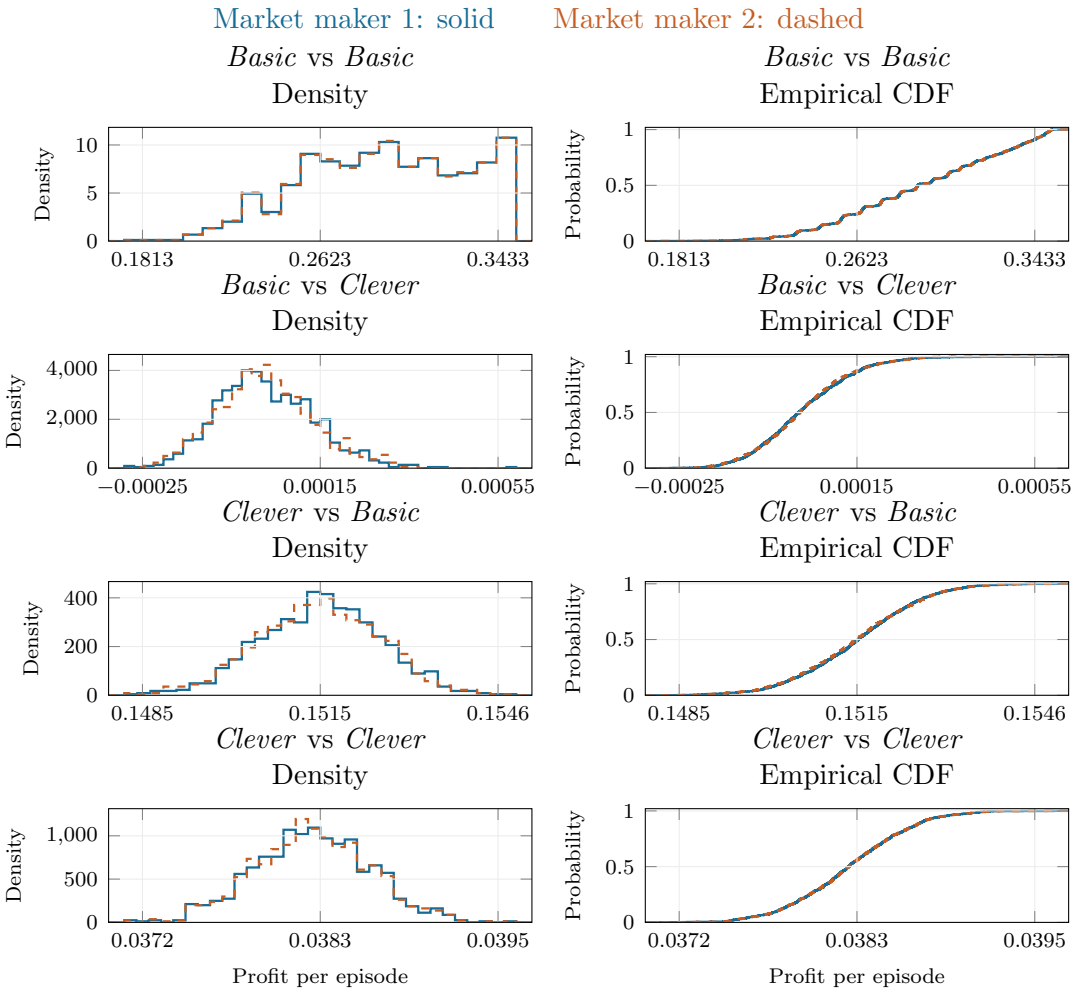
Notes: Baseline optimistic initialization, $\sigma = 5$, $\kappa = 0$, $K = 1{,}000$, and $T = 10^6$. Quantiles are empirical quantiles of the saved experiment-average profits. Dealer roles are retained. Monte Carlo standard errors measure uncertainty in the mean, whereas the quantiles describe dispersion across learning paths.

Figure A1: Baseline optimistic initialization: cross-experiment profit distributions.



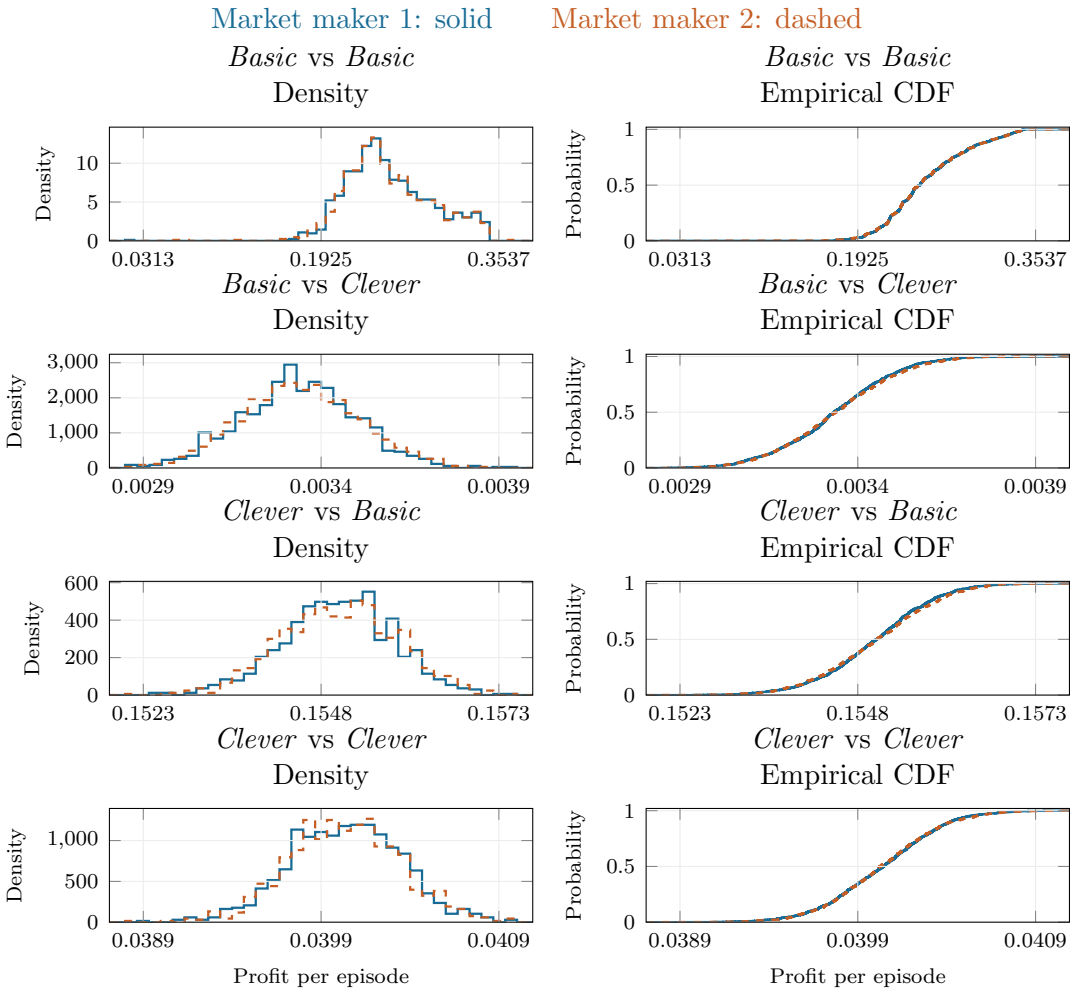
Notes: Full-path profit over $T = 10^6$ episodes with optimistic $U[3, 6]$ initial Q-values. Each series retains 1,000 experiment averages. Rows compare the dealer's algorithm with its rival's: the *Basic–Clever* row uses MM1 in BC and MM2 in CB; the *Clever–Basic* row uses MM1 in CB and MM2 in BC. Histograms integrate to one, using common bins for the two roles and the Freedman–Diaconis rule with 20–60 bins. Empirical CDFs retain every observation. Horizontal scales differ by row. These are outcome distributions, not confidence intervals.

Figure A2: Midstream adoption after a *Basic–Basic* history: cross-experiment profit distributions.



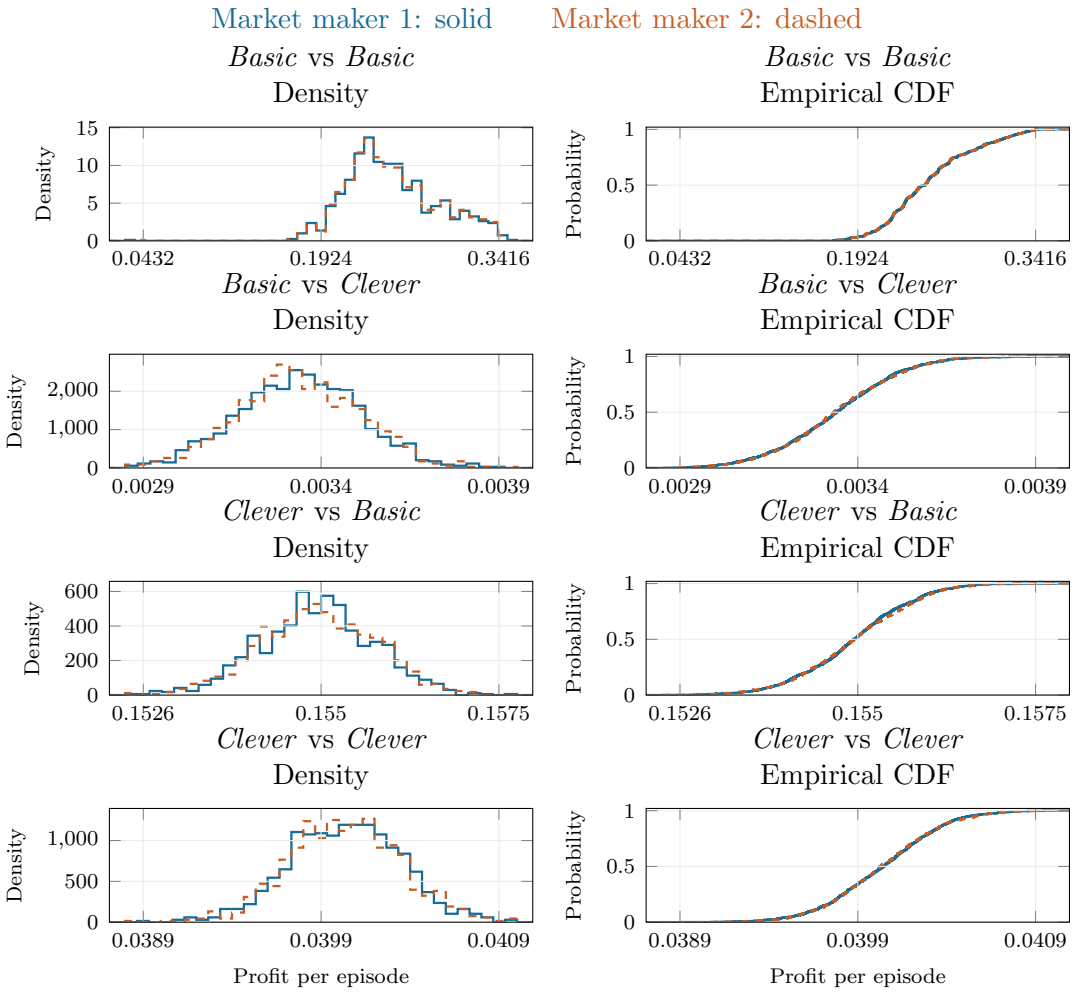
Notes: Profit per continuation episode over 500,000 episodes after a shared 500,000-episode *Basic–Basic* history. Initial Q-values and exploration clocks are inherited. Pre-adoption profits are excluded. Each series retains 1,000 experiment averages. Rows compare the dealer's algorithm with its rival's: the *Basic–Clever* row uses MM1 in BC and MM2 in CB; the *Clever–Basic* row uses MM1 in CB and MM2 in BC. Histograms integrate to one, using common bins for the two roles and the Freedman–Diaconis rule with 20–60 bins. Empirical CDFs retain every observation. Horizontal scales differ by row. These are outcome distributions, not confidence intervals.

Figure A3: Zero initialization: cross-experiment profit distributions.



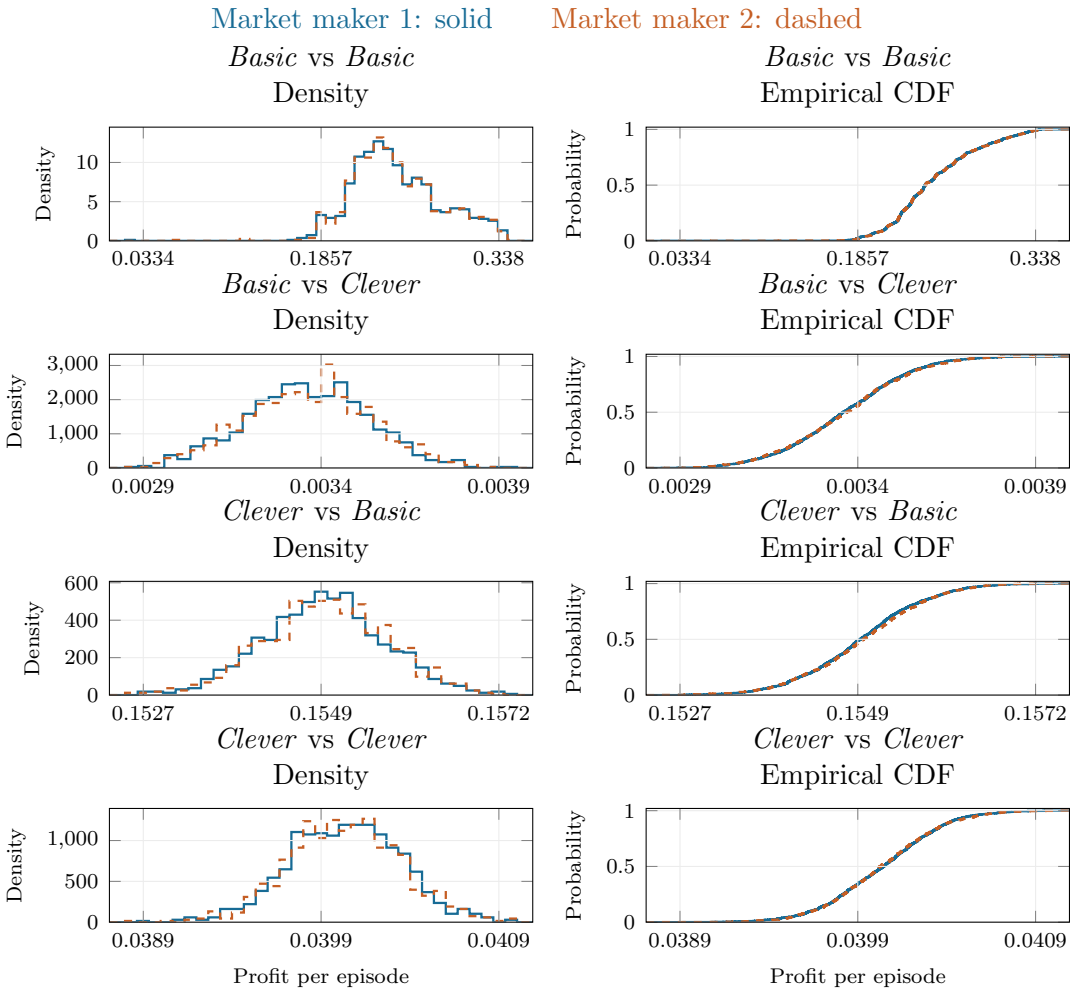
Notes: Full-path profit over $T = 10^6$ episodes with zero initial Q-values. Each series retains 1,000 experiment averages. Rows compare the dealer's algorithm with its rival's: the *Basic–Clever* row uses MM1 in BC and MM2 in CB; the *Clever–Basic* row uses MM1 in CB and MM2 in BC. Histograms integrate to one, using common bins for the two roles and the Freedman–Diaconis rule with 20–60 bins. Empirical CDFs retain every observation. Horizontal scales differ by row. These are outcome distributions, not confidence intervals.

Figure A4: Analytical initialization: cross-experiment profit distributions.



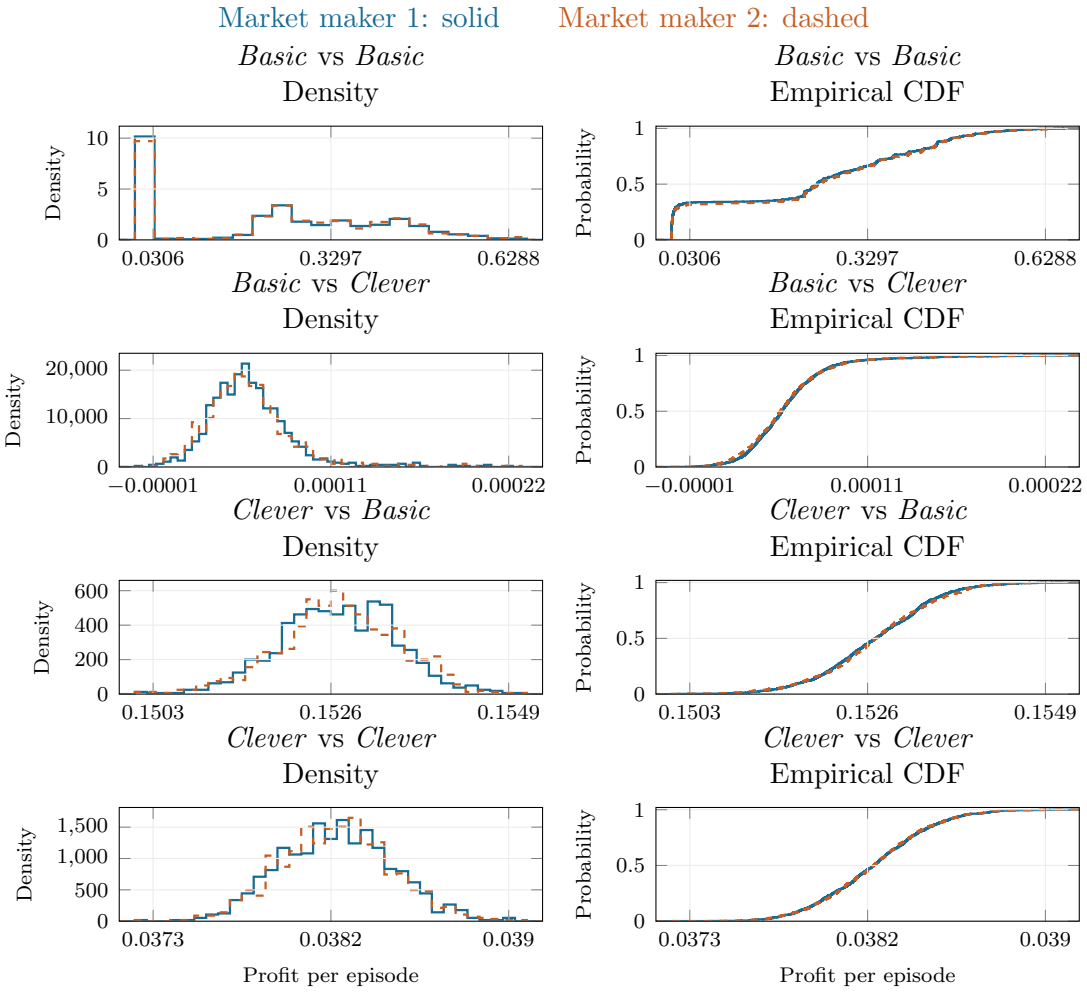
Notes: Full-path profit over $T = 10^6$ episodes with the analytically initialized Q-values of the robustness treatment. Each series retains 1,000 experiment averages. Rows compare the dealer's algorithm with its rival's: the *Basic–Clever* row uses MM1 in BC and MM2 in CB; the *Clever–Basic* row uses MM1 in CB and MM2 in BC. Histograms integrate to one, using common bins for the two roles and the Freedman–Diaconis rule with 20–60 bins. Empirical CDFs retain every observation. Horizontal scales differ by row. These are outcome distributions, not confidence intervals.

Figure A5: Supervised initialization: cross-experiment profit distributions.



Notes: Full-path profit over $T = 10^6$ episodes with the supervised initial Q-values of the robustness treatment. Each series retains 1,000 experiment averages. Rows compare the dealer's algorithm with its rival's: the *Basic–Clever* row uses MM1 in BC and MM2 in CB; the *Clever–Basic* row uses MM1 in CB and MM2 in BC. Histograms integrate to one, using common bins for the two roles and the Freedman–Diaconis rule with 20–60 bins. Empirical CDFs retain every observation. Horizontal scales differ by row. These are outcome distributions, not confidence intervals.

Figure A6: Supervised initialization with lower exploration: cross-experiment profit distributions.

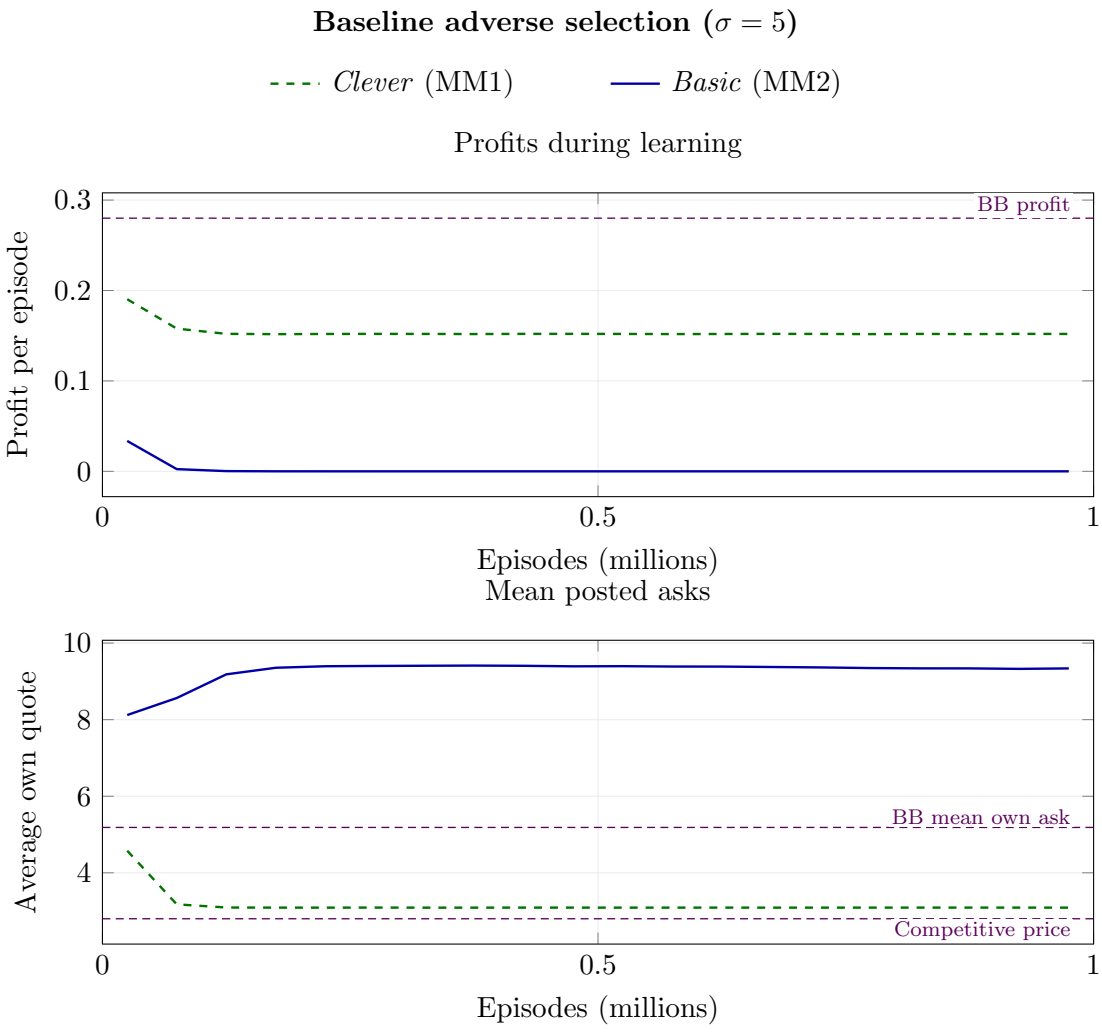


Notes: Full-path profit over $T = 10^6$ episodes with supervised initial Q-values and the recorded lower-exploration schedule. Each series retains 1,000 experiment averages. Rows compare the dealer's algorithm with its rival's: the *Basic–Clever* row uses MM1 in BC and MM2 in CB; the *Clever–Basic* row uses MM1 in CB and MM2 in BC. Histograms integrate to one, using common bins for the two roles and the Freedman–Diaconis rule with 20–60 bins. Empirical CDFs retain every observation. Horizontal scales differ by row. These are outcome distributions, not confidence intervals.

C Profit and Quote Paths

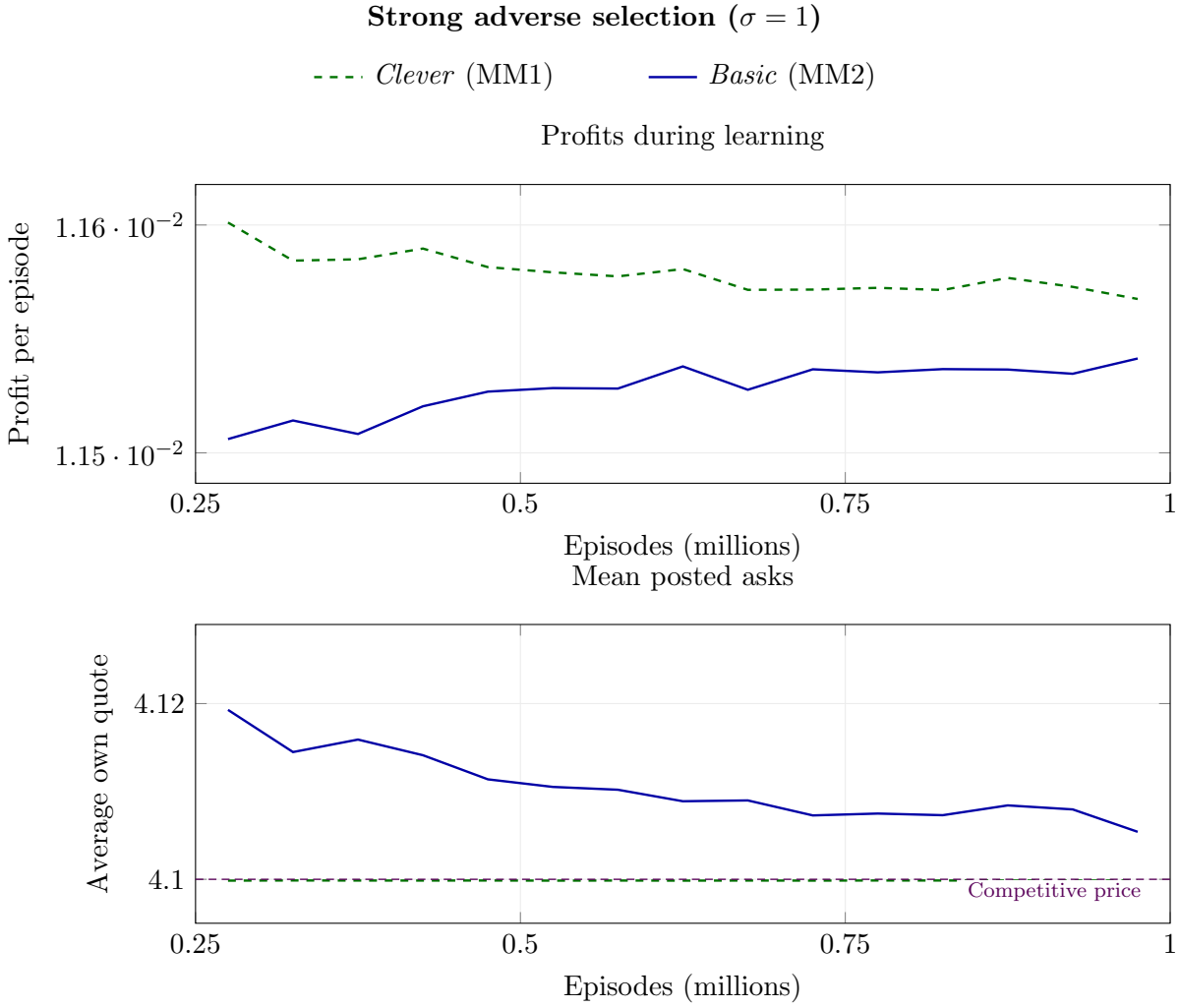
Figures A7–A9 show the asymmetric interaction used to interpret the *Basic–Clever* delegation game. In every panel, MM1 uses *Clever* and MM2 uses *Basic* in the same market. The three environments are baseline adverse selection ($\sigma = 5$), strong adverse selection ($\sigma = 1$), and weak adverse selection ($\sigma = 15$). Points average each algorithm's own monetary profit and posted ask within consecutive 50,000-episode blocks across 1,000 experiments. These within-CB comparisons explain the interaction; the adoption incentive compares the market maker's payoff across profiles holding the rival's delegate fixed.

Figure A7: MM1 plays *Clever* and MM2 plays *Basic*: baseline adverse selection ($\sigma = 5$).



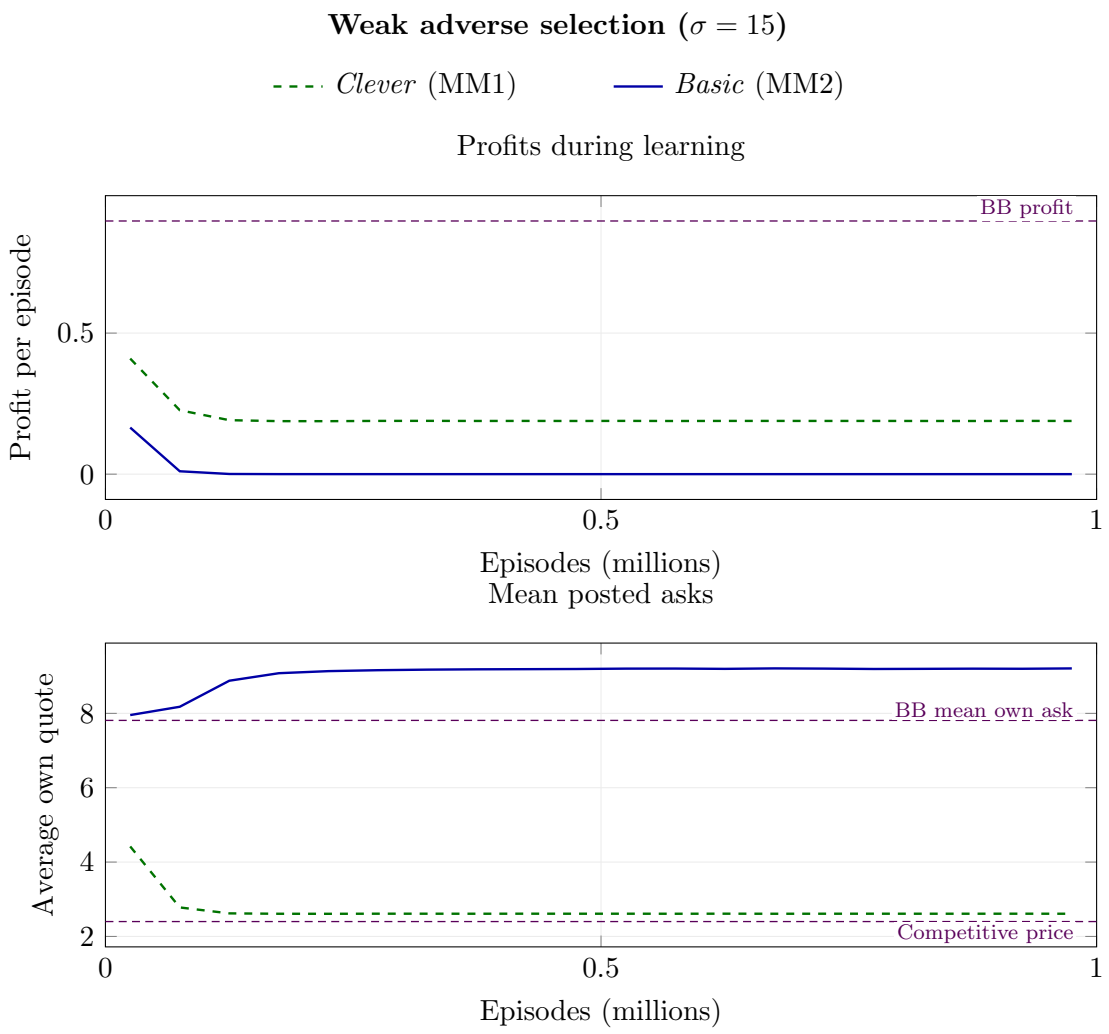
Notes: Means across 1,000 matched experiments in the same CB market. The upper panel shows each algorithm's monetary profit per episode; the lower panel shows its own posted ask. Green dashed is *Clever* (MM1), and blue solid is *Basic* (MM2). Both panels use all 10^6 episodes, with one point per 50,000-episode block. The competitive price is 2.8. Purple dashed lines show the BB profit reference (rounded to 0.28), BB mean own ask (5.1852), and competitive price.

Figure A8: MM1 plays *Clever* and MM2 plays *Basic*: strong adverse selection ($\sigma = 1$).



Notes: Means across 1,000 matched experiments in the same CB market. The upper panel shows each algorithm's monetary profit per episode; the lower panel shows its own posted ask. Green dashed is *Clever* (MM1), and blue solid is *Basic* (MM2). Both panels show late episodes, from 250,000 onward, to distinguish the nearby paths; each point covers 50,000 episodes. The competitive price is 4.1. The BB full-horizon references, profit 0.0203 and mean own ask approximately 4.64, lie outside these enlarged scales. Delegation payoffs continue to use all 10^6 episodes.

Figure A9: MM1 plays *Clever* and MM2 plays *Basic*: weak adverse selection ($\sigma = 15$).



Notes: Means across 1,000 matched experiments in the same CB market. The upper panel shows each algorithm's monetary profit per episode; the lower panel shows its own posted ask. Green dashed is *Clever* (MM1), and blue solid is *Basic* (MM2). Both panels use all 10^6 episodes, with one point per 50,000-episode block. The competitive price is 2.4. Purple dashed lines show BB full-horizon profit, BB mean own ask, and the competitive price.