The Q-Tables Remember: Persistence of Learned Collusion After Regulatory Disruption

Montaha Ghabri*
Tunis Business School (TBS)
Tunis, Tunisia
montaha.ghabri@tbs.u-tunis.tn

Abstract

Independent reinforcement learning agents deployed in competitive markets can converge to tacitly collusive pricing without explicit coordination. While the emergence of algorithmic collusion is well-documented, its stability once learned remains an open question of direct relevance to antitrust enforcement. This paper tests four regulatory-style interventions applied after Q-learning agents have converged to collusive pricing in a repeated Bertrand duopoly: forced competitive pricing ($k \in \{50, 100\}$ periods), an exploration shock ($\varepsilon = 0.5$), and a memory reset. Replicating Calvano et al. [4], our single baseline run yields an equilibrium price of 1.887 and a Profit Gain Index of $\Delta = 0.991$. In this experimental run, all four interventions reduce equilibrium prices but do not restore competitive outcomes. Recovery rates range from 54.5% (forced pricing, both durations) to 81.8% (memory reset), indicating partial disruption. Two patterns are observed. First, forcing Nash-level pricing for 50 and 100 periods produces the same post-intervention equilibrium in this run, suggesting a possible threshold effect rather than a dose-response relationship. Second, the memory reset achieves the least disruption, consistent with an experienced firm re-teaching collusive behaviour to the reset competitor. A sensitivity analysis across nine combinations of learning rate α and exploration decay β finds $\Delta \geq 0.63$ in all tested cells, with non-trivial dispersion. These results provide exploratory evidence that symmetric, preventive regulatory approaches may outperform firm-specific, reactive enforcement, while the need for broader multi-seed validation remains. Even so, the same partial-disruption pattern across four different interventions is a strong signal worth testing at larger scale.

Keywords

Algorithmic Collusion, Multi-Agent Reinforcement Learning, Q-Learning, Bertrand Competition, Regulatory Intervention

1 Introduction

Algorithmic pricing systems increasingly rely on autonomous reinforcement learning agents to adapt prices in competitive markets. Recent evidence demonstrates that independent agents can converge to supra-competitive outcomes resembling tacit collusion even without communication or explicit coordination [4]. Such outcomes raise serious regulatory concerns: coordinated pricing reduces consumer welfare and undermines the conditions that antitrust law is designed to protect.

While a growing literature documents the *emergence* of algorithmic collusion, far less attention has been paid to its *persistence* once learned. In practice, regulatory interventions occur only after harmful behaviour has been detected, so understanding whether learned collusion is fragile or entrenched is of direct importance for enforcement design. Calvano et al. [4] briefly examine single-period deviations and find quick reversion to collusive prices, but do not test systematic disruptions. No prior work evaluates multiple intervention types, measures quantitative recovery rates, or asks whether disruptions affect the new equilibrium symmetrically or asymmetrically across competing firms.

This paper fills that gap. We replicate the baseline environment of Calvano et al. [4] and apply four regulatory-style interventions to the converged game: forced competitive pricing simulating a consent decree, an exploration shock simulating an audit, and a memory reset simulating algorithm replacement. For each, the game is allowed to re-converge fully, and the resulting equilibrium price is compared to both the pre-intervention baseline and the Nash benchmark.

Research Questions.

- **RQ1:** How stable is algorithmic collusion against market disruptions of different types?
- **RQ2:** Does intervention duration affect the degree of disruption?
- **RQ3:** Do symmetric and asymmetric interventions differ in effectiveness?

Main Findings. In our experimental run, all four interventions reduce equilibrium prices but incompletely. Recovery rates range from 54.5% to 81.8%, meaning collusion persists at 55–82% of its original level. Forcing Nash-level pricing for 50 and 100 periods produces the same outcome in this run, consistent with a possible disruption threshold. The memory reset, targeting one firm only, achieves the smallest disruption, consistent with the intact firm anchoring re-convergence to collusion. A sensitivity analysis over nine parameter combinations finds high collusion ($\Delta \geq 0.63$) throughout the tested grid, but with variance that motivates additional runs.

Contributions.

- To the best of our knowledge, this is the first systematic comparison of four regulatory intervention types applied to converged algorithmic collusion, using full re-convergence as the measurement standard.
- Evidence of duration insensitivity in our run, consistent with a disruption-threshold interpretation for forced pricing.

*Supervisor: Dr. Sonia Rebai (Course Instructor, MBA 501: Advanced Decision and Game Theory, Tunis Business School).

- Evidence from this setup that asymmetric interventions are less disruptive than symmetric ones, with implications for algorithm-replacement remedies.
- A compact robustness check showing that the qualitative pattern is not confined to a single (α, β) choice.

2 Related Work

2.1 Emergence of Algorithmic Collusion

The foundational contribution is Calvano et al. [4], who showed that independent Q-learning agents in repeated Bertrand competition converge to supra-competitive prices without communication. Their key finding is that simple model-free learners, optimising only individual rewards, nonetheless discover and sustain tacit collusion. Klein [8] extended this to sequential pricing, confirming that collusion persists under asynchronous updating. Abada, Lambin, and Tóth [1] validated the finding across varied market structures and parameter ranges. On the empirical side, Assad et al. [2] documented pricing patterns in the German retail gasoline market consistent with algorithmic coordination, though causal identification remains difficult with observational data.

2.2 Multi-Agent Reinforcement Learning in Economics

The broader multi-agent reinforcement learning literature establishes that independent learners can develop cooperative strategies in repeated social dilemmas even when each agent optimises only its own reward [9]. As Dafoe et al. [6] note, however, such cooperation is not guaranteed and depends on the reward structure, state representation, and learning algorithm. The conditions under which algorithmic collusion emerges, and whether it is robust to perturbation, remain active questions.

2.3 Antitrust Policy and Algorithmic Markets

Ezrachi and Stucke [7] provided an early warning that pricing algorithms could sustain anticompetitive coordination in ways that existing law was not designed to address. Baker [3] argues that antitrust doctrine is in principle adequate but faces serious enforcement challenges with algorithmic conduct. Mehra [11] contends that coordination emerging from independent learning, rather than agreement, may require new regulatory frameworks.

2.4 Gap Addressed by This Paper

To the best of our knowledge, no study has systematically measured how established algorithmic collusion responds to regulatory disruptions after convergence, and compared symmetric versus asymmetric interventions under full re-convergence. This paper provides an initial test of that gap.

3 Background

3.1 Repeated Bertrand Competition

Each firm i chooses price p_i to maximise profit $\pi_i = (p_i - c_i)\, q_i(p_i, p_{-i})$, where c_i is marginal cost and q_i follows a multinomial logit demand:

$$q_i = \frac{\exp\big((a - p_i)/\mu\big)}{\sum_j \exp\big((a - p_j)/\mu\big) + \exp(a_0/\mu)}. \tag{1}$$

Here $\mu > 0$ captures horizontal differentiation and a_0 is the outside option. In repeated play, the static Nash equilibrium price p^N is the competitive benchmark, and the joint monopoly price p^M is the collusive ceiling. The folk theorem [10] guarantees that collusive outcomes can be sustained as subgame-perfect equilibria when agents are sufficiently patient, provided deviations are detectable and punishable.

3.2 Q-Learning

Q-learning [14] is a model-free algorithm in which agent i maintains a value function $Q_i(s, a)$ and updates it after each period:

$$Q_{i,t+1}(s_t, a_{i,t}) = (1 - \alpha)\, Q_{i,t}(s_t, a_{i,t}) + \alpha \Big[\pi_{i,t} + \delta \max_{a'} Q_{i,t}(s_{t+1}, a')\Big], \tag{2}$$

where α is the learning rate and δ the discount factor. In independent Q-learning, each agent treats competitors as part of the environment, updating only on own rewards. Despite this, Calvano et al. [4] demonstrated that independent Q-learners reliably converge to collusive outcomes in Bertrand games.

3.3 Collusion Metrics

The primary measure of collusion is the Profit Gain Index:

$$\Delta = \frac{\bar{\pi} - \pi^N}{\pi^M - \pi^N}, \tag{3}$$

where $\bar{\pi}$ is the average equilibrium profit, π^N the Nash profit, and π^M the monopoly profit. $\Delta = 0$ corresponds to competitive pricing; $\Delta = 1$ to full monopoly collusion. Post-intervention effectiveness is measured by the recovery rate:

$$R = \frac{\bar{p}_{\text{post}} - p^N}{\bar{p}_{\text{base}} - p^N} \times 100\%, \tag{4}$$

where $R = 100\%$ indicates a full return to the pre-intervention equilibrium and $R = 0\%$ indicates collapse to Nash.

4 Methodology

4.1 Experimental Approach

Controlled simulation is used in place of field data because the relevant counterfactual, identical firms subject to the same regulatory action, with and without a learning algorithm, is not observable empirically. The baseline replicates Calvano et al. [4], using the open-source Python implementation of Courthoud [5]. Four interventions are applied after convergence, each followed by full re-convergence of the game.

4.2 Economic Environment

Two symmetric firms compete in an infinitely repeated Bertrand duopoly. In each period t, firms choose prices simultaneously from a discrete grid; consumers purchase according to the logit demand in Equation (1); firms earn $\pi_{i,t} = (p_{i,t} - c)\, q_{i,t}$. Parameters follow Calvano et al. [4]: $c = 1$, $a - c = 1$, $a_0 = 0$, $\mu = 0.25$, $\delta = 0.95$. These yield $p^N \approx 1.473$ and $p^M \approx 1.925$. The price grid has $m = 15$ points on $[1.2, 2.0]$.

4.3 Q-Learning Agents

Each agent observes the previous period prices. The state is $s_t = (p_{1,t-1}, p_{2,t-1})$, giving $|S| = 225$. Q-values are updated by Equation (2) with $\alpha = 0.15$. Exploration follows $\varepsilon_t = \exp(-\beta t)$ with $\beta = 4 \times 10^{-6}$, satisfying the GLIE condition required for convergence [12]. At convergence ($t \approx 1.4 \times 10^6$), $\varepsilon_t \approx 0.002$. Q-values are initialised to the discounted average profit against a uniformly randomising opponent.

4.4 Training and Convergence

Training continues until the greedy policy $\pi^*(s) = \arg\max_a Q_i(s, a)$ is unchanged across all 225 states for 10^5 consecutive periods, or until 10^7 iterations. The trained game is saved to disk; all intervention scripts load the same saved object, so the baseline is trained exactly once.

4.5 Implementation Pipeline (Scripts and Trial Structure)

The empirical workflow is implemented in three scripts. script_01_baseline.py trains the baseline once, stores the converged game as baseline_game.pkl, and generates the impulse response figure. script_02_interventions.py loads that same baseline object, creates deep copies, applies each intervention, and re-runs the full convergence routine before measuring post-intervention equilibria and recovery rates. script_03_sensitivity.py re-trains fresh games on a 3×3 (α, β) grid with three sessions per cell.

This pipeline is transparent and reproducible, but it also implies an important scope condition: the headline intervention table is based on one baseline training realisation rather than an average across many seeds. We therefore interpret duration and asymmetry patterns as evidence from this run, not as universal constants.

4.6 Intervention Design

After convergence the game object is deep-copied for each intervention. The intervention is applied, the iteration counter is reset to zero (so ε restarts from 1.0), and the same `simulate_game()` convergence procedure is re-run. The post-intervention equilibrium price is read at the new convergence point. This design avoids the ambiguity of measuring prices during a transient recovery window.

Table 1 summarises the four interventions.

4.7 Robustness Analysis

To assess whether findings are sensitive to the choice of learning parameters, the baseline game is re-trained across a 3×3 grid: $\alpha \in \{0.10, 0.15, 0.25\}$ and $\beta \in \{4 \times 10^{-6},\ 10^{-5},\ 10^{-4}\}$, with three independent sessions per cell. The baseline parameterisation ($\alpha = 0.15$, $\beta = 4 \times 10^{-6}$) appears in the centre-left cell.

Table 1: Intervention Designs and Policy Analogues

Intervention	Mechanism	Policy analogue
Forced (k=50)	Firm 0 locked to p^N for 50 periods; Firm 1 updates freely	Consent decree
Forced (k=100)	Same mechanism, 100 periods	Extended decree
Expl. shock	Both agents set ε=0.5 for 100 periods; Q-tables update	Regulatory audit
Memory reset	Firm 0 Q-table reset to $Q_{i,0}$; Firm 1 retains memory	Algorithm replacement

5 Results

5.1 Baseline Replication

Table 2 reports the baseline outcomes. Agents converge in approximately 1.4×10^6 iterations to an equilibrium price of $p^* = 1.887$, close to the monopoly benchmark of 1.925. The Profit Gain Index $\Delta = 0.991$ indicates near-full collusion, above but close to the range of 0.90–0.96 reported by Calvano et al. [4], which is plausible under single-run variation. Because this is a single baseline realisation, the level of Δ should be interpreted as run-specific rather than as a point estimate of the population mean.

Table 2: Baseline Replication Results

Metric	Value
Nash price (p^N)	1.473
Monopoly price (p^M)	1.925
Equilibrium price (p^*)	1.887
Profit Gain Index (Δ)	0.991
Convergence (iterations)	1.4×10^6

Figure 1 shows the impulse response to a unilateral downward deviation by Firm 1, replicating Figure 3 of Calvano et al. [4]. Prices fall below the Nash benchmark for approximately three periods, a punishment phase, before both firms return to the collusive equilibrium by period six. This punishment-and-forgiveness pattern is the hallmark of tacit collusion in repeated games [10] and indicates that the learned policy encodes a credible deterrence mechanism.

5.2 Intervention Results

Table 3 reports the post-intervention equilibrium price and recovery rate for each intervention. In this run, all four reduce prices relative to the baseline of 1.887, but none come close to restoring competitive pricing at $p^N = 1.473$. Recovery rates range from 54.5% to 81.8%.

Forced competitive pricing. Locking Firm 0 to the Nash price for either 50 or 100 periods yields the same post-intervention equilibrium at $p^* = 1.699$ ($R = 54.5\%$, $\Delta p = -0.188$). This is the

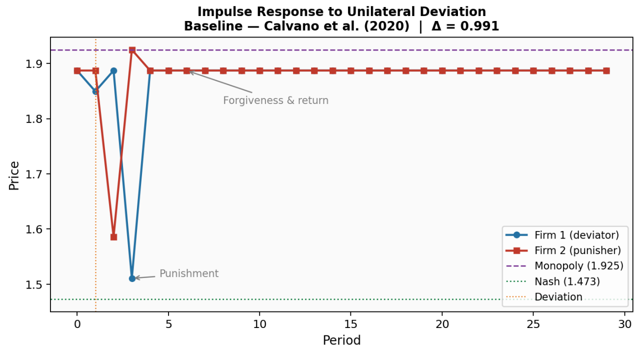

Figure 1: Impulse response to a unilateral one-step downward deviation by Firm 1 (period 1, orange dotted line). Prices drop below Nash (green dotted) for ≈ 3 periods before returning to the collusive equilibrium. Purple dashed: monopoly price (1.925). Baseline: $\Delta = 0.991$.

Table 3: Post-Intervention Equilibrium Prices and Recovery Rates

Intervention	$\bar{p}_{\text{post}}$	$R\,(\%)$	Δp
Forced ($k = 50$)	1.699	54.5	−0.188
Forced ($k = 100$)	1.699	54.5	−0.188
Expl. shock	1.756	68.2	−0.131
Memory reset	1.812	81.8	−0.075
Nash ($R = 0\%$)	1.473	0.0	N/A
Baseline ($R = 100\%$)	1.887	100.0	N/A

largest disruption observed in this run. The insensitivity to duration is therefore a suggestive empirical pattern: the disruption effect may saturate quickly, and extending enforcement beyond a threshold may add little. In this setup, the exact match is also mechanically plausible: both interventions start from the same saved baseline object and use the same deterministic convergence criterion, so once Q-value corruption crosses the disruption threshold, re-convergence can follow the same path. During the forced period, Firm 1 continues learning against a fixed Nash-pricing opponent, which corrupts its Q-values for high-price states. Once both firms resume free learning, neither fully recovers the original collusive coordination.

Exploration shock. Spiking both agents' exploration to $\varepsilon = 0.5$ for 100 periods yields $p^* = 1.756$ ($R = 68.2\%$, $\Delta p = -0.131$). The disruption is smaller than forced pricing because the shock is symmetric: both Q-tables are disturbed in the same direction and to a similar degree, which may facilitate re-coordination once the shock ends.

Memory reset. Resetting Firm 0's Q-table while Firm 1 retains its full history produces the smallest disruption: $p^* = 1.812$ ($R = 81.8\%$, $\Delta p = -0.075$). The intact Firm 1 acts as an anchor. As Firm 0 re-explores the price space, Firm 1's policy consistently rewards high-price responses and penalises deviations, effectively guiding

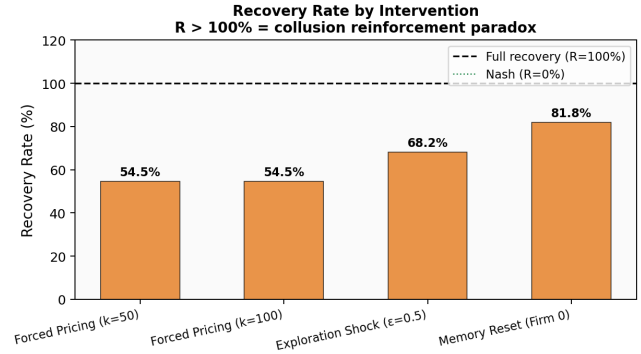

Figure 2: Recovery rates by intervention. All bars fall strictly below the $R = 100\%$ full-recovery line. Forced pricing achieves the most disruption ($R = 54.5\%$); memory reset achieves the least ($R = 81.8\%$). The grey shaded region above $R = 100\%$ (unused here) would indicate post-intervention prices exceeding the pre-intervention baseline.

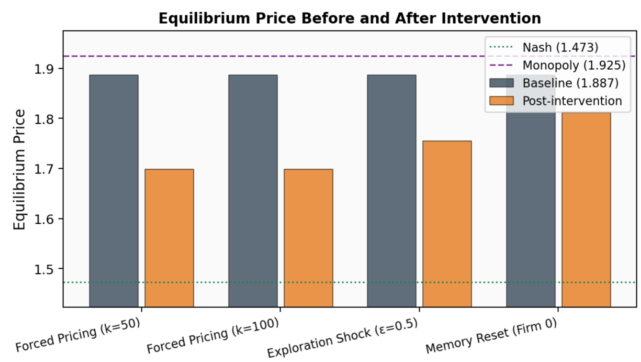

Figure 3: Equilibrium prices before (dark grey, baseline = 1.887) and after (orange) each intervention. Nash benchmark (green dotted, $p^N = 1.473$) and monopoly ceiling (purple dashed, $p^M = 1.925$) shown. All post-intervention prices remain substantially above Nash.

re-convergence toward collusion. This result has direct implications for algorithm-replacement remedies: if only one firm's algorithm is replaced, the incumbent's intact strategy may undo much of the regulatory benefit in this setting.

One clarification is important for interpretation. Symmetry plays different roles across interventions. In forced pricing, the temporary constraint is asymmetric and directly distorts one firm's learning path, which increases disruption. In the exploration shock, both firms are disturbed symmetrically, which can make re-coordination easier once the shock ends. In memory reset, asymmetry is again different: one firm keeps full collusive memory while the other is reset, which can reduce disruption because the intact firm can re-anchor behaviour.

Figures 2 and 3 visualise these results across all interventions.

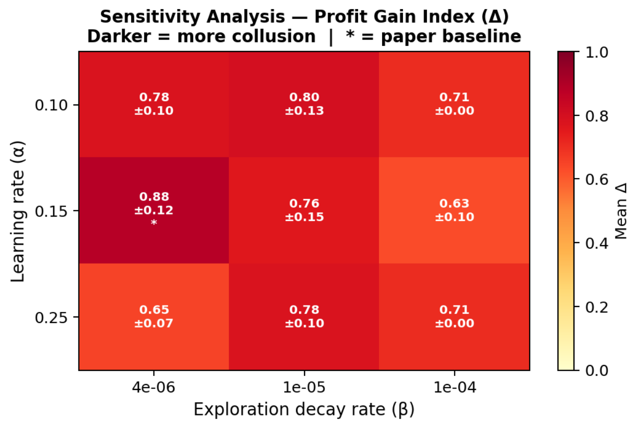

Figure 4: Sensitivity of the Profit Gain Index Δ to learning parameters α (learning rate) and β (exploration decay rate). Each cell: mean ± std across three sessions. The paper baseline ($\alpha = 0.15$, $\beta = 4 \times 10^{-6}$) is marked with an asterisk. All cells: $\Delta \geq 0.63$.

5.3 Robustness: Sensitivity Analysis

Figure 4 shows the mean Δ across the 3×3 parameter grid (three sessions per cell). Every cell yields $\Delta \geq 0.63$; the minimum is 0.63 ± 0.10 at ($\alpha = 0.15, \beta = 10^{-4}$) and the maximum is 0.88 ± 0.12 at the paper's baseline ($\alpha = 0.15, \beta = 4 \times 10^{-6}$). These results indicate that substantial collusion is not confined to a single parameter choice in the tested grid. However, with only three sessions per cell and visible standard deviations, this robustness exercise should be treated as indicative rather than definitive.

Two directional patterns are visible in the heatmap. Faster exploration decay (larger β, moving right) tends to reduce Δ, which is intuitive: agents that stop exploring early converge on less well-coordinated strategies. The relationship with α is non-monotone, reflecting the known tension between learning speed and stability in Q-learning [13].

6 Discussion

6.1 Partial Disruption in This Experimental Setup

The uniform pattern across mechanistically distinct interventions, all four reduce collusion and none eliminate it, suggests that partial recovery may be a persistent feature of Q-learning in repeated Bertrand competition. After convergence, the Q-tables encode a reinforced collusive strategy. Interventions perturb this encoding but, in our run, do not erase it. Even the most disruptive intervention (forced pricing, $R = 54.5\%$) leaves agents converging to an equilibrium well above Nash.

6.2 Possible Duration Thresholds

The identical outcomes for $k = 50$ and $k = 100$ forced pricing point to a possible threshold mechanism. Once a forced-pricing intervention is long enough to corrupt the Q-values in the states most relevant to collusive coordination, extending it further provides no additional benefit in this run. From a regulatory standpoint, this suggests that the practically relevant question may be not only *how long* to enforce a remedy, but also *whether* the intervention exceeds the minimum threshold needed to destabilise the learned strategy. This interpretation should be verified with additional seeded replications.

6.3 Symmetric versus Asymmetric Interventions

The memory reset, the only asymmetric intervention tested, achieves the smallest disruption ($R = 81.8\%$). Both exploration shock and forced pricing affect both firms either directly or structurally, and both achieve larger disruptions ($R = 54.5\%$ and 68.2% respectively). A plausible mechanism is that an intact firm facing a reset competitor has both the incentive and the informational advantage to steer re-convergence toward collusion. This finding challenges a common policy intuition, that replacing one firm's algorithm is sufficient to break coordination, and suggests instead that regulators should target both sides of a market simultaneously for stronger effects.

6.4 Policy Implications

Three practical conclusions follow from the results in this setup.

First, reactive interventions do achieve a meaningful reduction in collusion. A decline from $\Delta = 0.991$ to an effective Δ in the range 0.55–0.82 represents a real consumer welfare gain even if full competition is not restored. Regulators should not conclude from partial recovery that enforcement is futile.

Second, symmetric enforcement appears to outperform targeted enforcement in this experiment. Policies that simultaneously disrupt all competing algorithms, such as industry-wide audit requirements, mandatory periodic resets, or co-ordinated exploration floors, should produce larger and more durable reductions than firm-specific remedies.

Third, preventive approaches may dominate reactive ones. A single intervention does not restore competition, and repeated interventions carry escalating costs. Policies that prevent collusion from forming in the first place, algorithmic design requirements, transparency mandates, or ex-ante approval regimes, may produce better long-run outcomes than post-detection enforcement alone.

6.5 Limitations

The baseline uses a single training run rather than an average across many seeds, so the reported $\Delta = 0.991$ may reflect favorable realisation noise. Interventions are applied once in isolation; repeated or coordinated interventions may produce different dynamics. The sensitivity analysis uses only three sessions per cell, and the observed standard deviations indicate meaningful stochastic variation.

These design choices were constrained by project time and compute budget, and therefore this manuscript should be read as a careful exploratory study rather than a fully powered statistical exercise. The current evidence is best interpreted as sufficient grounds for continued testing rather than a final factual statement about all algorithmic markets. Finally, the model abstracts away from entry, asymmetric costs, demand uncertainty, and multi-product competition, all of which are present in real algorithmic markets. Extending the analysis to these settings is left for future work.

7 Conclusion

This paper examined whether regulatory-style interventions can disrupt algorithmic tacit collusion once it has been learned. Replicating the baseline of Calvano et al. [4], independent Q-learning agents converge in our baseline run to an equilibrium price of 1.887 and a Profit Gain Index of $\Delta = 0.991$. All four interventions tested, forced competitive pricing, an exploration shock, and a memory reset, reduce equilibrium prices but incompletely. Recovery rates range from 54.5% to 81.8%, meaning that in this run a substantial share of collusive pricing survives targeted disruption.

Two findings appear particularly policy-relevant. The duration insensitivity of forced pricing, 50 and 100 periods yield the same outcome in this run, is consistent with enforcement intensity beyond a threshold adding little incremental benefit. The higher recovery rate of the memory reset relative to symmetric interventions suggests that replacing one firm's algorithm while leaving the competitor's intact may be a weak long-term remedy.

A sensitivity analysis over a 3×3 parameter grid finds that $\Delta \geq 0.63$ across all tested combinations of learning rate and exploration decay, indicating that the qualitative pattern is not tied to one parameterisation within this grid.

Taken together, the evidence in this study points toward a possible need for preventive rather than purely reactive antitrust policy in algorithmic markets. Symmetric, ex-ante regulatory approaches, design requirements, exploration floors, mandatory periodic algorithmic resets applied industry-wide, are likely to outperform firm-specific, post-detection enforcement. As autonomous pricing agents become more prevalent, developing regulatory frameworks that account for the learning dynamics studied here becomes increasingly urgent. At the same time, broader multi-seed and multi-market validation is needed before making strong quantitative policy claims. In that sense, these results should be read as credible signals and as sufficient grounds to continue testing these mechanisms at larger scale.

References

[1] Ibrahim Abada, Xavier Lambin, and Balázs Tóth. 2023. Artificial Intelligence, Algorithmic Pricing, and Tacit Collusion: Evidence from Simulations. *Journal of Industrial Economics* 71, 2 (2023), 355–390. doi:10.1111/joie.12317

[2] Stephanie Assad, Robert Clark, Daniel Ershov, and Liting Xu. 2024. Algorithmic Pricing and Competition: Empirical Evidence from the German Retail Gasoline Market. *Management Science* 70, 1 (2024), 1–23. doi:10.1287/mnsc.2023.4736

[3] Jonathan B. Baker. 2021. Algorithms and Tacit Collusion: An Antitrust Analysis. *Antitrust Law Journal* 84 (2021), 621–662.

[4] Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. 2020. Artificial Intelligence, Algorithmic Pricing, and Collusion. *American Economic Review* 110, 10 (2020), 3267–3297. doi:10.1257/aer.20190623

[5] Matteo Courthoud. 2021. Algorithmic Collusion Replication. https://github.com/matteocourthoud/Algorithmic-Collusion-Replication. GitHub repository.

[6] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2020. Open Problems in Cooperative AI. *arXiv preprint arXiv:2012.08630* (2020).

[7] Ariel Ezrachi and Maurice E. Stucke. 2016. *Virtual Competition: The Promise and Perils of the Algorithm-Driven Economy.* Harvard University Press, Cambridge, MA.

[8] Timo Klein. 2021. Autonomous Algorithmic Collusion: Q-Learning under Sequential Pricing. *The RAND Journal of Economics* 52, 3 (2021), 538–558. doi:10.1111/1756-2171.12383

[9] Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-Agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th International Conference on Autonomous Agents and MultiAgent Systems*. 464–473.

[10] Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. 1995. *Microeconomic Theory.* Oxford University Press, New York.

[11] Salil K. Mehra. 2016. Antitrust and the Roboseller: Competition in the Time of Algorithms. *Minnesota Law Review* 100 (2016), 1323–1375.

[12] Satinder Singh, Tommi Jaakkola, Michael L. Littman, and Csaba Szepesvári. 2000. Convergence Results for Single-Step On-Policy Reinforcement-Learning Algorithms. *Machine Learning* 38, 3 (2000), 287–308. doi:10.1023/A:1007678930559

[13] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press, Cambridge, MA.

[14] Christopher J. C. H. Watkins and Peter Dayan. 1992. Q-Learning. *Machine Learning* 8, 3–4 (1992), 279–292. doi:10.1007/BF00992698