Less than meets the eye: simultaneous experiments as a source of algorithmic seeming collusion*

Xavier Lambin†

July 9, 2024

Abstract

This article challenges the mechanisms of algorithmic collusion as proposed in earlier works. Identifying the failure to pass a key robustness test, we question the idea that supracompetitive prices result from collusive mechanisms where high prices are sustained by reward and punishment strategies. Instead, our analysis suggests that both phenomena may originate from simultaneous experimentation and learning inertia inherent in reinforcement learning, without a causal link between them. Such seeming collusion can emerge in memoryless environments and with myopic agents, and much faster than previously thought. We caution against straightforward analogies with human collusion and recommend addressing high algorithmic prices with specific measures.

Keywords- Algorithmic decision-making, Machine learning, multi-agent reinforcement learning, tacit collusion.

*This project is supported by the French Government's Investment for the Future Program (PIA), now part of France 2030, operated by ADEME. The views expressed in this paper are those of the authors alone and might not represent the views of ADEME or any other institution.

†ESSEC Business School and THEMA, 3 Av. Bernard Hirsch, B.P. - 50105, Cergy, 95021, France. xavier.lambin@essec.edu

1 Introduction

Algorithmic decision-making has become ubiquitous in our lives, and its impact is increasing at an unprecedented rate. From our social media feeds to the stock market, from self-driving cars to medical diagnoses, algorithms are increasingly being used to automate decision-making processes.

In an influential paper Calvano et al. [2020b] (henceforth CCDP), show that basic and independent reinforcement algorithms, when trained simultaneously, consistently achieve supra-competitive outcomes. Furthermore, the responses of the algorithms to out-of-equilibrium stimuli resemble reward-punishment schemes that may be used to sustain collusion. The authors conclude that algorithms genuinely collude and provide policy recommendations and guidance to antitrust authorities. Following this, Calvano et al. [2020a] present tests, based on responses to stimuli, that regulators can employ to verify whether algorithms are autonomously engaging in collusion.[1] This research has been followed up by many other studies that have used and expanded the notion of algorithmic collusion, such as Hettich [2021], Dolgopolov [2021], Banchio and Skrzypacz [2022], Klein [2021], Werner [2022], Qiu et al. [2022], Xu et al. [2023], to name just a few. Should these findings be confirmed, they could have significant implications for antitrust regulations, necessitating urgent action. It is therefore no surprise that the issue of Artificial Intelligence (AI) collusion has garnered attention from regulatory agencies, with it being a top priority on the agendas of many organizations (see e.g. OECD [2017], Autoridade Da Concorrència [2019], ACB [2019], Ezrachi and Stucke [2018], McSweeny and O'Dea [2017], Competition Bureau [2018] and Petit [2017]).

The interpretation of the results of CCDP and subsequent works remain, however, a topic of debate. The specificity of the algorithmic design has been discussed by Meylahn et al. [2022], Abada et al. [2022], Eschenbaum et al. [2022], while the interpretation of the results has been examined by Epivent and Lambin [2024], Abada and Lambin [2023]. Asker et al. [2022, 2023] emphasize the critical role of algorithmic learning protocols on supracom-

[1] This literature, as well as the present paper, focuses on autonomous algorithmic collusion. A situation when managers would deliberately agree to implementing algorithms that are known to converge to high prices (see for example the "Amazon poster case" U.S. v. Topkins [2023]) would likely qualify as traditional collusion. Similarly, algorithms can lead to high prices because of commitment (Salcedo [2015], Brown and MacKay [2023]) or single supplier effects (Harrington [2022]). These discussions, though interesting, are not covered in the present paper.

petitive limit prices. Calvano et al. [2023] attribute the "spurious" collusion results of Asker et al. [2023] to the specific exploration mode they implement (synchronous learning), together with optimistic initialization of Q-matrices. Overall, the literature provides no formal explanation for the observational facts described in CCDP. This is mostly due to the fact it is notoriously difficult to draw theoretical results in multi-agent Q-learning processes.

This paper employs a simplified exploration procedure to elucidate the dynamics at play. It demonstrates that apparent collusion may arise due to the specific learning process inherent in simultaneously-trained reinforcement learning algorithms: by construction, the initial valuations of actions are based on experiments that are performed while the other agents are also experimenting. These valuations may differ significantly from the profits observed in "play" mode when all agents play only their preferred (or "greedy") actions. Still, the learning procedure is such that the erroneous valuations persist over time. When the rates of exploration decrease jointly, we show that agents may fail to identify profitable independent deviations and converge to prices (much) greater than Nash. Our theory is similar in spirit to that of Banchio and Mantegazza [2022], though they focus on explaining the cycles observed after learning, while we rationalize the convergence to singleton or fixed-point (supra-competitive) strategies, which represent 64% of the simulations in CCDP. We address much more general demand systems than the prisoner dilemma, with a specific application to the economic environment of CCDP. We use a mean-field assumption, eliminating the need for continuous time approximation of the learning process. Compared to Asker et al. [2023], we provide a complete characterization of the initial and final "greedy actions" in the Q-learning context, when endowed with a simple exploration procedure. We offer a comprehensive description of the underlying mechanism, including the characterization of the convergence point of the algorithm of CCDP. A notable contribution to the literature is our explanation for the apparent reward and punishment schemes identified in CCDP and subsequent works, clarifying that these schemes may not be the cause of the observed high prices.

Our theory is then compared to the findings CCDP, which we replicate faithfully. The first important step is to show that the memoryless version of CCDP also yields supra-competitive prices,[2] which challenges the claim that

[2]"Memoryless," or "stateless," algorithms refer to algorithms in which an agent's state space does not

the high prices are caused by reward and punishment schemes. In a second step, the results from our theoretical models are shown to provide a credible explanation to the main observations of this literature, with simultaneous learning causing the convergence to high prices. Finally, the apparent reward and punishment schemes are also explained. From these observations, we note that the interpretation of high prices and apparent punishment schemes as evidence of collusion similar to known human conduct may have led to misguided policy recommendations. We propose to refine the interpretation and to implement more straightforward policy interventions against supracompetitive prices.

The remainder of the paper is organized as follows. Section 2 faithfully replicates the experiments of CCDP. It reconsiders a previously overlooked robustness test, challenging the idea that the phenomenon may correspond to collusion, defined as high prices sustained by reward and punishment schemes. Section 3 rationalizes the observational findings of prior literature by providing a theoretical foundation for Q-learning seeming collusion, which is numerically tested in Section 4. Section 5 provides remedies to algorithmic supracompetitive prices and derives policy recommendations. Section 6 concludes.

2 The previous literature reexamined

CCDP claims that the supra-competitive profits observed in their setting are a consequence of collusive reward and punishment schemes. This conjecture is, however, based on the interpretation of observational facts, but no formal proof is provided. With an exact replication of CCDP, this section proves that this conjecture insufficiently substantiated since supra-competitive prices are maintained (and even strengthened) when agents have no memory of past actions, a situation when reward and punishment schemes cannot be applied, as recalled by the authors.[3]

include past observations, preventing the use historical price data from other agents to inform current actions.

[3]Following CCDP and much of the literature, we adopt the definition of Harrington [2019], p.6 as well as its preamble: "Collusion ties future rewards and punishments to current behavior and that is what sustains a supracompetitive outcome. It is this causal relationship that links a firm's current conduct with rival firms' future conduct that defines collusion, not the setting of supracompetitive prices. Supracompetitive prices are the product of that causal relationship."

2.1 The memoryless version of the algorithm of Calvano et al. [2020b] yields seemingly collusive outcomes

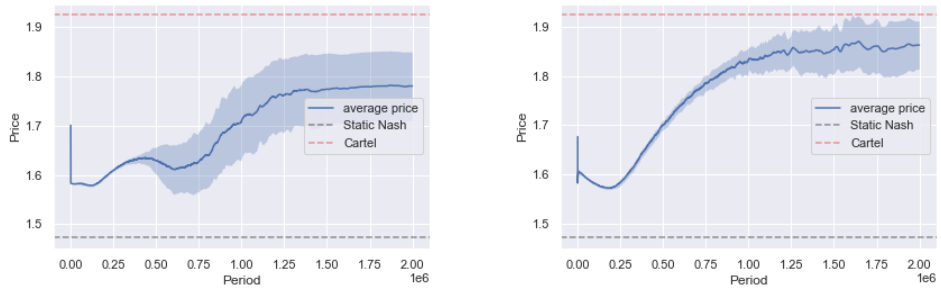
Waltman and Kaymak [2008] find that Q-learning agents systematically reach supra-competitive outcomes, a phenomenon they already call "collusion." The key criticism of CCDP to Waltman and Kaymak [2008] is that they "consider also the case where algorithms are myopic and have no memory of past actions –conditions under which collusion is either unfeasible or cannot emerge in equilibrium– and find that in these cases the output reduction is even larger. This raises doubts that what they observe may not be collusion but a failure to learn an optimal strategy" (CCDP, p.2). However, the case of memoryless algorithms is not treated in CCDP either, except in the very specific case when agents are myopic.[4] We now address this omission.

We first replicate faithfully the setting of CCDP. Figure 1(a) displays the greedy action in their main specification, where agents have a one-step memory of past prices. Then, we consider also the case when algorithms do not observe the other agent's prices (they are "memoryless"). As a consequence, the state space of the algorithms does not evolve over episodes (their perceived environment is static because they have no memory of past actions). This precludes reward-punishment schemes since there is no possibility of observing the price of the other agent, and therefore no possibility to reward or punish.[5] The one difference with the memoryless algorithms of CCDP, however, is that we don't assume that the agents also become myopic: they keep the standard discount factor of $\delta = 0.95$ that is used in the main specification in CCDP. We observe in Figure 1(b) that these memoryless agents very quickly learn to post supra-competitive prices. Comparing this result to the partial "collusion" of the main specification of CCDP (agents with memory, Figure 1(a)), the memoryless version of

[4]The authors state in the main text that they do test memoryless algorithms and that they converge on "approximately, the one-shot equilibrium." Yet, their Online Appendix A4.1, reveals that they only test it jointly with myopic agents, which the authors have shown earlier to be of itself sufficient to induce convergence to outcomes close to Nash. They state "With memoryless algorithms ($k = 0$), there is no loss of generality in setting the discount factor δ to 0." It turns out that this statement is incorrect. In the context of Q-learning, a positive discount factor does affect the dynamics updating of the Q-value of the actions experimented (see Section 3.2). The same methodological issue is present in Calvano et al. [2023]. In the present section, agents stick to the discount factor of $\delta = 0.95$, which is the value used in the baseline simulations of CCDP and of the present paper.

[5]Though the observation of profits, together with an accurate representation of the environment could in principle allow an agent to infer its opponent's action, the updating of Q-values as well as the impossibility to condition future actions on one's own past actions (Equation 1) does not allow for such a possibility.

the algorithms seem to collude even more effectively than algorithms with memory.[6] This is at odds with the interpretation that supra-competitive prices are due to reward and punishment schemes.



(a) one-step memory of past actions (replication of CCDP)

(b) no memory of past actions (stateless algorithm)

Figure 1: Average "greedy" price over learning episodes ($\delta = 0.95$)

The observation that memoryless or stateless algorithms can attain prices and profits surpassing those achieved by seemingly collusive algorithms with memory is not exclusive to this particular context. In Appendix A, we replicate the settings and main findings from Klein [2021] and Calvano et al. [2023], and we identify a similar effect in those studies as well. Previous literature has shown that stateless algorithms can autonomously reach supra-competitive prices in other contexts (Cooper et al. [2015], Hansen et al. [2021], Banchio and Mantegazza [2022]). Our replication effort shows that the stateless version of the algorithm of CCDP and related papers also reach high prices, which is a novel finding. The fundamental point we are making here, is that what has been qualified as "collusion" in the sense of Harrington [2019] in some of the recent literature cannot qualify as such: the reward and punishment schemes that should sustain the high prices simply cannot survive the complete unobservability of past actions.

2.2 Should we care about whether algorithmic supracompetitive prices stem from "collusion" or not?

Yes, we should, because incorrectly categorizing a phenomenon as "collusion" could lead to the application of an inadequate remedy. As we showed

[6] A qualitatively similar finding was presented in Asker et al. [2023], albeit in a very different setting.

in the previous subsection with a simple one-factor-at-a-time method (switching to memoryless algorithms), the phenomenon described in the existing literature cannot be "collusion," as defined by high prices sustained by reward-punishment strategies. The invalidation of this interpretation naturally calls for a reconsideration of the policy implications of the previous literature.

Indeed, many scholars and policy-markers have pointed out that market transparency rendered algorithmic collusion more likely (see e.g., ACB [2019], Gal [2019], Kühn and Tadelis [2018], Assad et al. [2020, 2021], Ezrachi and Stucke [2020]). Therefore, an intuitive remedy to algorithmic "collusion," inspired from real-life cases involving (human) collusion, would be to restrict algorithms' access to market data, such as their competitors' prices. This measure makes reward-punishment schemes, and hence collusion, impractical (see e.g. Kühn and Vives [1995], Albæk et al. [1997]). However, reducing market transparency is ineffective in addressing the issue at hand here, as supra-competitive outcomes are shown not to be sustained by reward-punishment schemes. In fact, the high prices can still arise even when competitors' prices are unobserved. In our specific application, Figure 1 shows that such a measure would *increase* prices even further in practice. Furthermore, preventing access to market data actually simplifies the relevant state space of the algorithms, potentially leading to even faster convergence to supra-competitive outcomes.

3 A simple theory

The previous section established that the phenomenon described in earlier literature does not align with the classic notion of collusion. However, it still bears striking resemblances to collusion, with high prices and observable patterns resembling price wars. In the following, we offer a straightforward explanation for these similarities. We show that both the high prices and the apparent price wars stem from the simultaneous learning process implemented in the previous literature. In particular, we show that there is no causal link between the prices wars and the high prices, which precludes a classification of the phenomenon as collusion in the sense of Harrington [2019].

3.1 Stage game and notations

There are two symmetric agents. They can choose any price out of a discrete set of prices $\mathcal{P}$ of size m: $m = card(\mathcal{P})$. The choice of a price is also named the "action" of the agent. The instantaneous profit of an agent posting a price p^l while the other agent posts a price p^{-l} is $\pi(p^l, p^{-l})$. We assume this game has (at least) one pure Nash equilibrium, which consists of agents setting a symmetric price $p_N \in \mathcal{P}$, yielding individual profit π_N.

Define $\tilde{\pi}(p) \equiv \mathbb{E}_q \pi(p, q)$, the expected profit of an agent when she posts price p and the other agent plays each action in $\mathcal{P}$ with uniform probability.

We index the actions available to each agent by decreasing order of expected profits, conditional on the other agent playing randomly: $\tilde{\pi}(a_1) > \tilde{\pi}(a_2) > ... > \tilde{\pi}(a_m)$. In particular, $a_1 \equiv \underset{p}{argmax}\, \tilde{\pi}(p)$ is the price that maximizes the expected profit. The associated profit is $\tilde{\pi} \equiv \underset{a_i}{max}\, \tilde{\pi}(a_i) = \tilde{\pi}(a_1)$.

3.2 Q-learning basics

Q-learning is a simple reinforcement learning algorithm that aims to learn the optimal action for an agent taking action a in a given state s of its observed environment. The estimation of the expected flow of future payoffs that follow a certain action in a certain state is called the Q-value. The algorithm works by iteratively updating the Q-value estimates for each state-action pair based on the observed rewards and the estimated future Q-values. At each time step t in the learning process, the Q-value is denoted as $Q^t(a_t, s_t)$ for each action a_t and state s_t. These individual Q-values are then stored in the Q-matrix Q^t. s_t is the state of the agent at time t, which could for example be the couple of last prices posted by Agent 1 and 2 $(p^{t-1,1}, p^{t-1,2})$ if these can be observed by the agents. This is sometimes called the "memory" of the agents. At each period t of the learning process, Agent l plays an action a_t and the other agent plays b_t. The updating of the Q-values of Agent l at

each period t proceeds as follows:

$$\begin{aligned} \text{if } a = a_t \text{ and } s = s_t\text{:}\ \ Q_l^{t+1}(a_t, s_t) &= (1-\alpha) Q_l^t(a_t, s_t) \\ &\quad + \alpha \left(\pi(a_t, b_t) + \delta \max_{a' \in \mathcal{P}} Q_l^t(a', s_{t+1}) \right) \\ \text{otherwise:}\ \ Q_l^{t+1}(a, s) &= Q_l^t(a, s) \end{aligned} \tag{1}$$

, and similarly for Agent 2. The second line simply means that only the Q-values of actions taken at time t in state s_t are updated at time t. We follow much of the reinforcement learning literature as well as CCDP by imposing an ϵ-greedy exploration process: with exogenous probability ϵ_t, agents select an action randomly over the action space, following a uniform distribution (they "explore"). With probability $1 - \epsilon_t$, agents play the action with the largest Q-value (they "exploit"). This action is called the "preferred", or "greedy" action. For further implementation details of Q-learning the reader is referred to the original paper of Watkins [1989].

The proof of convergence for Q-learning in a stochastic environment or multi-agent is generally based on the foundational work of Watkins [1989] and the subsequent analysis by Jaakkola et al. [1994] or Hu and Wellman [1998]. However, one key assumption that the learning rate α decreases to zero over time is violated in CCDP. Indeed, it uses a constant learning rate $\alpha > 0$, which means that the estimates may never completely converge as long as algorithms explore with positive probability, but continue to vary marginally in response to the most recently received rewards. Although this doesn't pose any practical implementation challenges, it can substantially complicate mathematical proofs.

To overcome this issue, we assume (in the theory Section 3.3 only) that algorithms consider their expected payoffs, rather than the realized ones: whenever they take an action, they receive the *expected* profit (according to the exploration process of the opponent), as opposed to their *actual* profit, which would be generated by a single draw of the opponent's possible actions. This "mean-field" assumption which is quite common in the literature simplifies the exposition considerably and allows us to derive clear results while abiding to the discrete time nature of Q-learning.[7] Numerical applica-

[7] It is for example derived from the fluid approximation mechanism of Banchio and Mantegazza [2022], which allows the Q-learning algorithms to become deterministic in the limit.

tions in Section 4 and Section 5 do not use this assumption, which is useful for theoretical exposition only in the present section.

3.3 Multi-agent Q-learning with a 2-steps exploration policy: high prices explained

It is notoriously difficult to establish theoretical results in multi-agent Q-learning, with slowly declining exploration rates. It turns out, however, that the phenomenon of seeming collusion can be explained quite simply, even with an exploration policy that is greatly simplified relative to what is sometimes done in practice. To build intuitions, we assume that agents adopt the most straightforward learning process, which simply consists of two phases. In the first phase, agents solely focus on exploration, setting $\epsilon_t = 1$. Following this "initialization" phase, after a sufficiently long time T_1 has elapsed and valuations have stabilized, the agents start fully exploiting their greedy action, setting $\epsilon_t = 0$. To keep notations concise and because our setting allows for any definition of states based on historical observations (including stateless algorithms), we omit the state argument in the definition of Q-matrices.

Q-values in the exploration phase ($\epsilon_t = 1$). Assume there is full exploration for T_1 periods. Denote $n_i(t)$ the number of times action i has been explored at time t, and $t(n_i)$ the time when action a_i is explored for the n_i-th time. Using (1) and the mean-field assumption, and dropping the subscript l for convenience, we get for both agents:

$$Q^{t(n_i+1)}(a_i) = \quad (1-\alpha)Q^{t(n_i)}(a_i) + \alpha\left(\tilde{\pi}(a_i) + \delta \max_{a\in\mathcal{P}} Q^{t(n_i+1)-1}(a)\right) \quad (2)$$

In our application as well as in the literature, Q-matrices are shown to always converge. Consequently, the following discussion assumes the convergence of Q-matrices, and Lemma 1 computes the Q-values when T_1 is large, assuming Q-matrices have converged.

Lemma 1. *Assume Q-matrices converge and $\epsilon_t = 1$. Then, $Q^{T_1}(a_i) \underset{T_1\to\infty}{\longrightarrow} \tilde{\pi}(a_i) + \frac{\delta}{1-\delta}\tilde{\pi}$*

Proof. Since Q-matrices have converged, each agent has a unique action with highest Q-value. This means that there exists a time T such that for all $t > T$,

there exists an action a_i such that $a_i \equiv \underset{a_j}{argmax}\, Q^t(a_j)$. As a consequence, $\underset{a\in\mathcal{P}}{max} Q^{t(n_i+1)-1}(a) = Q^{t(n_i+1)-1}(a_i)$. Using t_j as a shorthand for the number of times action a_j was visited after period T, from Equation (2) we observe that the sequence $Q^{t_i}(a_i)$ has a closed-form solution:

$$Q^{t_i}(a_i) = (1-\alpha^{t_i}(1-\delta))\left(Q^T(a_i) - \frac{\tilde{\pi}(a_i)}{1-\delta}\right) + \frac{\tilde{\pi}(a_i)}{1-\delta}$$

We note that with $\epsilon_t = 1$, in expectation, $\mathbb{E}(t_i) = T_1/m$. In turn, we obtain that $Q^{t_i}(a_i) \underset{T_1\to\infty}{\longrightarrow} \frac{1}{1-\delta}\tilde{\pi}(a_i)$, which implies that $\underset{a\in A}{max} Q^{t_j} \underset{T_1\to\infty}{\longrightarrow} \frac{1}{1-\delta}\tilde{\pi}(a_i)$. In turn, observing from Equation (2) that $Q^{t_j}(a_j)$ is upper and lower-bounded by arithmetico-geometric processes of ratio lower than 1, we straightforwardly get that $Q^{t_j}(a_j) \underset{T_1\to\infty}{\longrightarrow} \tilde{\pi}(a_j) + \frac{\delta}{1-\delta}\tilde{\pi}(a_i)$. By the definition of $\tilde{\pi} \equiv \underset{p}{argmax}\tilde{\pi}(p)$ and of our mapping a, we must have that $a_i = a_1$, which completes the proof. □

According to Lemma 1, when agents have explored adequately such that the Q-values of their actions are sufficiently close to their true values, their preferred action becomes $a_1 = argmax_p\tilde{\pi}(p)$. This first step can be viewed as the initialization of the Q-matrix.[8]

Q-values in the exploitation phase ($\epsilon_t = 0$). We assume now that at time T_1, which is large enough so the valuations of the first phase have converged, the exploration rate is decreased from 1 to 0, for T_2 periods until period $T = T_1 + T_2$. By design, this means that only the greedy action a_1, which is the action with the highest Q-value, is played during the exploitation phase until its Q-value –which is still being updated– possibly drops below the Q-value of another action. As long as a_1 is the greedy action, the Q-value follows the following process:

$$\begin{aligned} Q^{t_1+1}(a_1) &= (1-\alpha)Q^{t_1}(a_1) + \alpha\left(\pi(a_1,a_1) + \delta\max_{a\in\mathcal{P}} Q^{t_1}\right) \\ &= (1-\alpha(1-\delta))Q^{t_1}(a_1) + \alpha\left(\pi(a_1,a_1)\right) \end{aligned}$$

[8]We note that CCDP initialize at $Q^0(a_i) = \frac{\tilde{\pi}(a_i)}{1-\delta}$. Our random initialization is a linear and increasing transformation of the CCDP initialization. As such, our qualitative results exposed thereafter hold even with the CCDP initialization. We choose random initialization to align with the extensive exploration process of CCDP, which *de facto* implements random initialization.

The Q-value of other actions remain unchanged as there is no exploration and only the Q-value of the greedy action is updated.

Lemma 2. *With $\epsilon_t = 0$, and assuming a_1 remains the greedy action throughout the exploration phase:*

$$Q^T(a_1) \underset{T\to\infty}{\longrightarrow} \frac{1}{1-\delta}\pi(a_1, a_1) \tag{3}$$

$$\forall T \geq T_1\,,\, \forall i > 1\,, Q^T(a_i) = \tilde{\pi}(a_i) + \frac{\delta}{1-\delta}\tilde{\pi} \tag{4}$$

Proof. With an exploration rate of $\epsilon = 0$, both agents start playing a_1. To get Relation (3), we note that the updating process of $Q^t(a_1)$ is an arithmetico-geometric sequence that has a closed-form solution. For $t > T_1$:

$$Q^t(a_1) = \alpha^{t-T_1}\left(Q^{T_1}(a_1) - \frac{\pi(a_1, a_1)}{1-\delta}\right) + \frac{\pi(a_1, a_1)}{1-\delta}$$

This expression converges to $\frac{\pi(a_1,a_1)}{1-\delta}$ as t goes to infinity. The Relation (4) is obtained by noting again that the exploration rate ϵ_t is set to 0. As a consequence, as long as $a_1 = argmax_a Q^t(a)$, the Q-values of actions other than a_1 stay constant over time. □

From the comparison of Equations (3) and (4), we observe that if $\pi(a_1, a_1) < (1-\delta)\tilde{\pi}(a_2) + \delta\tilde{\pi}$, after a certain number of periods, a_2 becomes the preferred action. a_2 is then played with probability 1 and will converge on its true value $\frac{1}{1-\delta}\pi(a_2, a_2)$. The maximum between $\frac{1}{1-\delta}\pi(a_1, a_1)$ and $\frac{1}{1-\delta}\pi(a_2, a_2)$ is then compared to $Q^{T_1}(a_3)$. The process goes on to action a_3, a_4... Unraveling this reasoning to all actions, we obtain the following theorem:

Theorem 1. *Define J the smallest j such that:*

$$\max_{i\leq j} \pi(a_i, a_i) > (1-\delta)\tilde{\pi}(a_{j+1}) + \delta\tilde{\pi} \tag{5}$$

In case Condition 5 is never met, set $J = m$. With a two-steps learning procedure with T_1 and T_2 large enough, the agents asymptotically set action

$$I \equiv \underset{i\leq J}{argmax}\, \pi(a_i, a_i) \tag{6}$$

Proof. The exploitation process, after T_1, starts with agents persistently choosing a_1. The Q-value converges to $\frac{\pi(a_1,a_1)}{1-\delta}$. If this value is greater than

$\tilde{\pi}(a_2) + \frac{\delta}{1-\delta}\tilde{\pi}$, which is the Q-value of action a_2, then a_1 is the point of convergence. Otherwise, a_2 becomes the new greedy action and is played until its Q-value possibly becomes lower than that of a_1 or a_3, which becomes the new greedy action. The process stops when condition (5) is met or when $J = m$. □

Convergence to an action that satisfies condition (5) is guaranteed whenever there exists $p \in P$ such that $\pi(p, p) > \tilde{\pi}$, which is the case in CCDP and subsequent literature. From Equation (6) we also note that agents cannot converge on a price lower than a_1. Theorem 1 also guarantees that the equilibrium profit is always greater than $(1-\delta)\min_p \tilde{\pi}(p) + \delta\tilde{\pi}$. Because $\tilde{\pi} = \max_p \tilde{\pi}(p)$, the condition (5) is more restrictive as δ increases, which means agents tend to converge on higher profits when they are more patient.

Though it may be tempting to interpret this phenomenon through the lens of collusion, we remind the reader that no reward-punishment scheme was needed to obtain this result, and neither was any specific information about past actions. Furthermore, we note that condition (5) still results in supra-competitive convergence profits even when agents are myopic (i.e., $\delta = 0$). This finding sheds light on an interesting result from CCDP, which noted that myopic agents may exhibit apparent collusion (Section 4.A in CCDP).

Although our Q-learners do not use a model of their environment, the high initial Q-values are reminiscent of Cooper et al. [2015], where correlated experiments lead to biased estimates of the parameters of their underlying model. Similarly, while our model addresses much more general demand settings than the prisoners dilemma, and uses the classic Q-learning and ϵ-greedy exploration policy, the exploitation phase shares characteristics with Hansen et al. [2021], where a deterministic exploration rule from the Upper Confidence Bound algorithm leads to simultaneous and symmetric moves from the third period onwards. In both papers, insufficient independent exploration are key assumptions that enable us to derive theoretical results.

In Appendix B, we show that convergence in the exploitation phase may occur surprisingly fast. After a short period of "fumbling," reflecting the agents' joint exploration of disappointing actions, the algorithms quickly settle on action a_I. This means that T_2 need not be very large: as we elaborate in Section 4 our empirical applications require only a few dozen periods of exploration for stateless algorithms and a few hundred periods for 1-step

memory algorithms for them to permanently set a_I.

3.4 Algorithms with memory in play mode: seeming punishments explained.

We now turn to the explanation for the emergence of apparent punishment schemes. Denote $a^{t,l}$ the action chosen by Agent l at time t. Following the literature, assume the state at time t is defined as the set of agents' previous action: $s_t = \{a^{t-1,1}, a^{t-1,2}\}$. The symmetry of the agents and the deterministic learning process exposed earlier allows us to define an apparent "reaction function" $r(\cdot)$ which maps the latest actions of the agents to an agent's next action. Using Theorem 1 with $I = J$ (as is the case in our applications), and recalling that the Q-values of non-greedy actions are not updated, it is straightforward to derive:

$$r(a^{t-1,1}, a^{t-1,2}) \equiv \begin{cases} a_{i+1} & \text{if } a^{t-1,1} = a^{t-1,2} = a_i \text{ with } i < I \\ a_I & \text{if } a^{t-1,1} = a^{t-1,2} = a_i \text{ with } i = I \\ a_1 & \text{if } a^{t-1,1} = a^{t-1,2} = a_i \text{ with } i > I \text{ or } a^{t-1,1} \neq a^{t-1,2} \end{cases} \tag{7}$$

The first two relations follow naturally from the mechanism outlined in the proof of Theorem 1. Understanding the final line requires recognizing that, during the exploitation phase, the Q-matrix of unvisited states remains unchanged. Consequently, any decisions made after an unexpected visit to these states will be the same as those that would have been made during the learning phase. Specifically, if one agent unilaterally deviates, both agents will respond by selecting the optimal action based on the Q-matrix in an asymmetric state. This Q-matrix was last updated during the learning phase, with the greedy action being a_1. As a consequence, action a_1 will be chosen again. This reasoning applies similarly to actions with $i > I$.

This means that any out-of-equilibrium deviation to an action other than a_I will result in the system converging again to a_I possibly after several steps that go through a_1, a_2... until agents choose a_I again. This rationalizes what has sometimes been described as a reward-punishment scheme, whereby a deviation is followed by several periods of bad outcomes until the system slowly returns to the pre-deviation prices. Instead of being a sign of agents'

sophisticated understanding of collusive strategies, our calculations shows that these may be a simple and mechanical consequence of the learning process, which, by construction, keeps a memory of past decisions. Further description and application of this phenomenon to CCDP are provided in Section 4.3.

3.5 Partial exploration

The present section assumed for simplicity that agents were using either full exploration ($\epsilon_t = 1$), or full exploitation ($\epsilon_t = 0$). While this allowed conveying the main intuition, we acknowledge that most exploration policies including that of CCDP use partial exploration ($0 < \epsilon_t \leq 1$).

In this situation, the key mechanism remains exactly same: starting from an inflated perception of expected profits due to simultaneous exploration. However, as the exploration rate decreases and both agents increasingly favor their greedy actions, their actual payoffs fall short of expectations. Over time, the previously second-best action may become the new greedy action for at least one agent. This may again lead to low payoffs for both agents, whether the shift results from a unilateral deviation or both agents switching to the second-best action simultaneously. Subsequently, the decision-making process progresses to consider third-best, fourth-best actions, and so on, until reaching action a_J, where joint play yields higher payoffs than the Q-values for other actions. In other words, the determination of the convergence point does not only require that unilateral deviations from the greedy action are profitable. Because of the symmetric exploration policy, convergence to a specific action also requires that the payoff from a joint deviation yields (weakly) higher profit than the prevailing best profit, a consideration which is absent from standard best-response analysis.

Due to partial exploration, not only is the Q-value of the greedy action continuously updated, but so too are those of non-greedy actions, which leads to cycles (see Banchio and Mantegazza [2022]). The unilateral deviations are explored with probability ϵ. Once the Q-value of the deviating action exceeds that of the prevailing greedy actions, it becomes the greedy action, which is then played with probability $1 - \epsilon$. Consequently, a profitable unilateral deviation with low joint profit becomes the greedy action with infinitesimal probability as ϵ approaches zero.

4 Numerical application

In this section, we show that our theory describes faithfully the observational facts of CCDP.

4.1 Convergence to high prices

Figure 2 shows the valuation of each action by the agents in the stateless version of CCDP, when the exploration rate is $\epsilon_t = 1$ for the first 1000 steps (exploration phase), and then drops to 0 (exploitation phase). At the end of the exploration period, we confirm that the algorithms' greedy action is 1.58 (thick blue line), which is indeed the initial prices of the simulations of CCDP and coincides with our theoretical value a_1. The expected value of this action decreases sharply as soon as the agents start the exploitation phase (vertical line). The figure shows that the Q-values of the successive greedy actions start declining as soon as they become the greedy action (the payoffs are "disappointing"). After less than 30 periods of exploitation, the agents converge on a greedy action of $p_{conv} = 1.74$ (thick orange line), which is also the median point of convergence in CCDP (Table 2 and 3).In that case, the realized payoff when both agents play p_{conv} exceeds that of the expected payoff, causing the Q-value of this action to increase and $p_{conv} = 1.74$ to remain the optimal action indefinitely.

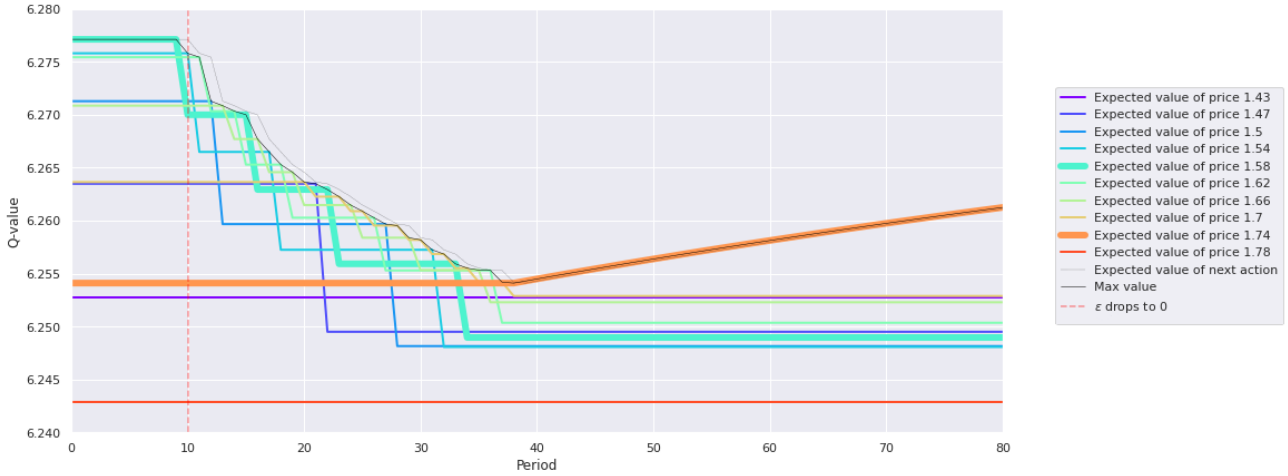


Figure 2: Q-value of actions over learning, with a 2 steps exploration policy (Stateless algorithms). Agents switch from full exploration to full exploitation at time $t = 10$.

The present section investigated the simpler case of memoryless algorithms, since we have already demonstrated that seemingly collusive out-

comes can occur in both memoryless and with-memory algorithms. To ensure completeness and comparability with CCDP, we also examine the CCDP case where agents observe past prices in Appendix D. As predicted by our theory, the results are qualitatively same, though convergence takes slightly longer due to a larger state space.

4.2 Extension to partial exploration

Lacking a formal proof, we now show numerically that our results are not driven by the two-steps approximation of the exploration policy. As we progressively increase the number of steps, we replicate the findings of CCDP in Figure 1(a). Figure 3 displays the average preferred price of agents over the course of learning when the exponential decline of the exploration rate is approximated using step functions. Specifically, Figure 3(b) corresponds to the 2-step exploration policy outlined in Section 3. To approach the exponential decay of exploration used in CCDP, the number of steps is gradually increased in Figures 3(d) through 3(j). In the final figure, which closely approximates CCDP's exploration policy, we observe a high degree of similarity between the replicated results and those of CCDP (shown in Figure 1(b)).

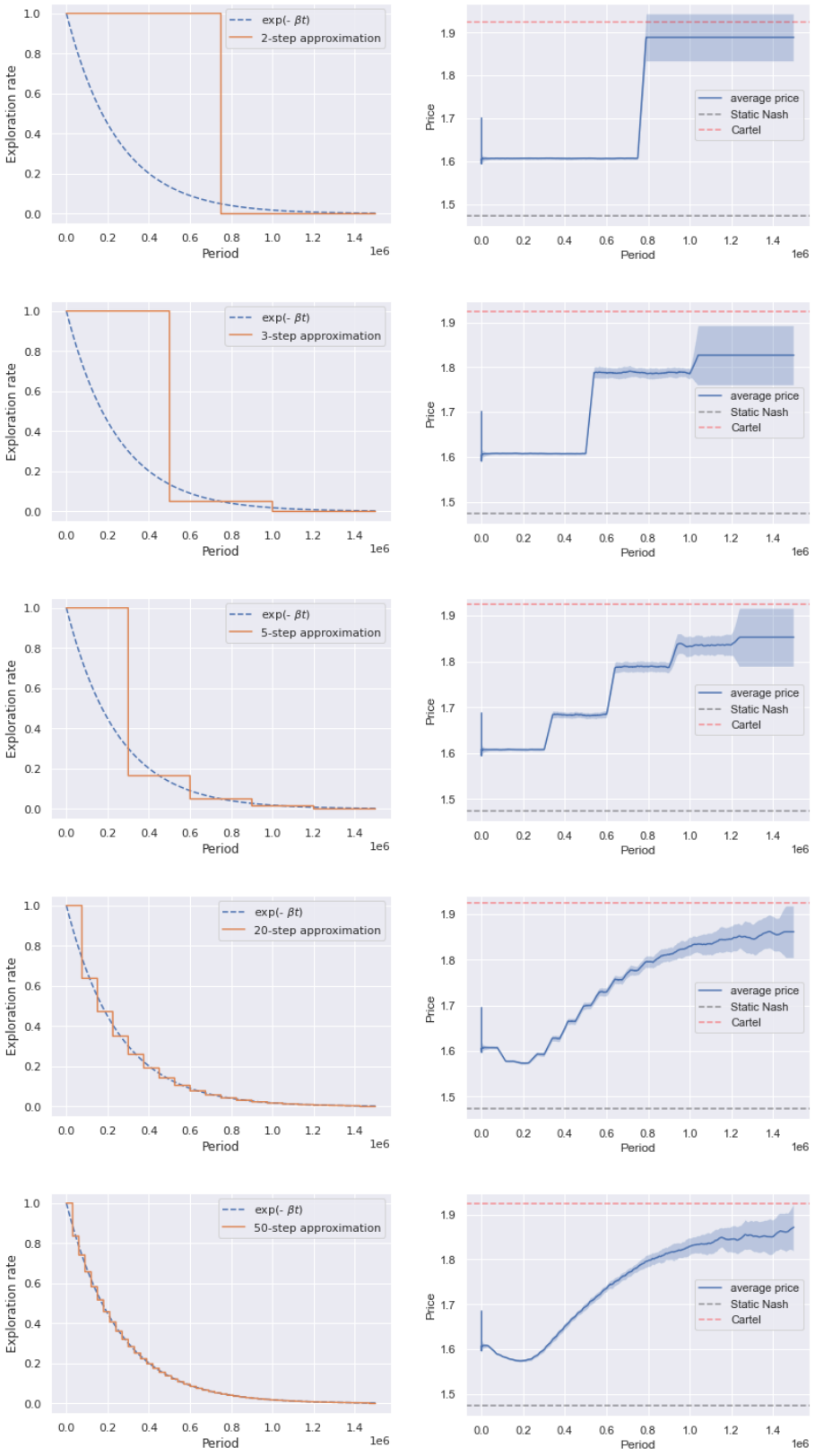
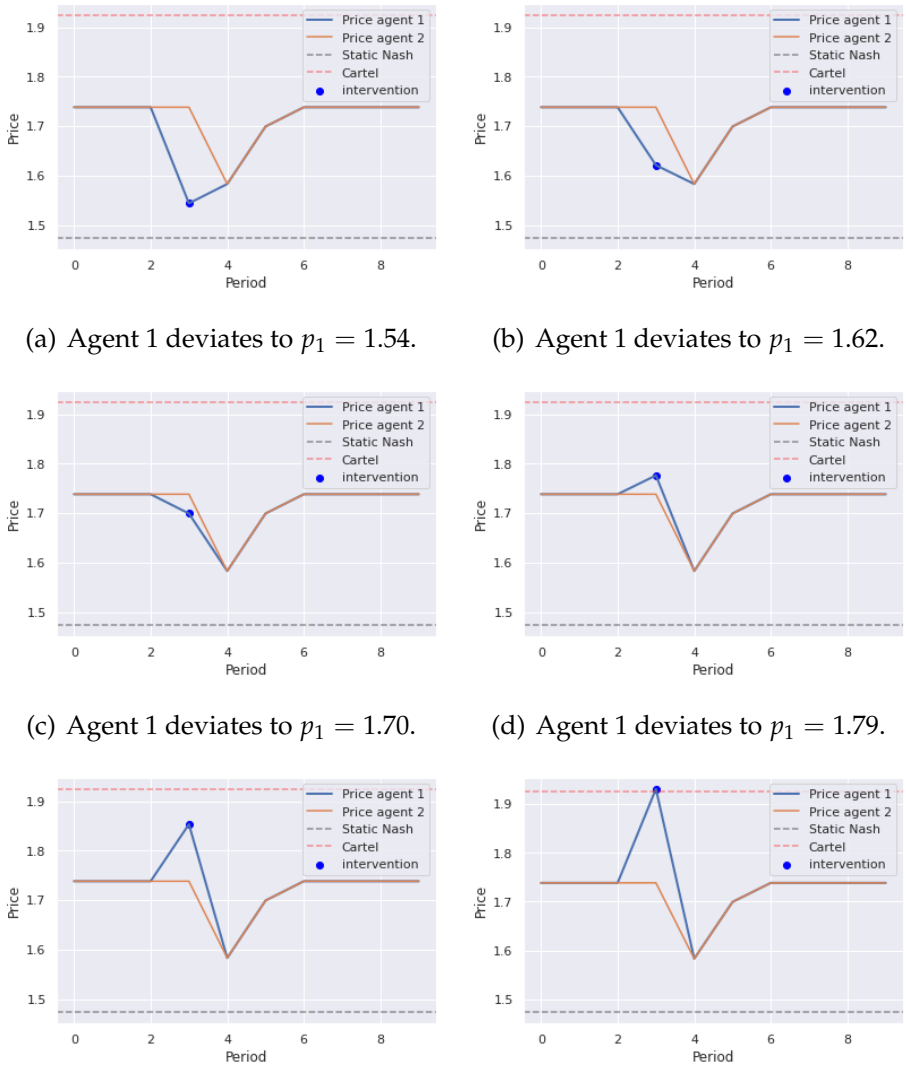


Figure 3: Right-hand side graphs show the greedy prices of the algorithms over 1.5 million learning episodes. Plots report the average over 1000 simulations. The corresponding exploration rate is on the left-hand side.

Again, the behavior of the algorithms has the appearance of collusion: we observe high prices, as well as pricing dynamics that replicate to some extent the punishment and reward strategies needed to sustain collusive prices according to economic theory. However, this appearance is misleading, in that it is not a sign of agent's ability to come up with a collusive agreement. Instead, it is a mechanical consequence of an imperfection in their learning process, which intrinsically embeds some inertia in valuations.

4.3 Seeming punishment schemes

The simple setting developed in Section 3 allows us to replicate the apparent reward and punishment schemes of CCDP. By construction indeed, the best response to any deviation from a_I to a_i with $i \neq I$ is a_1 (if only one agent has deviated or $i > I$) or a_{i+1} (if agents jointly deviated to a_i with $i < I$). Thus any unilateral deviation to an action other than a_I will cause the system to go through all actions $a_1, a_2, ..., a_I$. Of course, this concerns not only the negative price deviations (price cuts in Figures 4(a) to 4(c)) but also the positive ones (price increments in Figures 4(d) to 4(f)). The punishments for price increments may seem at odds with the interpretation of the phenomena as collusion. This issue is raised and described in further detail in Epivent and Lambin [2024], where it is also noted that the "punishment" action is always the same: it consists in setting a price of 1.58, which coincides with the punishment price found in the present paper. The present paper provides an explanation for this phenomenon, which is shown to be the result of a learning policy that freezes the valuation of actions over learning periods, even when those actions are no longer played. After the learning phase, the out-of-equilibrium deviations serve as a means of observing the learning process, as they allow us to witness the exploration steps that occurred during learning, even after learning is complete.



(a) Agent 1 deviates to $p_1 = 1.54$.

(b) Agent 1 deviates to $p_1 = 1.62$.

(c) Agent 1 deviates to $p_1 = 1.70$.

(d) Agent 1 deviates to $p_1 = 1.79$.

(e) Agent 1 deviates to $p_1 = 1.85$.

(f) Agent 1 deviates to $p_1 = 1.93$.

Figure 4: Prices chosen by the agents in play mode, when Agent 1 is forced to deviate at time $t = 3$ (blue dots).

The author of these lines does not believe these patterns can qualify as "collusive" reward-punishment schemes that sustain the high prices. Indeed these patterns are also observed when the discount factor δ is zero, a situation when collusion and the supporting reward punishment schemes are not an equilibrium. Most importantly, they are not needed to sustain the high prices, as Theorem 1 and its numerical application in Section 4.1 show. If anything, one may argue the reaction function induces the Q-values of non-preferred actions to consistently be driven even lower than those of the greedy action, but this scheme is pre-programmed through the specific

initialization of the Q-matrices, which biases the convergence price toward the best-responses to the other agent posting a random price.

5 Remedies to supra-competitive prices caused by simultaneous experiments

In Sections 3 and 4, we showed that algorithmic supra-competitive prices can stem from the combination of two factors: (1) simultaneous experiments during the exploration phase and (2) insufficient independent value updating during the exploitation phase. In this section, we verify empirically that altering either of these factors allows for convergence outcomes close to Nash. We then expose in detail the policy implications of our work.

5.1 Enforcing sequential entry

The first remedy targets the exploration phase. In the following experiment we allow only for sequential entry, in that entry may only occur once the incumbent has concluded their exploration phase. After entry, the incumbent may still update its Q-matrix and choose the greedy action accordingly, but is not allowed to explore. The average greedy price over this experiment is shown in Figure 5, with entry at period $T = 100,000$. We observe that when the entry in a market is sequential, agents converge to outcomes close to Nash.[9]

[9]Before entry the incumbent is a monopolist. Her problem is therefore much simpler than in the presence of a competitor who also experiments concomitantly. Therefore, to reduce computing time, we allow the exploration rate to decrease much faster than in the rest of the paper, for the purpose of this figure only. $\epsilon_t = exp(-\beta t)$ with $\beta = 4 * 10^{-5}$, instead of $\beta = 4 * 10^{-6}$ in the rest of the paper and in baseline CCDP.

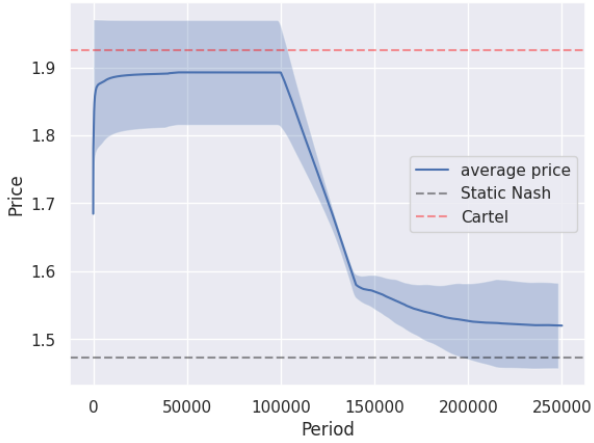
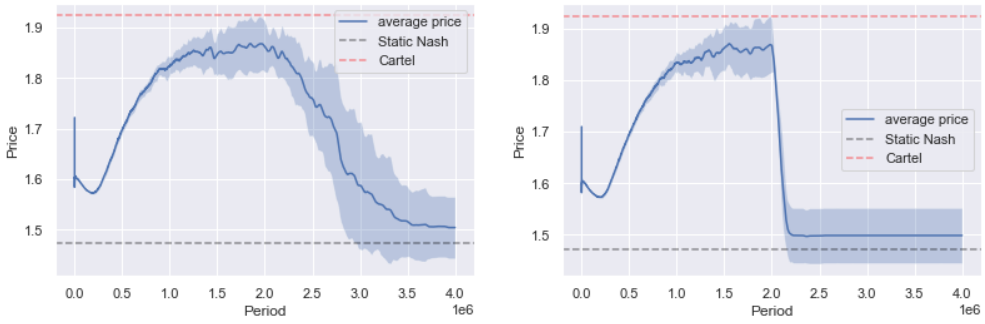


Figure 5: Incumbent cannot explore after entry, which occurs at period $T = 100,000$. Average price over 1000 simulations.

5.2 Enforcing "turn-based" exploration

The second remedy targets the exploitation phase. Figure 6 shows a situation when agents first behave as in CCDP for the first 2 million periods. As we discussed, at the end of this phase the agents converge to supra-competitive prices, and the exploration rate is virtually zero. However, from period 2 million onward, one of the agents is now constrained to explore (i.e. select an action randomly) with a small probability $\epsilon_{min} = 0.005$ (Figure 6(a)) or 0.05 (Figure 6(b)).[10] Again, we observe that the agents converge to outcomes close to Nash. A smaller exploration rate makes convergence slower but may also limit the inefficiency losses due to excessive exploration.



(a) $\epsilon_{min} = 0.005$ (b) $\epsilon_{min} = 0.05$

Figure 6: A minimal level of exploration is imposed on Agent 2 after the learning phase (from period T=2,000,000 onwards). Average price over 200 simulations.

[10]This is a simple application of "turn-based" exploration whereby agents explore one at a time, which formal definition and proof of convergence to Nash is provided in Appendix C.

5.3 Policy implications

The current paper shares with previous literature on AI supracompetitive pricing the concern that the use of basic AI algorithms may lead to abnormally high prices and low welfare. In addition, it sheds light on the mechanism behind at least some of the high prices observed.

Our first contribution, detailed in Section 2, challenges prevailing interpretations in the AI-Economics literature regarding algorithmic collusion. We challenge the idea that these algorithms genuinely collude by enforcing punishments for deviations, cautioning against qualifying a non-collusive mechanism as such. A proposed policy approach suggests simulating deviations for evidence (Calvano et al. [2020a]). Despite the merits and relative intuitiveness of this test, our research highlights the risk of regulators mistaking imperfect exploration for collusion, which may lead to inadequate antitrust remedies.

In a second contribution, we emphasize that the issue of supra-competitive profits may be more pressing than previously suggested. Researchers have noted that agents may require millions of iterations to reach supra-competitive outcomes, making the strategy unrealistically expensive, even for the largest platforms (Eschenbaum et al. [2022], Mellgren [2020], Meylahn et al. [2022]). Our research demonstrates that basic algorithms operating for a few hundred periods can yield similar results, making the strategy feasible for many firms operating in digital markets. This rapid convergence to supracompetitive prices may provide an explanation for the empirical findings in Assad et al. [2020], where high prices were observed alongside algorithm deployment. This phenomenon occurs through a much simpler mechanism than the classic notion of collusion.

There is however a silver lining. We showed that the supra-competitive prices may not have much to do with algorithmic sophistication, but are inherent to the rather simple learning process behind most reinforcement learning algorithms. As such, the issue is broader than previously thought, but may also be simpler to address. Algorithmic seeming collusion, as described in the existing economics literature, may not be a matter of antitrust. More broadly, it is an issue of market design. This leads to a simple policy recommendation: simultaneous learning from reinforcement-learning algorithms may need to be restrained: when agents of the same market explore in turn, they follow standard best-response dynamics.

6 Conclusion

We established that simple multi-agent reinforcement learning algorithms, when trained simultaneously, may swiftly converge on supra-competitive profits and policies that embed apparent price wars. We provided the first theoretical proof of the mechanism, for simplified though realistic algorithms already used in the literature. Contrary to the prevailing notion in the literature that collusive reward-punishment schemes drive the high prices, our research indicates that the root cause may be much simpler: simultaneous experiments, combined with a fundamental characteristic of reinforcement learning that involves keeping track, with some inertia, of the subjective expected value of future states. The convergence towards highly supra-competitive profits may occur much more quickly than previously believed, underscoring the need for greater scrutiny by regulators and researchers alike. These results have strong implications for the current discussions on antitrust for AI, and more importantly, for designing markets robust to AI decision-making.

Finally, this paper does not claim that the possibility of genuine autonomous collusion by AI algorithms, whereby algorithms independently, and with high probability, generate reward and punishment schemes that sustain high prices, is non-existent. Instead, our research demonstrates that there is currently no definitive evidence of such a phenomenon which, despite its plausibility, still lacks conclusive proof. Notably, Cartea et al. [2023, 2022] have taken interesting steps toward this direction. Although the assumptions remain fairly limiting, the core mechanisms they explore may inspire further exploration.

Acknowledgement

The author would like to express sincere gratitude to members of the research community who generously shared their time and engaged in insightful discussions on this research. The author also warmly thanks participants in various seminars and conferences, for providing valuable feedback. In particular, the Oxford-Man Institute Learning in Games and Algorithmic Collusion Workshop and the Roundtable on academic integrity (ESSEC) made invaluable contributions to this research. This project is supported by the French Government's Investment for the Future Program (PIA), now part of France 2030, operated by ADEME. The views expressed in this paper

are those of the author alone and might not represent the views of ADEME or any other institution. Remaining errors and omissions are my own.

References

I. Abada and X. Lambin. Artificial intelligence: Can seemingly collusive outcomes be avoided? *Management Science*, 2023.

I. Abada, X. Lambin, and N. Tchakarov. Collusion by mistake: Does algorithmic sophistication drive supra-competitive profits? *Available at SSRN 4099361*, 2022.

ACB. Autorité de la Concurrence Bundeskartellamt : Algorithms and Competition. (November), 2019. URL `www.bundeskartellamt.de`.

S. Albæk, P. Møllgaard, and P. B. Overgaard. Government-assisted oligopoly coordination? a concrete case. *The Journal of Industrial Economics*, 45(4): 429–443, 1997.

J. Asker, C. Fershtman, and A. Pakes. Artificial intelligence, algorithm design, and pricing. *AEA Papers and Proceedings*, 112:452–56, May 2022. doi: 10.1257/pandp.20221059. URL `https://www.aeaweb.org/articles?id=10.1257/pandp.20221059`.

J. Asker, C. Fershtman, and A. Pakes. The impact of artificial intelligence design on pricing. *Journal of Economics & Management Strategy*, 2023.

S. Assad, R. Clark, D. Ershov, and L. Xu. Algorithmic Pricing and Competition : Empirical Evidence from the German Retail Gasoline Market. *Working Paper*, 2020.

S. Assad, E. Calvano, G. Calzolari, R. Clark, V. Denicolò, D. Ershov, J. Johnson, S. Pastorello, A. Rhodes, L. Xu, et al. Autonomous algorithmic collusion: Economic research and policy implications. *Oxford Review of Economic Policy*, 37(3):459–478, 2021.

Autoridade Da Concorrència. Digital ecosystems, Big Data and Algorithms Issues Paper. *Autoridade Da Concorréncia*, (July), 2019. URL `http://www.concorrencia.pt/vEN/Estudos{_}e{_}Publicacoes/Estudos{_}Economicos/Other/Pages/Digital-Ecosystems-Executive-Summary{_}.aspx?lst=1`.

M. Banchio and G. Mantegazza. Adaptive Algorithms and Collusion via Coupling. *SSRN Electronic Journal*, 2022.

M. Banchio and A. Skrzypacz. Artificial intelligence and auction design. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, page 30â31, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391504. doi: 10.1145/3490486.3538244. URL https://doi.org/10.1145/3490486.3538244.

Z. Y. Brown and A. MacKay. Competition in pricing algorithms. *American Economic Journal: Microeconomics*, 15(2):109–156, 2023.

E. Calvano, G. Calzolari, V. Denicolò, J. E. Harrington, and S. Pastorello. Protecting consumers from collusive prices due to ai. *Science*, 370(6520), 2020a.

E. Calvano, G. Calzolari, V. Denicolò, and S. Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 2020b. ISSN 19447981. doi: 10.1257/aer.20190623.

E. Calvano, G. Calzolari, V. Denicolo, and S. Pastorello. Algorithmic collusion: Genuine or spurious? *International Journal of Industrial Organization*, 90:102973, 2023. ISSN 0167-7187. doi: https://doi.org/10.1016/j.ijindorg.2023.102973. URL https://www.sciencedirect.com/science/article/pii/S0167718723000541. The 49th Annual Conference of the European Association for Research in Industrial Economics, Vienna, 2022.

A. Cartea, P. Chang, J. Penalva, and H. Waldon. The algorithmic learning equations: Evolving strategies in dynamic games. *Available at SSRN*, 2022.

Á. Cartea, P. Chang, J. Penalva, and H. Waldon. A folk theorem from learning in games. In *A Folk Theorem from Learning in Games: Cartea, Álvaro | uChang, Patrick | uPenalva, José | uWaldon, Harrison*. [Sl]: SSRN, 2023.

A. C. Chapman, D. S. Leslie, A. Rogers, and N. R. Jennings. Convergent learning algorithms for unknown reward games. *SIAM Journal on Control and Optimization*, 2013. ISSN 03630129. doi: 10.1137/120893501.

Competition Bureau. Big data and innovation: Implications for competition policy in canada. Technical report, Discussion paper, 2018.

W. L. Cooper, T. Homem-de Mello, and A. J. Kleywegt. Learning and pricing with models that do not explicitly incorporate competition. *Operations Research*, 2015. ISSN 15265463. doi: 10.1287/opre.2014.1341.

A. Dolgopolov. Reinforcement learning in a prisoner's dilemma. *Available at SSRN 4240842*, 2021.

A. Epivent and X. Lambin. On algorithmic collusion and reward-punishment schemes. *Available at SSRN 4227229*, 2024.

N. Eschenbaum, F. Mellgren, and P. Zahn. Robust algorithmic collusion. *arXiv preprint arXiv:2201.00345*, 2022.

A. Ezrachi and M. E. Stucke. Sustainable and Unchallenged Algorithmic Tacit Collusion. *SSRN Electronic Journal*, 2018. doi: 10.2139/ssrn.3282235.

A. Ezrachi and M. E. Stucke. Sustainable and unchallenged algorithmic tacit collusion. *Northwestern Journal of Technology and Intellectual Property*, 17: 217–241, 2020.

M. S. Gal. Algorithms as illegal agreements. *Berkeley Technology Law Journal*, 34(1):67–118, 2019.

K. T. Hansen, K. Misra, and M. M. Pai. Frontiers: Algorithmic collusion: Supra-competitive prices via independent algorithms. *Marketing Science*, 40(1):1–12, 2021.

J. E. Harrington. Developing competition law for collusion by autonomous artificial agents. In *Journal of Competition Law and Economics*, 2019. doi: 10.1093/joclec/nhy016.

J. E. Harrington. The effect of outsourcing pricing algorithms on market competition. *Management Science*, 68(9):6889–6906, 2022.

M. Hettich. Algorithmic Collusion: Insights from Deep Learning. CQE Working Papers 9421, Center for Quantitative Economics (CQE), University of Muenster, Nov. 2021. URL `https://ideas.repec.org/p/cqe/wpaper/9421.html`.

J. Hu and M. P. Wellman. Multiagent reinforcement learning: Theoretical framework and an algorithm. *Proceedings of the fifteenth international conference on machine learning*, 1998. doi: 10.1.1.42.1313.

T. Jaakkola, M. I. Jordan, and S. P. Singh. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in Neural Information Processing Systems*, pages 703–710, 1994.

T. Klein. Autonomous algorithmic collusion: Q-learning under sequential pricing. *The RAND Journal of Economics*, 52(3):538–558, 2021. doi: https://doi.org/10.1111/1756-2171.12383. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1756-2171.12383.

K.-U. Kühn and S. Tadelis. The economics of algorithmic pricing: Is collusion really inevitable. *Working paper*, 2018.

K.-U. Kühn and X. Vives. *Information exchanges among firms and their impact on competition*. European Communities, 1995.

McSweeny and O'Dea. The Implications of Algorithmic Pricing for Coordinated Effects Analysis and Price Discrimination Markets in Antitrust Enforcement. *Antitrust*, 2017.

F. Mellgren. Tacit collusion with deep multi-agent reinforcement learning, 2020.

J. Meylahn, M. Schinkel, and A. Van den Boer. Artificial Collusion: Examining Supracompetitive Pricing by Q-Learning Algorithms. *SSRN Working paper*, 2022.

OECD. Algorithms and collusion: Competition policy in the digital age. Technical report, 2017.

N. Petit. Antitrust and Artificial Intelligence - A Research Agenda. *SSRN Electronic Journal*, 2017. ISSN 1556-5068. doi: 10.2139/ssrn.2993855.

L. Qiu, Y. Huang, and P. V. Singh. Product rankings, ai pricing algorithms, and collusion. *AI Pricing Algorithms, and Collusion (June 9, 2022)*, 2022.

B. Salcedo. Pricing Algorithms and Tacit Collusion. *Pennsylvania State University Working Papers*, 2015.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

U.S. v. Topkins. U.S. District Court, Northern District of California, 2023. URL https://www.justice.gov/atr/case-document/file/513586/download. No. 15-cr-00201 (Accessed October 30, 2023).

L. Waltman and U. Kaymak. Q-learning agents in a Cournot oligopoly model. *Journal of Economic Dynamics and Control*, 2008. ISSN 01651889. doi: 10.1016/j.jedc.2008.01.003.

Y. Wang and L. Pavel. A modified Q-learning algorithm for potential games. In *IFAC Proceedings Volumes (IFAC-PapersOnline)*, 2014. ISBN 9783902823625. doi: 10.3182/20140824-6-za-1003.02646.

C. J. C. H. Watkins. Learning from delayed rewards. 1989.

T. Werner. Algorithmic and human collusion. *Available at SSRN 3960738*, 2022.

X. Xu, S. Lee, and Y. Tan. Algorithmic collusion or competition: the role of platforms' recommender systems. *arXiv preprint arXiv:2309.14548*, 2023.

APPENDIX

A Replication of Klein [2021] and Calvano et al. [2023]

Figure 7 replicates the setting of Klein [2021]. The firms sell homogeneous goods with linear demand. They delegate their pricing decisions to Q-learning algorithms with 6 equally spaced feasible prices. The figure confirms the supracompetitive profits found by the author (black line). The blue line shows an extension using the same algorithms but without a memory. The profits are even greater in the latter case.

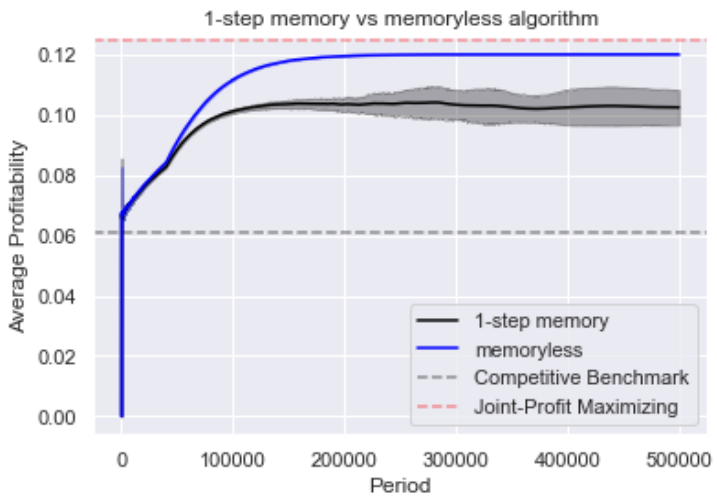


Figure 7: Black line: Replication of Klein [2021]. Blue line: Stateless version. Shaded areas show standard errors.

We now replicate the setting of Calvano et al. [2023]: a homogenous-good duopoly with rectangular demand. Consumers' willingness to pay is 2, and firms have a marginal cost of 1, which means that collusion is an equilibrium if and only if $\delta \geq 1/2$. Firms delegate their pricing decisions to Q-learning algorithms with 15 equally spaced feasible prices.[11] Figure 8 shows the final prices as a function of the discount factor, as in Calvano et al. [2023] (Figure 2, bottom). We confirm the supracompetitive profits found in Calvano et al. [2023] (black line). The blue line shows an extension with the same algorithms but without a memory. Notably, at the base discount rate

[11]This setting is intentionally similar to Asker et al. [2023]. The settings differ in the level of the willingness to pay of consumers, the marginal costs and the discretization of prices.

of $\delta = 0.95$, the profits are even greater when algorithms have no memory compared to when they do.

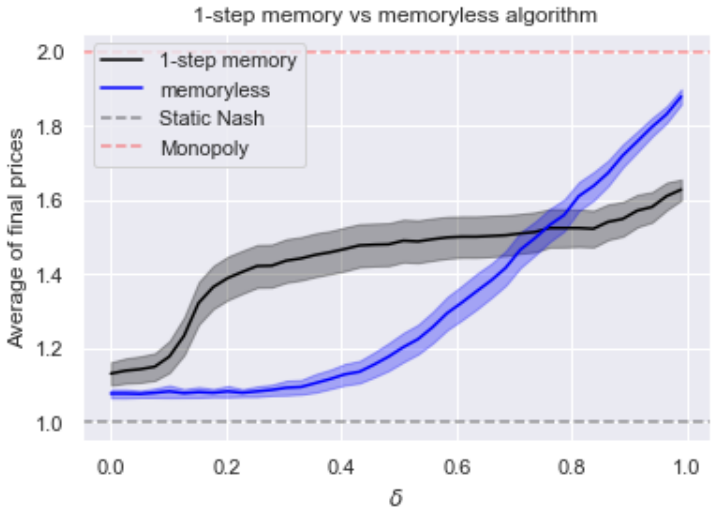


Figure 8: Black line: Replication of Calvano et al. [2023]. Blue line: Stateless version. Shaded areas show standard errors.

B Exploitation phase: counting the number of fumbling periods

As indicated in Theorem 1, the exploitation phase is characterized by a sequence of periods during which algorithms may opt for actions other than the ultimate choice a_I. We refer to this phenomenon as "fumbling," reflecting the agents' joint choice of actions that yield low payoffs before they settle on a_I. In this section, we show that fumbling may actually occur in a very short number of periods, meaning the algorithms may converge in actions very fast.

When actions are symmetric, as it turns out to be the case in practice in our setting, the Q-value of action a_i after $t(n_i)$ visits in the exploration phase, is an arithmetico-geometric sequence, with its progression indexed by n_i. Its value at each period can be expressed in a closed form:

$$Q^{t(n_i)}(a_i) = \alpha^{t(n_i)} \left(Q^{T_1}(a_i) - \frac{\pi(a_i, a_i)}{1-\delta} \right) + \frac{\pi(a_i, a_i)}{1-\delta} \tag{8}$$

By Theorem 1, Equation (6), for $i \neq I$ the total number of visits $\tilde{t}_i$ of second-best action a_i in the exploitation phase is such that $Q^{\tilde{t}_i}(a_i) < \frac{\pi(a_I, a_I)}{1-\delta}$. With

this and recalling the expression of $Q^{T_1}(a_i)$ in Lemma 1, we can express an upper bound for $\tilde{t}_i$:

$$\tilde{t}_i < \left\lfloor \frac{1}{ln(1/\alpha)} ln\left(\frac{(1-\delta)\tilde{\pi}(a_i) + \delta\tilde{\pi} - \pi(a_i, a_i)}{(1-\delta)(\pi(a_I, a_I) - \pi(a_i, a_i))} \right) \right\rfloor + 1 \tag{9}$$

where $\lfloor \cdot \rfloor$ denotes the integer part. The total number of periods of fumbling is bounded from above by :

$$T^{fum} \equiv \sum_{\substack{i=1 \\ i \neq I}}^{J} \tilde{t}_i \tag{10}$$

We observe that as the weight α given to the most recent observation approaches 1, the time spent fumbling T^{fum} tends towards $J-1$. On the other hand, when α is close to 0, T^{fum} diverges to infinity.[12]

C Turn-based exploration

For the purpose of finding a remedy to the issue of high algorithmic prices, we introduce the notion of "turn-based" exploration. It consists in an exploration policy when agents may only explore one at a time.

Definition 1 (Turn-based exploration)**.** *Take $\epsilon > 0$ and two agents indexed by 1 and 2, respectively. Denoting $\lfloor \cdot \rfloor$ the integer part, $\mod$ the modulo and $b \equiv \left\lfloor \frac{t}{T} \right\rfloor \mod 2$, turn-based exploration with cycles of length T are such that:*

$$a^{t,b} = \underset{a_i \in \mathcal{P}}{argmax} Q^{t,b}(a_i)$$

$$a^{t,1-b} \sim \begin{cases} \underset{a_i \in \mathcal{P}}{argmax} Q^{t,1-b}(a_i), & \text{with probability } 1-\epsilon \\ Uniform(\mathcal{P}), & \text{with probability } \epsilon \end{cases}$$

We operate under the assumption that a small level of exploration is maintained, ensuring that the order of Q-values remains consistent with that of zero exploration.

Definition 2 (small exploration rate)**.** *The exploration rate ϵ is positive and small enough that, for all actions of the other agent, the order of the payoffs of actions*

[12] In practical applications, Q-learning algorithms typically use α values between 0.05 and 0.2, as noted in Sutton and Barto [2018]. The CCDP algorithm, which we replicate in Section 4, uses a constant learning rate of 0.15.

coincides with the order of the payoffs of actions when there is no exploration.

We also define $\underline{\eta}_\epsilon$ and $\overline{\eta}_\epsilon$, respectively the minimum and maximum difference between the expected payoffs of two distinct actions:

$$\underline{\eta}_\epsilon \equiv \min_{i \neq j,l} |\pi_\epsilon(a_i, a_l) - \pi_\epsilon(a_j, a_l)|$$
$$\overline{\eta}_\epsilon \equiv \max_{i \neq j,l} |\pi_\epsilon(a_i, a_l) - \pi_\epsilon(a_j, a_l)|$$

Our result presented in Theorem 2 provides an interesting contrast to the findings of Section 3.3. It shows that when (1) a small exploration rate is maintained for a sufficiently long time, (2) agents converge to a recurrent class of size one (fixed point), and (3) they explore in turn, then the agents converge on a Nash equilibrium.

Theorem 2. *Assume the exploration rate is small and agents explore in turn over periods of length* $T > \frac{ln(\frac{2\overline{\eta}_\epsilon}{\underline{\eta}_\epsilon})}{ln(\frac{1}{\alpha})}$. *If the agents converge to fixed points, they converge to the Nash equilibrium of the static game.*

Proof. Assume without loss of generality that at time t, Agent 1 starts exploring while Agent 2 cannot and keeps playing a^2. After T periods, the Q-values of Agent 1 are such that:

$$Q^{t+T}(a_i) = \alpha^T \left(Q^t(a_i) - \frac{\pi(a_i, a^2)}{1-\delta} \right) + \frac{\pi(a_i, a^2)}{1-\delta}$$

from which we derive that

$$|Q^{t+T}(a_i) - \frac{\pi(a_i, a^2)}{1-\delta}| < \frac{\underline{\eta}_\epsilon}{2}$$

This last expression guarantees that the Q-values of actions are in the same order as the expected profits if there were no exploration. In particular, this means that the greedy action a^1 of Agent 1 is the best response to action a^2. We can then apply the same reasoning to Agent 2, who chooses the best response to Agent 1's action, and so on. Therefore we must have that a^2 is also a stage-game best response to a^1. In turn, if agents converge to a fixed point, (a^1, a^2) is a Nash equilibrium of the stage game, which exists by assumption. □

Theorem 2 is similar in spirit to some results of the computer science literature (see e.g., Hu and Wellman [1998], Chapman et al. [2013], Wang

and Pavel [2014]), though we use a much simpler exploration policy and constrain the algorithms to explore in turn, as opposed to simultaneously. It proves that when an exploration rate close to 0 is maintained sufficiently long and agents explore in turn, agents converging on a fixed point converge on a Nash equilibrium when it exists.[13] The result of Theorem 2 is exploited to construct and test a simple remedy to algorithmic supra-competitive prices in Section 5.

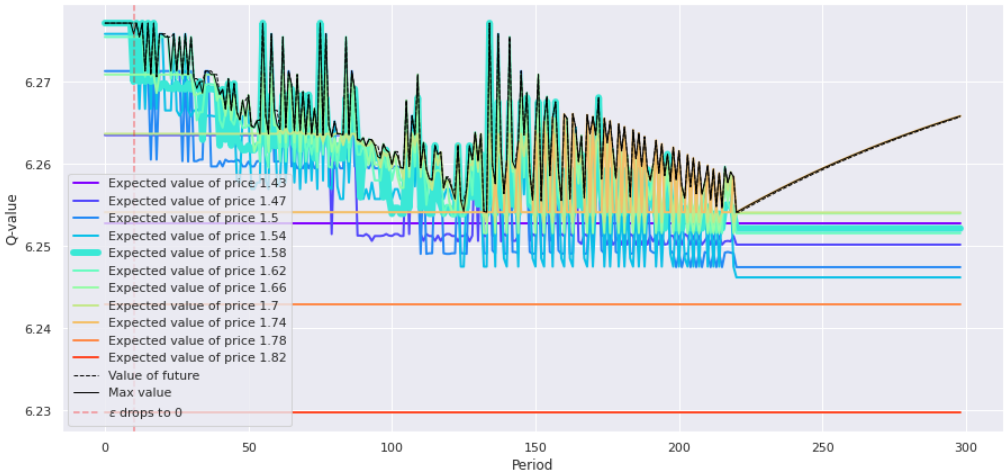
D Robustness check: algorithms with a memory

Section 4 shows that the behavior of AI agents can be represented accurately with memoryless agents. For the sake of completeness, we report here the results for agents with a 1-step memory, which is the specific setting of CCDP: the state of the agents at time t is the couple of prices chosen by the agents at time $t-1$.

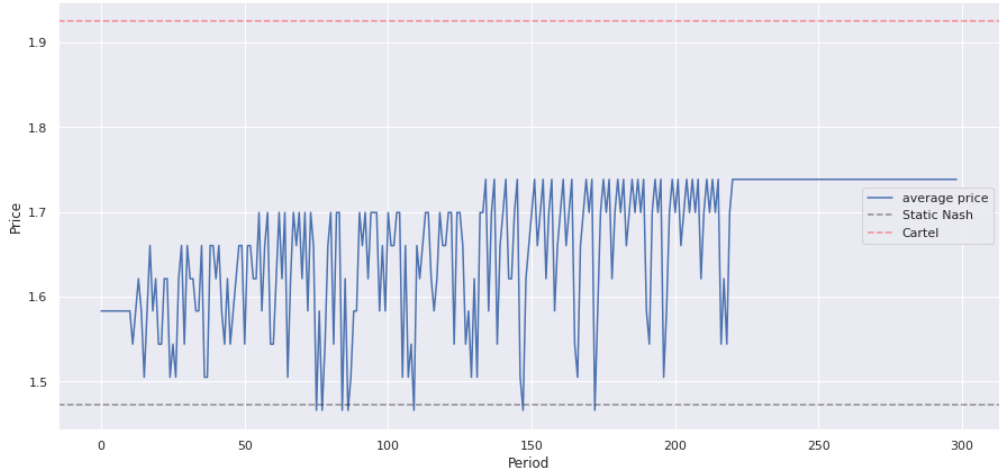
D.1 Analysis of the learning process

Figure 9(a) shows the valuation of each action by agents endowed with a 1-step memory, when the exploration rate is at 1 for 1000 steps (exploration phase), and then drops to 0 (exploitation phase). At the end of the exploration period, the agents' greedy action is $a_1 = 1.58$. After about 250 periods of exploitation, the agents converge on a greedy action of $p_{conv} = 1.74$. The results are similar to Figure 2, even though they are noisier and convergence is much slower due to the increased size of the state space.

[13] The theorem focuses on the cases when agents converge on a fixed point (constant action), as is observed in 64 % of the simulations in CCDP. In the context of a prisoner's dilemma and assuming near-symmetry of Q-values among agents throughout learning, Banchio and Mantegazza [2022] provides a rationale for Q-learners converging on the same estimates for both cooperation and defection, resulting in cyclic patterns that involve the cooperative action.



(a) Stateless version of CCDP, with 2 steps exploration policy.



(b) Price played

Figure 9: Valuation of actions over learning, with a 2 steps exploration policy. Agents switch from full exploration to full exploitation at time $t = 10$. Agents have a 1-step memory.

Figure 9(b) displays the preferred price of the agents over periods. From the beginning of the exploitation phase until period $t = 224$, a noisy pattern arises due to the execution of several actions that fail to satisfy Equation (5) until their Q-value drops below that of another action. Additionally, because the value of the learning rate α is less than 1, multiple rounds of exploitation may be necessary before the Q-values of the optimal actions reach their true value, which may involve alternating between several actions with an index $i < I$.

D.2 Smoothing the exploration policy towards exponential decay

In Figure 10, we show the counterpart to Figure 3 for the case of algorithms with a 1-step memory (i.e. we replicate exactly the setting of CCDP): the state space at time t is $(p^{t-1,1}, p^{t-1,2})$. Again, we observe in the last figure, which corresponds to a very close approximation of the exploration policy of CCDP, that we replicate perfectly the results of CCDP (reported in Figure 1(a)).

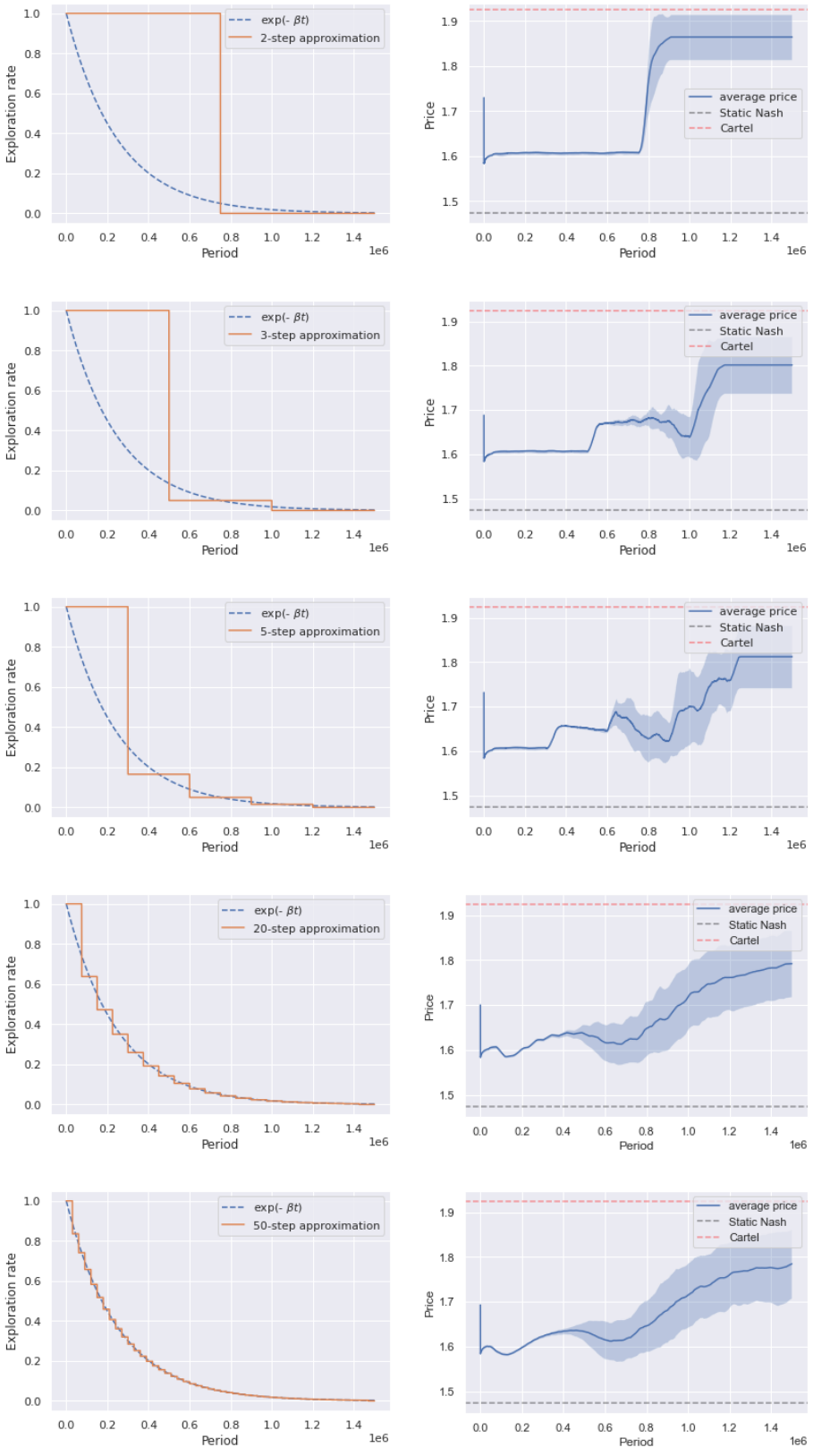


Figure 10: Right-hand side graphs show the greedy prices of the algorithms over 1.5 million learning episodes. Plots report the average of 1000 simulations. The corresponding exploration rate is on the left-hand side. Algorithms with 1-step memory.