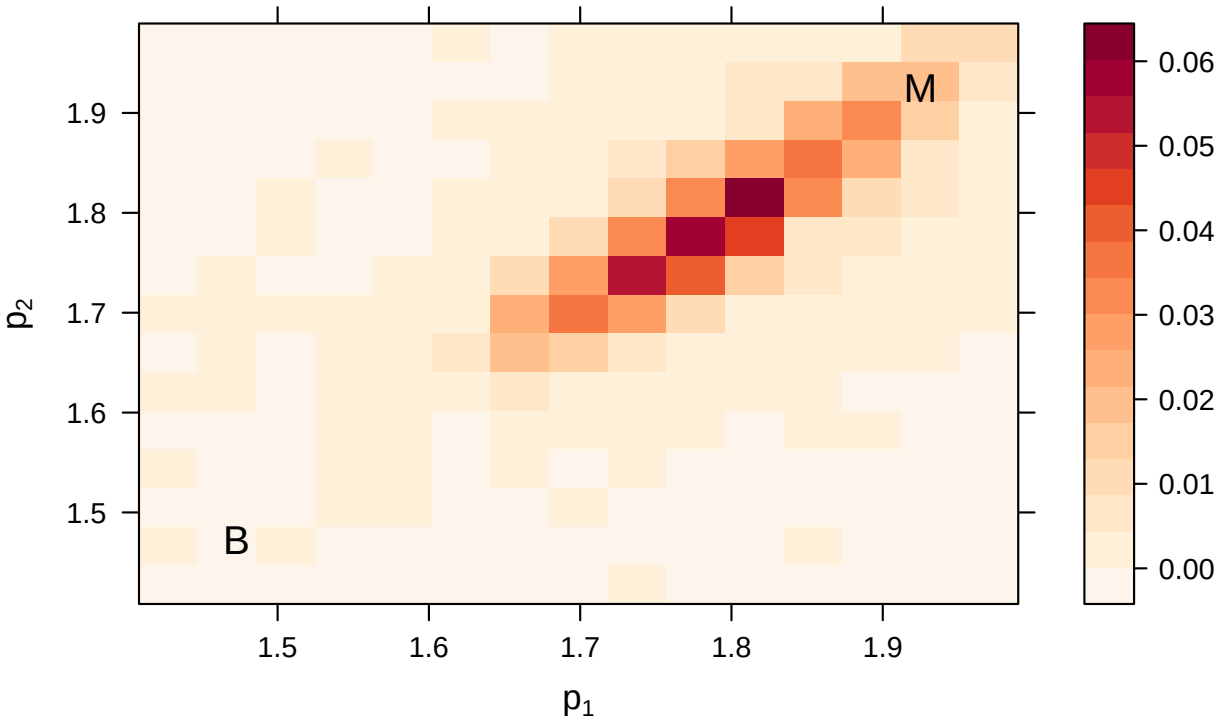


Figure 3: Distribution of states visited in the last 100,000 iterations pooling all data from 1000 sessions with $\alpha = 0.15$, $\beta = 4 \times 10^{-6}$. "M" corresponds to the fully collusive price, "B" to the Bertrand equilibrium price.

intermediate level (partial collusion). This is confirmed by Figure 3, which shows the relative frequency of the different price combinations eventually charged for a representative experiment. Prices are rarely as high as under monopoly but are almost always higher than in the Bertrand-Nash equilibrium. Price dispersion is low, and firms tend to price symmetrically.

In sum, our algorithms consistently and symmetrically charge supra-competitive prices, obtaining a sizable profit gain.

In view of the robustness of the results with respect to the learning and experimentation parameters, to ease exposition we shall henceforth focus on one representative experiment, corresponding to $\alpha = 0.15$ and $\beta = 4 \times 10^{-6}$. (This is the experiment illustrated in Figure 3.) With these parameter values, sub-optimal cells are visited on average about 20 times, and the initial Q-value of such cells counts for just 3% of their final value. A Nash equilibrium is learned 54% of the times, and the average profit gain is 85%. None of these values seems extreme. But at any rate, we have systematically conducted robustness analyses with respect to α and β not only for the baseline case but also for the extensions considered below so as to confirm that our results are not sensitive to these parameters.

5.6. *Strategies*

Let us now turn to the issue of what strategies underpin the documented non-competitive outcomes. The key question is whether the high prices are the result of the algorithms' failure to learn the static Bertrand-Nash equilibrium or of genuine collusion. The policy implications would be radically different: the former means that the algorithms are not smart enough, the latter that they are already, in a sense, "too smart." As AI technology advances, in the former case the problem is likely to fade away; in the latter, to worsen.

At this point, it may be useful to spell out exactly what we mean by collusion. Following Harrington (2017), we define collusion as "a situation in which firms use a reward-punishment scheme to coordinate their behavior for the purpose of producing a supra-competitive outcome." That is, what is crucial is not the level of profits as such but the way the supra-competitive result is achieved. Even extremely high profit levels may be regarded as collusive only insofar as deviations that are profitable in the short run would trigger a punishment.

That Q-learning algorithms actually do learn to collude is suggested by the fact that the profit gain tends to increase with the amount of equilibrium play. But the correlation is far from perfect,[34] so here we set forth two additional, perhaps more compelling arguments. First, in economic environments where collusion cannot arise in equilibrium, we find that the algorithms learn instead to set competitive prices. Second, going back to settings where collusion is possible, we consider exogenous defections and observe how the algorithms react. We find that such defections gets punished, and that the punishment makes the defections unprofitable. This is perhaps the most direct possible evidence of collusive behavior where collusion is tacit.

5.6.1. *Competitive environments*

In certain environments, collusion is impossible by default; in others, it can never arise in equilibrium. If the supra-competitive prices that we find were the result of erratic choices, or of something other than collusion, then they should also be expected in settings where collusion is impossible. In particular, collusion is impossible when $k = 0$ (the algorithms have no memory), and it cannot arise in equilibrium when $\delta = 0$ (the immediate gain from defection cannot be outweighed by the loss due to future punishments). As noted in Section 3, the previous literature has indeed found supra-competitive prices even in these

[34] In particular, increasing exploration initially improves learning and increases profits but eventually backfires; the same is true for increasing α. But while for the profit gain the downside of decreasing β and increasing α is already evident in Figures 1 and 2, for equilibrium play it only appears for values of α and β that lie outside of the interval shown in the figures.

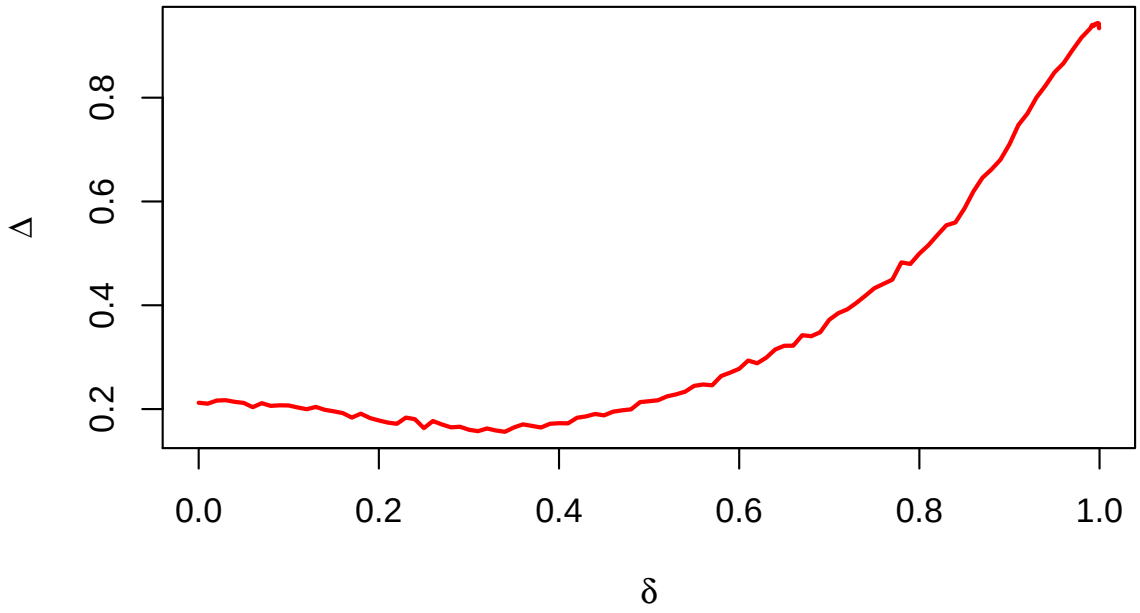


Figure 4: Average Profit Gain Δ for $\alpha = 0.15, \beta = 4 \times 10^{-6}$

settings, which poses the question of how to interpret the findings in economic terms.

In this respect, our results are quite different.[35] Consider first what happens when the algorithms get short-sighted. Figure 4 shows how the average profit gain varies with δ. The theoretical postulate that lower discount factors, i.e. less patient players (or else less frequent interaction), impede collusion, is largely supported by our simulations. The profit gain indeed decreases smoothly as the discount factor falls, and when $\delta = 0.35$ it has already dropped from over 80% to a modest 16%.

At this point, however, something perhaps surprising happens: the average profit gain starts increasing as δ decreases further. Although the increase is small, it runs counter to theoretical expectations. This "paradox" arises because changing δ affects not only the relative value of future versus present profits, but also the effective rate of learning. This can be seen from equation (4), which implies that the relative weight of new and old information depends on both α and δ.[36] In particular, a decrease in δ tends to increase the effective speed of the updating, which as noted may impede learning when exploration is extensive.[37] Figure 4 suggests that if one could abstract from this spurious effect, collusion

[35]The difference might be due to the fact that we use a different economic model, and our algorithms are allowed to learn more effectively, than in previous studies.

[36]Loosely speaking, new information is the current reward π_t, and old information is whatever information is already included in the previous Q-matrix, $\mathbf{Q}_{t-1}$. The relative weight of new information in a steady state where $Q = \frac{\pi}{1-\delta}$ then is $\alpha(1-\delta)$.

[37]A similar problem emerges when δ is very close to 1. In this case, we observe another "paradox," i.e.,

would tend to disappear when agents become short-sighted.

Turning to the case of memoryless algorithms, we find profit gains of less than 5%. These are almost negligible and do not vanish altogether simply because our discretization of the strategy space implies that the one-shot Bertrand equilibrium can at best be approximated.

All of this means that the algorithms do learn to play the one-shot equilibrium when this is the only equilibrium of the repeated game. If they do not play such competitive equilibrium when other equilibria exist, it must be because they have learned more sophisticated strategies.

5.6.2. *Deviations and punishments*

To look into how these strategies are structured, we perturb the system once the learning is completed. That is, upon convergence we step in and manually override one agent's choice, forcing it to defect. We impose not only defections lasting for a single period but also defections lasting several periods; and defections both to the static best-response and to smaller price cuts. For all of these cases, we then examine the reaction of both agents in the subsequent periods. In a word, we derive impulse-response functions.

Figure 5 shows the average of the impulse-response functions derived from this exercise for all 1000 sessions of our representative experiment. It shows the prices chosen by the two agents τ periods after the deviation. In particular it depicts the evolution of prices (top) and profits (bottom) following agent 1's one-period deviation to the static best-response to the pre-deviation price.

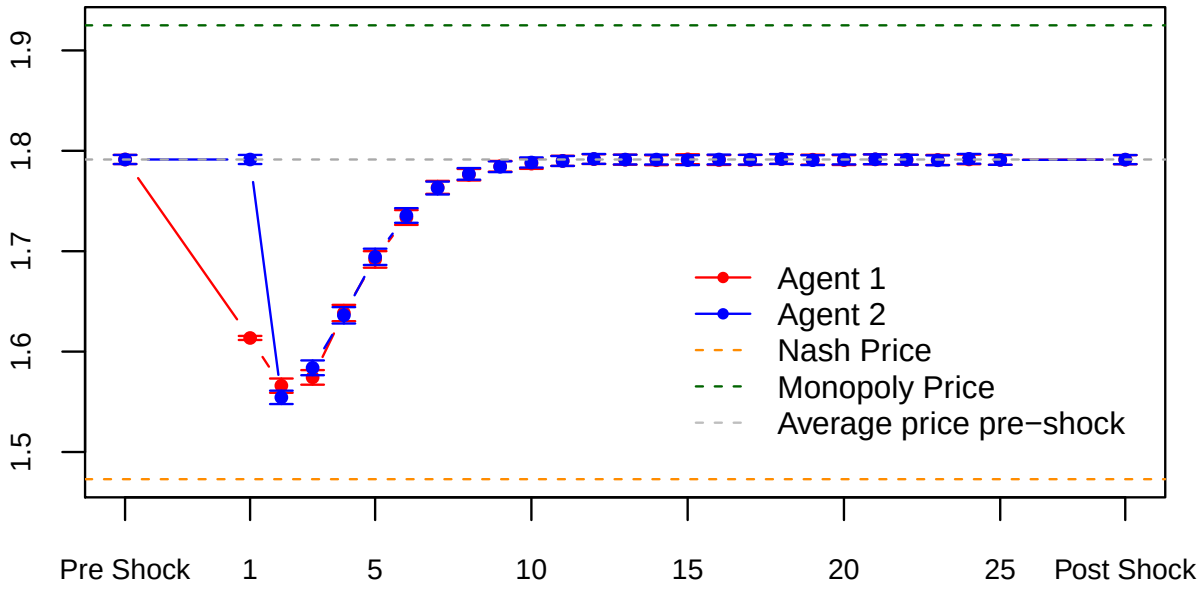
Clearly, the exogenous deviation gets punished. The punishment is not as harsh as it could be (e.g., a reversion to the static Bertrand-Nash equilibrium), and it is only temporary: in the subsequent periods, the algorithms gradually return to their pre-deviation behavior.[38] This pattern seems natural for Q-learning algorithms. These algorithms experiment widely, at least initially, so it would be difficult to sustain collusion if any defection triggered a permanent switch to the one-shot Bertrand equilibrium.

But in any case, the punishment is harsh enough to make the deviation unprofitable as illustrated by the evolution of profits after the shock in Figure 5 (bottom). Evidently, agent 2's retaliation wipes out agent 1's profit gain already in the very next period: that is, incentive compatibility is verified.

---

the average profit gain eventually starts decreasing with δ. This failure of Q-learning for $\delta \approx 1$ is well known in the computer science literature.

[38] In fact, prices stabilize on a new level that on average is slightly lower than the pre-deviation one.



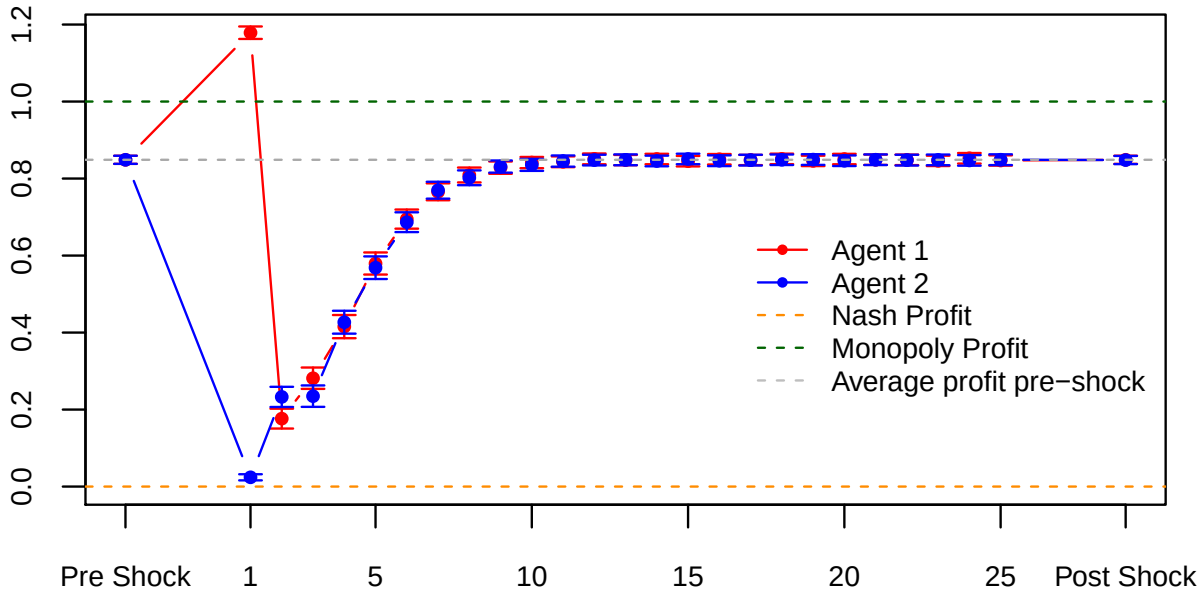


Figure 5: Prices (top) and Profits (bottom) impulse response, $\alpha = 0.15, \beta = 4 \times 10^{-6}, \delta = 0.95$.

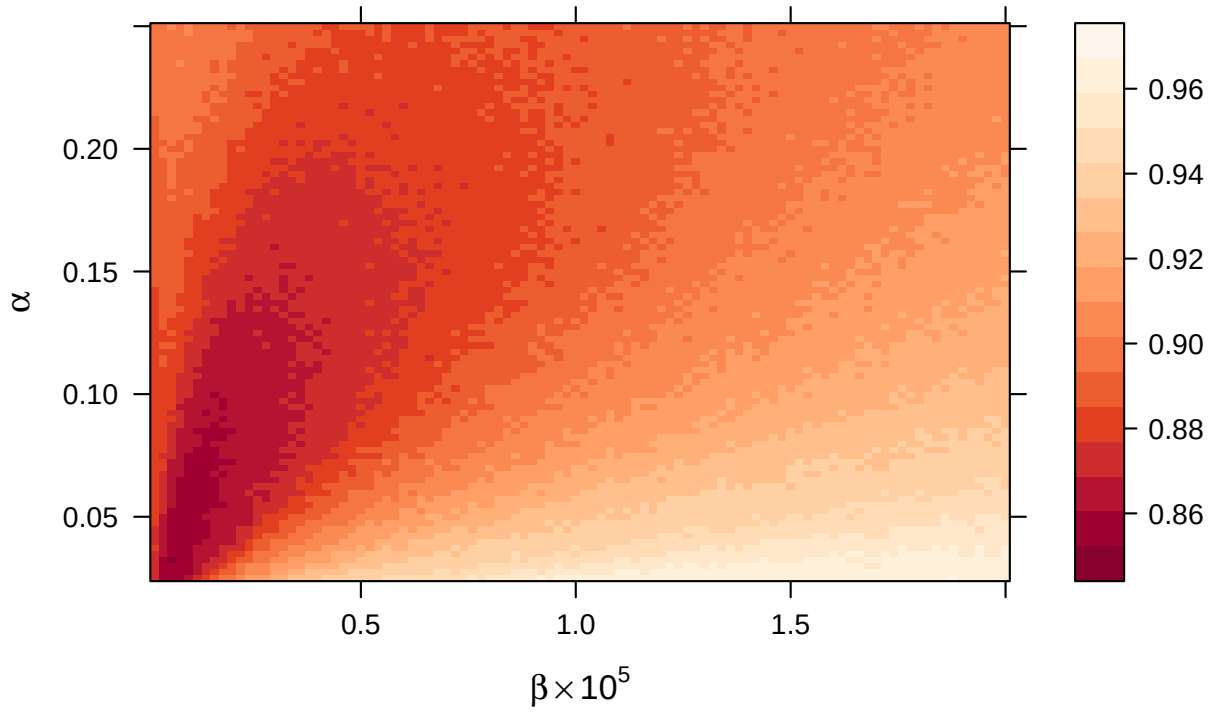


Figure 8: Average % price drop of non deviating agent 2 one period after deviation.

and $\beta = 4 \times 10^{-6}$.

6.1. *Number of players*

Theory predicts that collusion is harder to sustain when the market is more fragmented. This is indeed one reason why antitrust authorities regulate mergers: apart from the unilateral effects, the concern is that more concentrated markets may be conducive to tacit collusion.

The experimental literature provides some support for this thesis, showing that without explicit communication it is very difficult for more than two agents to collude. With three agents, prices are typically pretty close to the static Bertrand-Nash equilibrium, and with four agents or more they may even be lower.[40]

In this respect, Q-learning algorithms would appear to be different. In simulations with three firms, the average profit gain Δ decreases from 85% to 64%. With four agents, the profit gain is still a substantial 56%. The fact that Δ decreases accords with the theoretical predictions; the fact that it decreases so slowly seems to be a peculiar and

[40]See for instance Huck et al. (2004). Cooper and Khun (2014) stress the importance of communication to sustain collusion in the lab.

In principle, the impact of changes in substitutability on the likelihood of collusion is ambiguous: on the one hand, when products are more substitutable the gain from deviation increases, but at the same time punishment can be harsher. This ambiguity is confirmed by the theoretical literature.[43]

In our setting, we test the consequences of changing parameter μ from 0.25 (baseline) up to 0.5 and down to 0.01, where products are almost perfect substitutes. The average profit gain decreases slightly when μ decreases, but when the products are almost perfect substitutes ($\mu = 0.01$) it is still greater than 80%.

6.7. *Linear demand*

We repeated the analysis for the case of duopoly with linear demand functions derived from a quadratic utility function of the Singh and Vives (1984) type, i.e.

$$u = q_1 + q_2 - \frac{1}{2}(q_1^2 + q_2^2) - \gamma q_1 q_2 \tag{11}$$

for various values of the horizontal differentiation parameter γ. The average profit gain is non monotone in γ: it is well above 80% when γ is below 0.4 or above 0.9 (when the products are fairly good substitutes) but reaches a minimum of 65% when γ is around $\frac{3}{4}$. The impulse-response functions are almost identical to those observed with logit demand. Other alternative demands could also be tested, of course, but it would seem that the results are robust to the demand specification.

6.8. *Finer discretization*

As Section 3 notes, one reason why the repeated prisoner's dilemma may be a misleading approach to the analysis of collusion is that coordination is much easier when there are only few possible actions. With two actions only, for instance, there is basically only one way to cooperate. As the number of actions and hence strategies increases, the strategy on which the players should coordinate is no longer self-evident. Failing explicit communication, this could engender misunderstandings and prevent cooperation.

To determine whether a richer strategy space facilitates or impedes cooperation, we run experiments with 50 or 100 feasible prices, not just 15. The average profit gain decreases slightly, but with $m = 100$ it is still a substantial 70%. We conjecture that Δ decreases because a larger number of actions and states necessitates more exploration to achieve

[43]For example, Tyagi (1999) shows that greater substitutability hinders collusion when demand is linear or concave but may either hinder or facilitate it when demand is convex.

the same amount of learning as in baseline model.

We have also considered the case of parameter ξ higher than 10%, but greater flexibility in price setting - below Bertrand-Nash or above monopoly - turns out to be immaterial. This is not surprising, given that the players never converge on these very low or very high prices.

6.9. *Asymmetric learning*

Going back to the baseline environment, we consider the case in which the two algorithms have different learning rates α, or different intensity of experimentation. Collusion appears to be robust to these changes. For example, when α_2 is set at 0.05 or 0.25 with α_1 constant at 0.15, the average profit gain is 82% in both cases. In both cases, the firm with lower α gains more, suggesting that setting $\alpha = 0.15$ still gives too much weight to new information.

We also halved and doubled the value of β for one algorithm only, keeping the other fixed at $\beta = 4 \times 10^{-6}$. In both cases, asymmetry reduces the average profit gain but only marginally, remaining well above 75%. The algorithm that explores more underperforms.

These highly preliminary results for situations in which different algorithms follow different learning and exploration strategies suggest that diversity among the algorithms does not significantly affect the degree of collusion.

7. CONCLUSION

We have shown that in stationary environments Q-learning pricing algorithms systematically learn to collude. Collusion tends to be partial and is sustained by punishment in case of defection. The punishment is of finite duration, with a gradual return to pre-deviation prices. The algorithms learn to play these strategies by trial and error, requiring no prior knowledge of the environment in which they operate. They leave no trace whatever of concerted action: they do not communicate with one another, nor have they been designed or instructed to collude.

From the standpoint of competition policy, these findings should clearly ring a bell. They suggest that with the advent of Artificial Intelligence and algorithmic pricing, tacit collusion may become more prevalent, heightening the risk that tolerant antitrust policy may produce too many false negatives and so possibly calling for policy adjustments.

However, more research is needed to confirm the external validity of our findings. Three issues stand out: the realism of the economic environment, the speed of learning, and the

diversity of the competing algorithms. As for the first issue, we have considered a good many extensions of the baseline model, but all separately, the model thus remaining quite highly stylized. To produce a more realistic setting for analysis, one should perhaps posit a model with several firms, longer memory, stochastic demand and possibly also structural breaks.

Such more realistic environments may however defy the learning capacity of our simple, tabular Q-learners. It might therefore be necessary to use algorithms whose learning is more efficient - say, deep learning algorithms. This point is related to the second issue mentioned above, i.e. the speed of learning. Besides managing more complex environments, deep learning can indeed also speed the learning process. This is important, because the training of the algorithms cannot always be conducted entirely off-line, and in the short run experimentation is costly. On-the-job learning seems necessary, in particular, when the economic environment is changeable, or when the training environment does not exactly reflect the reality of the markets where the algorithms are eventually deployed.

One reason why training environments may not be fully realistic is that it is difficult to guess what specific algorithms competitors are using. In this respect, here we have restricted attention mostly to training in self-play mode. In reality, there are many different forms of reinforcement learning, and Q-learning algorithms themselves come in different varieties. It would therefore seem necessary to extend the analysis to the case of heterogeneous algorithms more systematically. Our robustness exercises in this direction have just scratched the surface of the problem.

Addressing these issues is clearly an important task for future work. But whatever may be done or left undone in the abstract, skeptics may always doubt that algorithms actually collude in the real world. Ultimately, the issue can only be decided by antitrust agencies and the courts. Unlike academic scholars, they can subpoena and extensively test the algorithms firms actually use, in environments that closely replicate the specific industry under investigation. What academic research can do is help make a preliminary assessment: that is, whether opening such investigations is a waste of resources, perhaps with the risk of many false positives, or instead may be necessary to fight collusion in the age of Artificial Intelligence. This paper is one contribution in this direction.

REFERENCES

[1] Andreoli-Versbach, P. and U. Franck, J. (2015). Econometric Evidence to Target Tacit Collusion in Oligopolistic Markets. *Journal of Competition Law and Economics*, **11** (2), 463–492.

[2] Arthur, W. B. (1991). Designing Economic Agents that Act like Human Agents: A Behavioral Approach to Bounded Rationality. *The American economic review*, **81** (2), 353–359.

[3] Barfuss, W., Donges, J. F. and Kurths, J. (2019). Deterministic limit of temporal difference reinforcement learning for stochastic games. *Physical Review E*, **99** (4), 043305.

[4] Barlo, M., Carmona, G. and Sabourian, H. (2016). Bounded memory Folk Theorem. *Journal of economic theory*, **163**, 728–774.

[5] Beggs, A. W. (2005). On the convergence of reinforcement learning. *Journal of economic theory*, **122** (1), 1–36.

[6] Benveniste, A., Metivier, M. and Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximations*. Springer.

[7] Bergemann, D. and Välimäki, J. (2008). Bandit Problems. In S. N. Durlauf and L. E. Blume (eds.), *The New Palgrave Dictionary of Economics: Volume 1 – 8*, London: Palgrave Macmillan UK, pp. 336–340.

[8] Bloembergen, D., Tuyls, K., Hennes, D. and Kaisers, M. (2015). Evolutionary Dynamics of Multi-Agent Learning: A Survey. *Journal of Artificial Intelligence Research*, **53**, 659–697.

[9] Börgers, T. and Sarin, R. (1997). Learning Through Reinforcement and Replicator Dynamics. *Journal of economic theory*, **77** (1), 1–14.

[10] Chen, L., Mislove, A. and Wilson, C. (2016). An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, pp. 1339–1349.

[11] Cooper, D. J. and Kühn, K.-U. (2014). Communication, Renegotiation, and the Scope for Collusion. *American Economic Journal: Microeconomics*, **6** (2), 247–278.

[12] Cross, J. G. (1973). A Stochastic Learning Model of Economic Behavior. *The quarterly journal of economics*, **87** (2), 239–266.

[13] Decarolis, F. and Rovigatti, G. (2018). From Mad Men to Maths Men: Concentration and Buyer Power in Online Advertising.

[14] Duffy, J. (2006). Agent-based models and human subject experiments. *Handbook of computational economics*.

[15] Erev, I. and Roth, A. E. (1998). Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *The American economic review*, **88** (4), 848–881.

[16] Ezrachi, A. and Stucke, M. E. (2016). Virtual Competition. *Journal of European Competition Law & Practice*, **7** (9), 585–586.

[17] — and — (2017). Artificial intelligence & collusion: When computers inhibit competition. *University of Illinois law review*, p. 1775.

[18] Fudenberg, D. and Levine, D. K. (2016). Whither Game Theory? Towards a Theory of Learning in Games. *The journal of economic perspectives*, **30** (4), 151–170.

[19] Gata, J. E. (2019). Controlling Algorithmic Collusion: Short Review of the Literature, Undecidability, and Alternative Approaches.

[20] Greenwald, A., Hall, K. and Serrano, R. (2003). Correlated Q-learning. In *ICML*, aaai.org,

vol. 3, pp. 242–249.

[21] Greenwald, A. R., Kephart, J. O. and Tesauro, G. J. (1999). Strategic Pricebot Dynamics. In *Proceedings of the 1st ACM Conference on Electronic Commerce*, EC '99, New York, NY, USA: ACM, pp. 58–67.

[22] Harrington, J. E., Jr (2017). Developing Competition Law for Collusion by Autonomous Price-Setting Agents.

[23] Ho, T. H., Camerer, C. F. and Chong, J.-K. (2007). Self-tuning experience weighted attraction learning in games. *Journal of economic theory*, **133** (1), 177–198.

[24] Hopkins, E. (2002). Two competing models of how people learn in games. *Econometrica*.

[25] Hu, J., Wellman, M. P. and Others (1998). Multiagent reinforcement learning: theoretical framework and an algorithm. In *ICML*, vol. 98, pp. 242–250.

[26] Huck, S., Normann, H.-T. and Oechssler, J. (2004). Two are few and four are many: number effects in experimental oligopolies. *Journal of economic behavior & organization*, **53** (4), 435–446.

[27] Kianercy, A. and Galstyan, A. (2012). Dynamics of Boltzmann Q learning in two-player two-action games. *Physical review. E, Statistical, nonlinear, and soft matter physics*, **85** (4 Pt 1), 041145.

[28] Klein, T. (2018). Assessing Autonomous Algorithmic Collusion: Q-Learning Under Short-Run Price Commitments.

[29] Könönen, V. (2006). Dynamic pricing based on asymmetric multiagent reinforcement learning. *International Journal of Intelligent Systems*, **21** (1), 73–98.

[30] Kühn, K.-U. and Tadelis, S. (2018). The Economics of Algorithmic Pricing: Is collusion really inevitable?

[31] Leufkens, K. and Peeters, R. (2011). Price dynamics and collusion under short-run price commitments. *International Journal of Industrial Organization*, **29** (1), 134–153.

[32] Maskin, E. and Tirole, J. (1988). A Theory of Dynamic Oligopoly, II: Price Competition, Kinked Demand Curves, and Edgeworth Cycles. *Econometrica*, **56** (3), 571–599.

[33] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, **518** (7540), 529–533.

[34] Rodrigues Gomes, E. and Kowalczyk, R. (2009). Dynamic Analysis of Multiagent Q-learning with E-greedy Exploration. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, New York, NY, USA: ACM, pp. 369–376.

[35] Roth, A. E. and Erev, I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and economic behavior*, **8** (1), 164–212.

[36] Salcedo, B. (2015). Pricing Algorithms and Tacit Collusion.

[37] Shoham, Y., Powers, R., Grenager, T. and Others (2007). If multi-agent learning is the answer, what is the question? *Artificial intelligence*, **171** (7), 365–377.

[38] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. and Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, **529** (7587), 484–489.

[39] —, Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M.,

Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K. and Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, **362** (6419), 1140–1144.

[40] —, Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T. and Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, **550** (7676), 354–359.

[41] Singh, N. and Vives, X. (1984). Price and quantity competition in a differentiated duopoly. *The Rand journal of economics*, pp. 546–554.

[42] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

[43] Tampuu, A., Matiisen, T., Kodelja, D., Kuzovkin, I., Korjus, K., Aru, J., Aru, J. and Vicente, R. (2017). Multiagent cooperation and competition with deep reinforcement learning. *PloS one*, **12** (4), e0172395.

[44] Tesauro, G. and Kephart, J. O. (2002). Pricing in Agent Economies Using Multi-Agent Q-Learning. *Autonomous agents and multi-agent systems*, **5** (3), 289–304.

[45] Tyagi, R. K. (1999). On the relationship between product substitutability and tacit collusion. *Managerial and Decision Economics*, **20** (6), 293–298.

[46] Waltman, L. and Kaymak, U. (2008). Q-learning agents in a Cournot oligopoly model. *Journal of economic dynamics & control*, **32** (10), 3275–3293.

[47] Watkins, C. J. C. H. (1989). *Learning from delayed rewards*. Ph.D. thesis, King's College, Cambridge.

[48] — and Dayan, P. (1992). Q-learning. *Machine learning*, **8** (3), 279–292.

[49] Wunder, M., Littman, M. L. and Babes, M. (2010). Classes of multiagent q-learning dynamics with epsilon-greedy exploration. *Proceedings of the 27th International Conference on Machine Learning*.