



Pigouvian algorithmic platform design[☆]

Thomas W.L. Norman

Magdalen College, High Street, Oxford OX1 4AU, United Kingdom



ARTICLE INFO

Article history:

Received 20 June 2022

Revised 30 January 2023

Accepted 15 May 2023

JEL classification:

C73

K21

L40

Keywords:

Algorithms

Reinforcement learning

Collusion

Platform design

replicator dynamics

Pigouvian taxation

ABSTRACT

There are rising concerns that reinforcement algorithms might learn tacit collusion in oligopolistic pricing, and moreover that the resulting 'black box' strategies would be difficult to regulate. Here, I exploit a strong connection between evolutionary game theory and reinforcement learning to show when the latter's rest points are Bayes–Nash equilibria, but also to derive a system of Pigouvian taxes guaranteed to implement an (unknown) socially optimal outcome of an oligopoly pricing game. Finally, I illustrate reinforcement learning of equilibrium play via simulation, which provides evidence of the capacity of reinforcement algorithms to collude in a very simple setting, but the introduction of the optimal tax scheme induces a competitive outcome.

© 2023 The Author(s). Published by Elsevier B.V.
This is an open access article under the CC BY license
(<http://creativecommons.org/licenses/by/4.0/>)

1. Introduction

As the technology of artificial intelligence develops, there is an increasing tendency for firms to employ algorithms to set prices, raising the possibility that such algorithms might collude.¹ This collusion could be deliberate, in which case it is well covered by the existing game-theoretic literature, but novel possibilities arise for collusion as a bi-product of the use of algorithms and specifically the learning process in which they engage. Whilst tacit collusion is a well-understood possibility in the absence of algorithms, their presence could overcome the severe equilibrium selection problem in infinitely repeated games ('supergames') that hitherto may have necessitated explicit communication. 'Reinforcement learning' algorithms respond to a game's history of play by increasing the frequency with which successful strategies are played, and decreasing that of unsuccessful strategies. Such algorithms do not require the players to know the payoffs of the game, nor to construct a model of opponents' play, removing possible impediments to collusion.² Moreover, 'deep' reinforcement algorithms

[☆] This research was partly funded by the economic consultancy Compass Lexecon. The views expressed in this paper are the sole responsibility of the author and cannot be attributed to Compass Lexecon or any other parties.

E-mail address: thomas.norman@magd.ox.ac.uk

¹ 'Revenue management' has a long history in hotel and airline pricing, but more generally, in 2015 more than a third of the vendors of bestselling items on Amazon had already automated their pricing (Chen et al., 2012).

² Salcedo (2015) finds collusive outcomes to be an inevitable consequence of optimized algorithms that can periodically observe and "decode" the others before they are changed, but these assumptions are perhaps too strong for practical relevance.

in particular can produce strategies that are complex and opaque (resembling a ‘black box’), with convergence properties unknown even to the designers, impeding the detection and regulation of any resulting anti-competitive conduct.³

In this paper, I exploit a connection between reinforcement learning and evolutionary game theory to allow the modelling of algorithmic pricing via Taylor (1979) ‘replicator dynamics’. Under the replicator dynamics, the strategies used in a population of players increase in frequency if and only if they outperform the average population strategy—a process with an intuitive ‘reinforcement’ flavour to it, made precise by Börgers and Sarin (1997). I use this connection to show that limiting rest points of the Cross (1973) model of reinforcement learning are Bayes–Nash equilibria in a two-firm pricing supergame. Beyond this though, I use Sandholm (2002, 2005, 2007) evolutionary implementation to derive a system of Pigouvian taxation of pricing algorithms that is guaranteed to lead such Cross-learning firms to an (unknown) socially optimal outcome.⁴ This implementation could most naturally be performed through an appropriately designed platform for algorithmic pricing. I illustrate these theoretical results through simulations, which show evidence both of equilibrium play under the relevant conditions and of a collusive equilibrium selection being made. That this occurs under the simplest strategic representation of collusion (the prisoner’s dilemma) and under the simplest reinforcement algorithm (the Cross learning process) highlights the ease with which such algorithms may collude in oligopoly. Finally, I impose the optimal tax scheme for the game and show that it induces a non-collusive outcome in the simulations—a seemingly essential intervention given the strength of the foregoing collusive equilibrium selection.

Theoretical support for the collusive possibilities of algorithm use is weak at present. Reinforcement learning (rooted in the work of Bush and Mosteller, 1951, 1955) has been one of the most successful machine learning techniques of recent years, with the ‘Q-learning’ algorithm being a simple and popular example that is representative of more complex approaches. Q-learning is an iterative procedure for estimating the value of the actions available in different states of a decision problem, without full knowledge of its payoffs or dynamic structure. It combines myopically sub-optimal ‘exploration’ of the strategy space with optimal ‘exploitation’ of its accumulated learning—the ‘exploitation’ component being closely related to the replicator dynamics (Tuyls et al., 2003b; 2006)—and under certain conditions (e.g. vanishing exploration) it converges to the optimum in single-agent problems (Watkins and Dayan, 1992). However, it lacks general convergence (let alone equilibrium) results in infinitely repeated games, owing to the intrinsically non-stationary problem of strategic interaction.⁵ Nonetheless, Calvano et al. (2020) examine the interaction of Q-learning algorithms experimentally, and find that they consistently learn to collude without communicating with one another, enforced by time-limited punishment of defections that gradually gives way to renewed collusion.⁶ The degree of collusion decreases with the number of firms, but remains substantial if there are three or four.⁷

The Cross model is a simplified reinforcement learning process under which the agent raises the probability of his chosen strategy in proportion to the resulting payoff received, with all other choice probabilities reduced proportionally. In a seminal finding, Börgers and Sarin (1997) establish that a continuous time limit of this model converges to the replicator dynamics.⁸ The Cross model is seemingly special, and in particular does not ‘explore’ currently unused strategies, which distinguishes it from popular reinforcement algorithms such as Q-learning. However, the main results in this paper sidestep this feature by assuming that no strategies are initially unused, and indeed some Q-learning processes (that converge to single-agent optima) have the resulting property of vanishing exploration. Börgers et al. (2004) show that a generalized form of Cross learning is a necessary condition for a learning rule’s performance to improve from one period to the next in a constant environment—a seemingly basic requirement for such a rule. Moreover, this generalized class—under which payoffs are subject to certain affine transformations before Cross learning is applied—also admits a close relationship with the replicator dynamics, which has been extended to other reinforcement learning processes (Bloembergen et al., 2015). Since the convergence and stability of evolutionary dynamics are well-studied, they offer a powerful toolkit for the analysis of complex algorithmic dynamics, shedding light on the AI black box.

Furthermore, the connection with evolutionary game theory offers a possible framework for the design of regulatory mechanisms. Sandholm (2002, 2005, 2007) provides a framework for implementing socially optimal outcomes as stable solutions of a broad class of evolutionary dynamics, using a system of transfer payments reminiscent of Pigouvian taxation to internalize externalities (such as those caused to consumers by collusive prices). Whilst standard Pigouvian taxes equilibrate an efficient outcome that is assumed to be known to the social planner, Sandholm’s evolutionary implementation has the planner ignorant of the preference information that would enable it to calculate the efficient state; instead, his optimal tax schemes equilibrate a *learning process* guaranteed to reach a social optimum (wherever it may be). Equilibrium is hence

³ A reinforcement algorithm is ‘deep’ (e.g. Silver et al., 2016; Mnih et al., 2015; Silver et al., 2018) if it tracks the success of strategies via multiple non-linear function approximators called ‘neural networks’.

⁴ In general, implementation theory characterizes mechanisms whose equilibrium outcomes satisfy some social optimality criterion.

⁵ For a survey of responses to this problem in the multi-agent systems literature, see Hernandez-Leal et al. (2017). An early reference is Sandholm and Crites (1996). Nowé et al. (2012, §14.2.2) provides a review of some of the surrounding issues. There are equilibrium convergence results for reinforcement learning processes in normal-form games and extensive-form games (see, e.g., Shoham and Leyton-Brown, 2008, §7.4), but not for infinitely repeated games.

⁶ See also Waltman and Kaymak (2008), as well as Klein’s (2021) investigation of a sequential Maskin and Tirole (1988) model, and Calvano et al.’s (2021) exploration of the imperfect monitoring case.

⁷ Calvano et al. (2020, §V) explore whether other classic ‘plus factors’ for collusion are at work in the Q-learned strategies; such plus factors offer evidence of collusion over and above the parallel movement of prices, and include the number and symmetry of firms, and product substitutability.

⁸ Other papers exploring this relationship include Posch (1997), Börgers and Sarin (2000), Hopkins (2002), Ianni (2002), Sato and Crutchfield (2003) and Beggs (2005).

obtained as a conclusion rather than an assumption of the approach, and moreover, it is certain to be efficient. His most general results rely on concavity of the social welfare function, which here implies a social preference for heterogeneous over homogeneous pricing rules, and is quite plausible if an algorithm is more likely to collude with itself (in ‘self play’) than with another algorithm.⁹ This in turn is quite plausible since many successful reinforcement algorithms ‘propose, then adapt’ to a pattern of play (e.g. a particular equilibrium), allowing symmetric algorithms to coordinate on the same outcome.¹⁰

There is a small emerging literature on mechanism design for algorithmic firms: Johnson et al. (2020) explore a welfare-improving intervention in e-commerce platforms, showing how market designers can improve competition between Q-learning pricing algorithms by steering demand (through product prominence) as a reward for apparent defection from a collusive agreement. Dong et al. (2017), meanwhile, offer a scheme for using ‘smart contracts’ in cloud computing “to stimulate tension, betrayal and distrust between the clouds, so that rational clouds will not collude and cheat”. But whilst traditional mechanism design is a developed field of economics, its wider application in a complex oligopolistic setting beset by multiple equilibria seems likely to be challenging. Evolutionary implementation offers the advantages of both an explicit behavioral link with reinforcement learning, and a unique, stable outcome.

2. The model

There are two firms, R (Row) and C (Column), each of which chooses a *pricing rule* from a finite set $S^r = \{1, \dots, n^r\}$, $r \in \{R, C\}$.¹¹ A pricing rule $i \in S^r$ maps from its own set Φ_i of *features*—which summarize properties of the market that may be important for choosing a price—into the set \mathbb{R}_+ of possible prices. Features play an important role in machine learning algorithms; for instance, in chess obvious features are the number and type of pieces held by each player, whereas in oligopoly pricing relevant features might include current inventory levels or recent price cuts by competitors.¹² The firms may use mixed strategies, which belong to the simplices $X = \{x \in \mathbb{R}_+^{n^R} : \sum_{i \in S^R} x_i = 1\}$ and $Y = \{y \in \mathbb{R}_+^{n^C} : \sum_{i \in S^C} y_i = 1\}$ of probability mixtures, metrized by the standard Euclidean distance. In order to define derivatives on X and Y , I will also have use for their supersets $\bar{X} = \{x \in \mathbb{R}_+^{n^R} : \sum_{i \in S^R} x_i \in I\}$ and $\bar{Y} = \{y \in \mathbb{R}_+^{n^C} : \sum_{i \in S^C} y_i \in I\}$, where I is an open interval containing 1. A mixed strategy is *fully mixed* if it places strictly positive probability on each possible pricing rule; such a strategy belongs to the interior, X° or Y° , of the relevant simplex.

Each firm chooses a pricing rule repeatedly through iterations $m \in \mathbb{N}$, observing at each iteration only its own strategy and profit (and not the other firm’s pricing rule). Between iterations m and $m + 1$, I suppose that the firms compete infinitely often in some pricing stage game (e.g. prisoner’s dilemma or differentiated Bertrand) via the pricing rules chosen at m , say at times $(m + t)_{t \in \mathbb{Q} \cap (0, 1)}$. This means that pricing rules are adjusted more slowly than prices themselves, and that the full set of folk-theoretic equilibria (including collusion) can be achieved by pricing rules.¹³ However, this timing need not be taken as descriptive of the oligopolistic interaction, but rather could be interpreted as the process of ‘training’ algorithms prior to their real-world use; each iteration is then an adaptation of the pricing rules in light of their success in a single simulation of the game.

Infinitely repeated games are employed in preference to finitely repeated games since they represent the standard and simplest model of collusion in oligopolistic competition.¹⁴ Whilst the idea of an infinitely repeated game occurring between each iteration may seem unrealistic, the standard hazard-rate interpretation of repeated games with discounting will see each supgame terminate in finite time with probability one.¹⁵ Game theoretically, the reason for the use of this structure here is that the Cross (1973) learning process below might be described as a ‘normal-form’ learning process, because it acts on mixed strategies of the repeated game’s normal form as opposed to the behavior strategies of the extensive form; Kuhn’s theorem (Kuhn, 1953; Aumann, 1964) reassures us of the outcome equivalence of these two approaches, although the resulting algorithms may differ in terms of computational efficiency (as I discuss further in the next section).

A firm’s expected profit in iteration m is the sum of an *interaction profit*, which depends on both firms’ pricing rules, and an *idiosyncratic profit*, which varies from firm to firm and only depends on the firm’s own pricing rule. Firm r ’s interaction profit when firm R uses pricing rule i and firm C uses pricing rule j is f_{ij}^r , and captures any external effects that the firms impose on one another through their choice of pricing rule. For instance, if a firm chooses a highly competitive pricing rule, then it imposes a negative externality on its opponent via interaction profits, whereas a collusive pricing rule would impose a positive externality. Meanwhile, each firm r ’s idiosyncratic profit from using each pricing rule is determined by its *type* θ , which belongs to a finite set $\Theta^r \subseteq \mathbb{R}^{n^r}$. For instance, a firm might be a small start-up that values a highly responsive pricing rule, or it might be a large blue-chip company that values a predictable and transparent pricing rule.

⁹ The ‘self play’ paradigm in reinforcement learning (see DiGiovanni and Zell, 2021) involves an algorithm playing the game against itself, either as a method of learning or as a performance criterion.

¹⁰ Indeed, an algorithm achieving Pareto efficiency in self play is offered by Powers and Shoham (2004).

¹¹ The restriction to two players is for the sake of expositional simplicity.

¹² Such features play an implicit part in Jehiel and Samet’s (2007) ‘similarity classes’.

¹³ Barlo et al.’s (2016) bounded memory folk theorem, for instance, would be implementable by pricing rules.

¹⁴ However, whilst departures from stage Nash play are subject to the well-known unravelling argument in finitely repeated games, with the bounded complexity inherent in pricing rules we could approximate all folk-theoretic equilibria of the prisoner’s dilemma in a large finitely repeated game (Neyman, 1985). Thus, the focus on infinite repetitions is inessential.

¹⁵ The hazard-rate interpretation of discounting in supgames is that the discount factor represents the probability of the game ending after each repetition.

The pricing rule game played at each iteration $m \in \mathbb{N}$ is then defined by a triple (π, μ^R, μ^C) , where: $\mu^r \in M^r = \{v \in \mathbb{R}_+^{\Theta^r} : \sum_{\theta \in \Theta^r} v_\theta = 1\}$ is a type distribution, from which firm r 's type is drawn at the start of the game; and $\pi_{\theta,i,j}^r = f_{ij}^r + \theta_i \in [0, 1]$ is a type- θ firm r 's profit from using pricing rule i against pricing rule j . Of course, this game of incomplete information nests complete information as the special case where μ^R and μ^C are degenerate. The bounds placed on profits are important only for their scaling with probabilities under Cross learning (below); without them, we would simply require some mapping of profits into the closed unit interval. Since there are finitely many possible types, the set of pure Bayesian strategies in this game is finite, and the pricing rule game is hence a finite normal-form game.

I assume that the firms update their strategies in each iteration of the pricing rule game according to the Cross (1973) model of reinforcement learning. Specifically, if a type- θ firm R 's strategy in iteration m places probability $x_{\theta,i}(m)$ on pricing rule i , and it then chooses i and receives profit $\pi_{\theta,i,j}^r$, then it updates its strategy in iteration $m+1$ such that

$$\begin{aligned} x_{\theta,i}(m+1) &= \pi_{\theta,i,j}^R + (1 - \pi_{\theta,i,j}^R)x_{\theta,i}(m) \\ x_{\theta,i'}(m+1) &= (1 - \pi_{\theta,i,j}^R)x_{\theta,i'}(m) \quad \text{for all } i' \neq i. \end{aligned}$$

Thus, the probability of the chosen pricing rule is updated to be a weighted average of its old probability and the maximum probability 1, with the weight on the latter being the profit that the pricing rule achieved; this weight is also used to scale down all other choice probabilities appropriately.¹⁶ A type- θ firm C updates its iteration- m strategy $y_\theta(m)$ in a similar manner. Given initial values $(x_\theta(1), y_\theta(1))$, these equations define the Cross learning process $\{x_\theta(m), y_\theta(m)\}_{m \in \mathbb{N}}$. A continuous time limit of this process is obtained from the equations

$$\begin{aligned} \tilde{x}_{\theta,i}(m+1) &= \tau \pi_{\theta,i,j}^R + (1 - \tau \pi_{\theta,i,j}^R)\tilde{x}_{\theta,i}(m) \\ \tilde{x}_{\theta,i'}(m+1) &= (1 - \tau \pi_{\theta,i,j}^R)\tilde{x}_{\theta,i'}(m) \quad \text{for all } i' \neq i, \end{aligned}$$

where $\tau \in (0, 1]$ measures the time between successive iterations of the game, and thus scales the size of strategy adjustment made in response to current payoffs; the new probabilities approach the old as $\tau \rightarrow 0$, in order to yield a smooth process. Similar equations hold for firm C . Given the initial values $(\tilde{x}_\theta(1), \tilde{y}_\theta(1))$, the limit process $\{\tilde{x}_\theta(m), \tilde{y}_\theta(m)\}_{m \in \mathbb{N}}$ for a sequence of pairs (τ, m) such that $\tau \rightarrow 0$ and $m\tau \rightarrow t$ describes the firms' strategies at the point t in continuous time.

Börger and Sarin (1997) establish a close relationship between this limit process and Taylor (1979) two-population replicator dynamics $\{\hat{x}_\theta(t), \hat{y}_\theta(t)\}_{t \in \mathbb{R}_+}$, where \hat{x}_θ and \hat{y}_θ are differentiable functions on X and Y satisfying

$$\begin{aligned} \frac{d\hat{x}_{\theta,i}}{dt} &= \hat{x}_{\theta,i}(t)(e_i \pi_{\theta,i,j}^R \hat{y}_{\theta,i}(t) - \hat{x}_{\theta,i}(t) \pi_{\theta,i,j}^R \hat{y}_{\theta,i}(t)) \\ \frac{d\hat{y}_{\theta,j}}{dt} &= \hat{y}_{\theta,j}(t)(\hat{x}_{\theta,j}(t) \pi_{\theta,j}^C e_j - \hat{x}_{\theta,j}(t) \pi_{\theta,j}^C \hat{y}_{\theta,j}(t)), \end{aligned} \quad (1)$$

for all $t \in \mathbb{R}_+$, $i \in S^R$ and $j \in S^C$, and e_i is the unit vector with a one in the i th row and zeros elsewhere. Given initial values $(\hat{x}_\theta(0), \hat{y}_\theta(0))$, these equations yield the solution $(\hat{x}_\theta, \hat{y}_\theta)$ of the continuous-time replicator equation.

Lemma 1 (Börger and Sarin, 1997). Suppose that for all τ , $(\tilde{x}_\theta(1), \tilde{y}_\theta(1)) = (\hat{x}_\theta(0), \hat{y}_\theta(0))$ with probability 1. Consider some t with $0 \leq t < \infty$ and assume $\tau \rightarrow 0$ and $m\tau \rightarrow t$. Let \tilde{x}_θ and \tilde{y}_θ be the solution of the continuous-time replicator equation for initial values $\hat{x}_\theta(0)$ and $\hat{y}_\theta(0)$. Then $(\tilde{x}_\theta(m), \tilde{y}_\theta(m))$ converges in probability to $(\hat{x}_\theta(t), \hat{y}_\theta(t))$.

Thus, if τ is small and $m\tau$ is close to t , then with high probability $(\tilde{x}_\theta(m), \tilde{y}_\theta(m))$ will take values close to the solution of the continuous-time replicator equation at time t . Convergence in probability captures the idea of a process being probabilistically approximated by a sequence of other processes, with the degree of approximation becoming closer as the sequence progresses. In particular, the probability of any given 'error' in this approximation can be made to lie below any given value by taking a sufficiently high element of the sequence. Of course, between each iteration an infinitely repeated game is still being conducted; thus, rather than compressing the timescale of firm interaction, this convergence should be viewed as identifying the amount of reinforcement learning (or 'training' of the algorithm) required for the replicator dynamics to provide a good approximation.

We may therefore use the replicator dynamics to describe the behavior of two firms updating their strategies according to Cross' reinforcement learning. In particular, we can use the results of Sandholm (2005) to derive a socially optimal tax scheme for pricing rules. Whilst I do not invoke the standard population-game interpretation of the replicator dynamics, I do need some of the associated notation in order to draw on Sandholm's model. Viewing (1) as capturing a population game then, there would be a unit mass of firms for each role, Row and Column, and X and Y would be the sets of distributions of firms over pricing rules. Repeated random pairwise matching of the firms would yield a linear common profit function for the Row population of $F^R : Y \rightarrow \mathbb{R}^R$, where $F_i^R(y)$ is just the expected interaction profit f_{ij}^R to pricing rule i when the Column population's pricing rule distribution is y . A similar function F^C determines the Column population's common profit.

¹⁶ Whilst reinforcement learning in general recognises the dilemma of balancing 'exploitation' of previous learning with 'exploration' of the strategy space, the Cross learning model has rather limited exploration, in the sense that it does not experiment with actions that are currently unused.

The type distributions (μ^R, μ^C) in this context just record the types present in each population. The set of states under type distribution $\mu \equiv (\mu^R, \mu^C)$ is $Z_\mu = \{(z^R, z^C) \in \mathbb{R}_+^{\Theta^R \times S^R \times \Theta^C \times S^C} : \sum_i z_{\theta,i}^r = \mu_\theta^r \text{ for all } r \in \{R, C\}, \theta \in \Theta^r\}$, where $z_{\theta,i}^r$ is the mass of population- r firms of type θ that choose pricing rule i . Letting $x(z) \equiv \sum_{\theta \in \Theta^R} z_\theta^R \in X$ and $y(z) \equiv \sum_{\theta \in \Theta^C} z_\theta^C \in Y$ denote the pricing rule distributions in state $z \equiv (z^R, z^C) \in Z_\mu$, the profit function $\Pi : Z_\mu \rightarrow \mathbb{R}^{\Theta^R \times S^R \times \Theta^C \times S^C}$ is such that the population- R profit from pricing rule i for firms of type θ when the state is z is given by

$$\Pi_{\theta,i}^R(z) = F_i^R(y(z)) + \theta_i,$$

for each $\theta \in \Theta^R$ and $i \in S^R$ (and similarly for $\Pi_{\theta,i}^C(z)$). A state z is a *Bayes–Nash equilibrium* of (π, μ^R, μ^C) if all firms choose a best reply to the play of their opponents:

$$i \in \arg \max_{i' \in S^r} \Pi_{\theta,i'}^r(z) \quad \text{if } z_{\theta,i}^r > 0, \quad r \in \{R, C\}.$$

Again, it is important to note that, because the pricing rules are used to conduct a supergame at each iteration, the Bayes–Nash equilibria of the pricing rule game (π, μ^R, μ^C) are those attainable under the folk theorem (including collusion); if μ^R and μ^C are degenerate, for instance, the classic two-player, complete-information folk theorem applies.

We can now establish a result linking Cross learning with equilibrium.

Proposition 1. *Suppose that each of two firms updates its pricing rule according to a Cross learning process $\{\tilde{x}_\theta(m), \tilde{y}_\theta(m)\}_{m \in \mathbb{N}}$, starting from a fully mixed strategy profile $(\tilde{x}_\theta(1), \tilde{y}_\theta(1)) \in X^\circ \times Y^\circ$. Consider a sequence of pairs (τ, m) such that $\tau \rightarrow 0$ and $m\tau \rightarrow t$ for some $t \in \mathbb{R}_+$. If (x^*, y^*) is a rest point of $\{\tilde{x}_\theta(m), \tilde{y}_\theta(m)\}_{m \in \mathbb{N}}$ for all but finitely many of the sequence of pairs (τ, m) , then (x^*, y^*) is a Bayes–Nash equilibrium of (π, μ^R, μ^C) .*

Proof. Suppose that for all τ , $(\tilde{x}_\theta(1), \tilde{y}_\theta(1)) = (\tilde{x}_\theta(0), \tilde{y}_\theta(0))$ with probability 1. Let \hat{x}_θ and \hat{y}_θ be the solution of the continuous-time replicator equation for initial values $\hat{x}_\theta(0)$ and $\hat{y}_\theta(0)$. If (x^*, y^*) is a rest point of $\{\tilde{x}_\theta(m), \tilde{y}_\theta(m)\}_{m \in \mathbb{N}}$ for all but finitely many of the sequence of pairs (τ, m) , then it is also a rest point of $\{\hat{x}_\theta(t), \hat{y}_\theta(t)\}_{t \in \mathbb{R}_+}$ by Lemma 1. And since interior trajectories of the replicator dynamics satisfy Sandholm's Sandholm (2005) criteria for admissible evolutionary dynamics (Sandholm, 2002), the rest points of such trajectories are Bayes–Nash equilibria by Sandholm (2005) Proposition 2.1. \square

Limiting rest points of the Cross learning process are thus Bayes–Nash equilibria of the pricing rule game.¹⁷ The fully mixed starting point of the learning process is required to rule out the replicator dynamics' boundary rest points, and provides a form of 'exploration' to the otherwise 'exploitative' Cross learning process.

But we would like to go further and implement an efficient equilibrium, which we can do using Sandholm (2005) social planner. This social planner seeks to impose a mechanism ensuring an efficient outcome to the game, but is subject to two constraints in doing so:

1. hidden information—the planner knows the interaction profits $(f_{ij}^r)_{i,j,r}$, but has no information about idiosyncratic profits (as captured by μ);
2. anonymity—the planner's mechanism may only condition on the firms' chosen pricing rule strategies.

These translate the constraints of the Sandholm model into the two-firm Cross learning context. The hidden information constraint is typical of mechanism design problems, and makes the imposition of an efficient outcome more difficult for the planner to achieve, but is likely to be faced in reality: Since interaction profits capture the external effects of pricing rules on other firms, it is reasonable that the planner could observe this through testing of pricing rules, but idiosyncratic profits would seem harder to observe. Anonymity, meanwhile, is imposed by Sandholm in order to confine attention to mechanisms that are easy to administer, in that they do not depend on the firms' identities but only on their choices. Whilst this seems a desirable constraint to impose on public policy, it does allow the planner to observe the firms' strategies in addition to their actions, which may be unrealistic in many cases. However, one context in which this is a natural assumption is where firms are required to submit their strategies to a platform for implementation (as in the Google AdWords auction, for instance).

The notion of efficiency that I employ is concerned, not just with firms (for whom of course collusion would constitute an efficient outcome), but also with consumers. Hence, it is measured both in terms of the profits obtained by the firms in a given state,

$$\bar{\Pi}^r(z) = \sum_{\theta \in \Theta^r} \sum_{i \in S^r} z_{\theta,i}^r \Pi_{\theta,i}^r(z),$$

and in terms of the welfare $\bar{U}^O : X \times Y \rightarrow \mathbb{R}$ of outside parties. These are used to define an *efficient social choice correspondence* incorporating the external effects on consumers:

$$\phi^O(\mu) = \arg \max_{z \in Z_\mu} \bar{\Pi}^R(z) + \bar{\Pi}^C(z) + \bar{U}^O(x(z), y(z)).$$

¹⁷ It follows that, in cases where the replicator dynamics converge, we have equilibrium convergence of the limiting reinforcement learning processes. Of course, there is the important proviso that the replicator dynamics themselves converge, but this is not a problem for Theorem 1 below, as discussed in Remark 3.

For every type distribution μ , $\phi^O(\mu)$ specifies the set of population states that maximize total welfare among those that are feasible under μ , and is nonempty if \bar{U}^O is continuous (on the compact Z_μ) by Berge's maximum theorem. The Row (or Column) profit function in ϕ^O can be split into its total common profits $\bar{F}^R(x, y)$ and its total idiosyncratic profits $\bar{I}(z^R)$:

$$\begin{aligned}\bar{\Pi}^R(z) &= \sum_{i \in S^R} x_i(z) F_i^R(y(z)) + \sum_{\theta \in \Theta^R} \sum_i z_{\theta,i}^R \theta_i \\ &= \bar{F}^R(x(z), y(z)) + \bar{I}(z^R).\end{aligned}$$

A firm- R tax scheme is a map $T^R: X \times Y \rightarrow \mathbb{R}^{n^R}$, with $T_i^R(x, y)$ giving the tax to be paid by firm R when choosing pricing rule i under strategy profile (x, y) . Its imposition shifts the firm- R common profit from F^R to $F^R - T^R$, whilst making no use of any type information (in the sense that T^R does not depend on any firm's θ). Sandholm (2005) explores tax schemes that globally implement a social choice correspondence ϕ , in the sense that for each type distribution μ , the set $\phi(\mu)$ is globally stable under any of a broad set of admissible evolutionary dynamics. A successful tax scheme must serve two roles, ensuring both that socially optimal play is always an equilibrium, and that this equilibrium is always essentially unique and globally stable.

Theorem 1. Suppose: that each of two firms updates its pricing rule according to a Cross learning process $\{\tilde{x}_\theta(m), \tilde{y}_\theta(m)\}_{m \in \mathbb{N}}$, starting from a fully mixed strategy profile $(\tilde{x}_\theta(1), \tilde{y}_\theta(1)) \in X^\circ \times Y^\circ$; that the function $\bar{F}^R + \bar{F}^C + \bar{U}^O$ is concave; and that pricing rules are subject to the tax scheme

$$\begin{aligned}T_i^R(x, y) &= - \left(\sum_{j \in S^C} y_j \frac{\partial F_j^C}{\partial x_i}(x) + \frac{\partial \bar{U}^O}{\partial x_i}(x, y) \right) \\ T_j^C(x, y) &= - \left(\sum_{i \in S^R} x_i \frac{\partial F_i^R}{\partial y_j}(y) + \frac{\partial \bar{U}^O}{\partial y_j}(x, y) \right).\end{aligned}$$

Then, for every $\varepsilon > 0$, there exists a $t \in \mathbb{R}_+$ and a sequence of pairs (τ, m) with $\tau \rightarrow 0$ and $m\tau \rightarrow t$ such that, for every $\delta > 0$, the firms' strategies $(\tilde{x}_\theta(m), \tilde{y}_\theta(m))$ are within ε of the efficient social choice $\phi^O(\mu^R, \mu^C)$ with probability at least $1 - \delta$.

Proof. Suppose that for all τ , $(\tilde{x}_\theta(1), \tilde{y}_\theta(1)) = (\hat{x}_\theta(0), \hat{y}_\theta(0))$ with probability 1. Let \hat{x}_θ and \hat{y}_θ be the solution of the continuous-time replicator equation for initial values $\hat{x}_\theta(0)$ and $\hat{y}_\theta(0)$. Given any $\varepsilon > 0$, there exists a $t \in \mathbb{R}_+$ such that $(\hat{x}_\theta(t), \hat{y}_\theta(t))$ is within $\varepsilon/2$ of $\phi^O(\mu^R, \mu^C)$ by Sandholm (2005) Theorem 5.1. Consider a sequence of pairs (τ, m) such that $\tau \rightarrow 0$ and $m\tau \rightarrow t$. Then, for every $\delta > 0$, $(\tilde{x}_\theta(m), \tilde{y}_\theta(m))$ is within $\varepsilon/2$ of $(\hat{x}_\theta(t), \hat{y}_\theta(t))$ with probability at least $1 - \delta$ for all but finitely many of the sequence of pairs (τ, m) by Lemma 1. \square

This is effectively Sandholm's Sandholm (2005) Theorem 5.1, modified for the two-population setting and combined with Lemma 1. Under the given tax scheme, the tax that a firm pays for choosing pricing rule i is equal to the marginal impact that the firm currently has on the other firm's profit and the welfare of outside parties by choosing this pricing rule. The proof then establishes something akin to convergence in probability of strategies to the efficient social choice, but where the relevant sequence of random variables is constructed anew for each ε -neighborhood of the latter. Thus, true convergence applies only 'in the limit' as $\tau \rightarrow 0$.

Remark 1. Under standard Pigouvian taxation, the planner sets taxes equal to the marginal externalities created at the efficient state, thereby rendering it an equilibrium. The planner is unable to do this here, as it knows neither the type distribution nor hence the efficient state, so it sets taxes on the basis of current externalities. The effect is that the learning trajectories leading to the efficient state (wherever it may be) are implemented; the mechanism is that the taxes transform the pricing rule game into a 'potential game' (Monderer and Shapley, 1996). Intuitively, in such games, it is as if each player were trying to maximize a common payoff function, yielding attractive stability properties under evolutionary dynamics.

Remark 2. Without the concavity condition, the game $(F - T, \mu^R, \mu^C)$ could admit multiple stable equilibria, preventing global convergence of play. Concavity of $\bar{F}^R + \bar{F}^C + \bar{U}^O$ means that average pricing rule distributions are socially preferred to extreme ones on interaction grounds (ignoring idiosyncratic profits)—i.e. that a greater variety of pricing rules is beneficial to welfare. This is quite plausible if pricing rules are drawn from a common set ($S^R = S^C$) and are more likely to collude in self play than against different pricing rules.¹⁸

Remark 3. As discussed in Börgers and Sarin (1997, §5), Lemma 1 applies to arbitrary but finite points in time, and is not true through infinite time; the asymptotic behavior of the Cross learning process may be quite different from the asymptotic behavior of the replicator dynamics if the latter fails to converge. This is the reason for Proposition 1's focus on rest points. Such asymptotic differences cannot arise in Theorem 1, however, since the replicator dynamics converge under Sandholm's optimal tax scheme.

¹⁸ Of course, whilst \bar{F}^r is bilinear, \bar{U}^O is not in general.

	<i>C</i>	<i>D</i>
<i>C</i>	0.4	0.5
	0.4	0.1
<i>D</i>	0.1	0.2
	0.5	0.2

Fig. 1. Prisoner's dilemma.

Remark 4. In addition to Cross learning, there is a more general relationship between reinforcement learning and the replicator dynamics. In particular, the replicator dynamics has been connected to reinforcement learning with an endogenous aspiration level (Börger and Sarin, 2000), to Learning Automata (Tuyls et al., 2002), and to the popular Boltzmann Q-learning model (Tuyls et al., 2003b; 2006). In the case of Q-learning, however, whilst the replicator dynamics describe the selection (or ‘exploitation’) process at work, there is also an additive mutation process capturing the ‘exploration’ aspect of reinforcement learning. This mutation process favors strategies with higher entropy, and hence is inconsistent with Nash convergence of the learning process, even for interior starting points; at best, convergence under a persistently explorative process would be to a perturbed equilibrium concept such as quantal response equilibrium (McKelvey and Palfrey, 1995). This takes Q-learning with non-vanishing exploration outside of the class of learning processes straightforwardly represented by Sandholm (2005) admissible evolutionary dynamics. Some Q-learning models are, however, ‘Greedy in the Limit with Infinite Exploration (GLIE)’, meaning that their exploration recedes to optimal exploitation over time, so that the replicator dynamics describe their limiting behavior.¹⁹

3. Simulations

In this section, I illustrate the Nash rest points result of the previous section through simulations of the Cross learning process on the prisoner's dilemma. These simulations also endorse the concern that reinforcement algorithms might learn to collude in an oligopoly setting, for they do so reliably, especially under the conditions yielding equilibrium play. Whilst evidence of this has previously been offered for Q-learning in a repeated differentiated Bertrand game (Calvano et al., 2020), the fact that it extends to such a simple strategic representation of collusion as the prisoner's dilemma, under such a simple reinforcement algorithm as the Cross learning process, highlights the ease with which collusion may arise. Finally, I show how the tax scheme of Theorem 1 induces a socially optimal outcome.

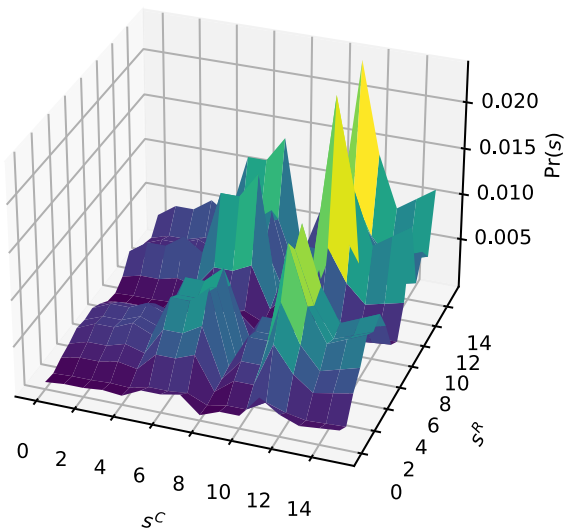
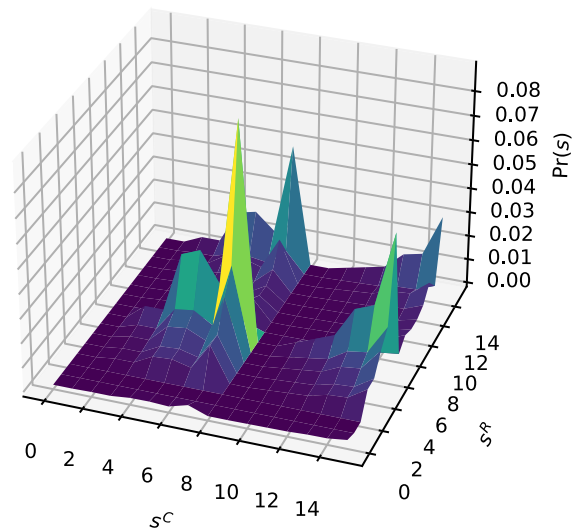
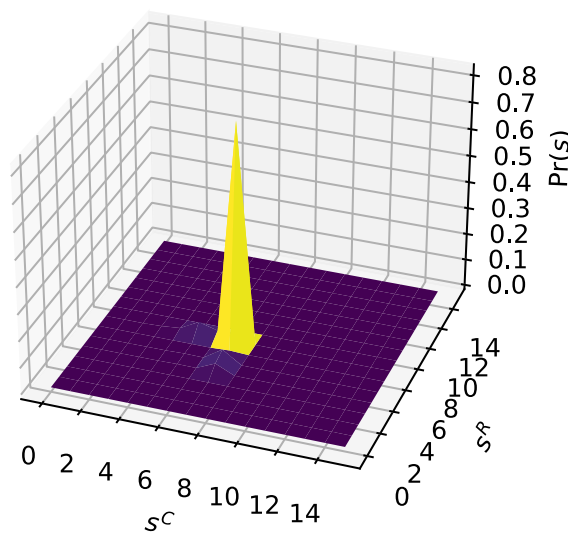
In particular then, I consider two firms of only one possible type ($\Theta^R = \Theta^C = \{0\}$) repeatedly playing the prisoner's dilemma in Fig. 1 (where the payoffs are normalized to lie in the interval (0,1) for the sake of their probabilistic scaling under the Cross learning process).

Since I wish to explore the full collusive possibilities of this game without focusing on the role of the discount factor, I use time-average payoffs, which can be thought of as the limiting case where firms become arbitrarily patient. I take the space of pricing rules to be defined by restricting the firms to a memory of length 1, as in Calvano et al. (2020). There being only four possible action profiles in the prisoner's dilemma, the resulting set of pricing rules conditions on just four features, yielding a common set $S^R = S^C$ of cardinality 16 (listed in full in the Appendix).

The reasons for choosing the prisoner's dilemma over (say) the more complex differentiated Bertrand game studied by Calvano et al. (2020) are twofold. First, the prisoner's dilemma is the most parsimonious representation of the incentive to collude in oligopoly competition, capturing in a reduced form a variety of strategic settings. Second, and relatedly, the Cross learning process as elaborated in the previous section is a normal-form learning process, in that it acts on the entire repeated-game strategy profile at each iteration. By contrast, the extensive-form Q-learning process employed by Calvano et al. (2020) acts on the space of possible action profiles available in the realized subgames of a single repeated game. Whilst the two approaches are equally valid for the theoretical characterization of convergence behavior, the extensive-form approach economizes significantly on the information it records at each iteration, and hence makes drastic gains in computational efficiency (a point made stark in Jehiel and Samet, 2005). This means that implementation of the Cross process would be computationally prohibitive on Calvano et al. (2020) pricing game.

Proposition 1 establishes Nash rest points for the Cross learning process as $\tau \rightarrow 0$, and the effect of reducing τ is apparent from a comparison of the charts in Fig. 2.

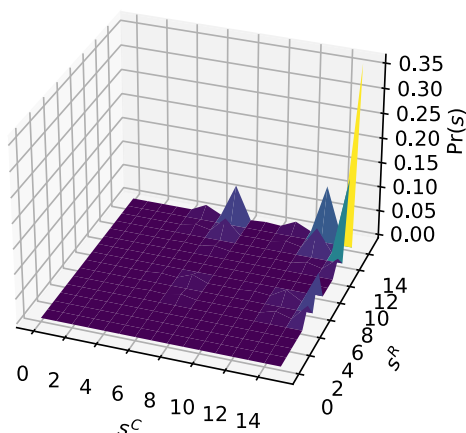
¹⁹ Tuyls et al. (2003a) develop a reinforcement learning algorithm that always attains a stable Nash equilibrium, based on an extension of the replicator dynamics.

Average strategy profile, $\tau = 1$ Average strategy profile, $\tau = 0.1$ Average strategy profile, $\tau = 0.01$ **Fig. 2.** Simulation results for high, medium and low τ .

These plots show the frequency of the pricing rules $s = (s^R, s^C)$ across 100 simulations of 50,000 iterations of the Cross learning process (starting from an equal-weight mixture), for τ values of 1, 0.1 and 0.01. Whereas there is significant weight placed on disequilibrium pricing rules when $\tau = 1$, this diminishes when τ is reduced to 0.1 and especially 0.01. In the latter case, the support of the average strategy profile is almost entirely confined to pricing rule s^7 , which is the ‘grim-trigger’ strategy that plays C if and only if both players have never played D. The $\tau = 1$ case tallies with the notable incidence of disequilibrium play under [Calvano et al. \(2020\)](#) Q-learning process, whereas $\tau = 0.01$ and [Proposition 1](#) point a way towards equilibrium convergent reinforcement learning processes.²⁰

The other obvious feature of these simulations is that they make a collusive equilibrium selection; the ‘grim-trigger’ pricing rule s^7 of course sustains cooperative play when paired against itself, yielding full collusion from a coopera-

²⁰ An alternative approach to guaranteeing equilibrium convergence under reinforcement is to introduce a gradient ascent element into the process, as in [Tuyls et al. \(2003a\)](#).

Average post-tax strategy profile, $\tau = 0.01$ Fig. 3. Post-tax simulation results for low τ .

tive initial history and slightly over a quarter of excess monopoly profit from a random initial history. This agrees with Calvano et al. (2020) evidence of systematic collusion under Q-learning; here the equilibrium play afforded by a vanishing τ focuses almost all weight on a collusive pricing rule, corresponding to the almost perfectly collusive strategies achieved under Q-learning with a fully mixed initialization (Calvano et al., 2020, p. 3292). The fact that collusion arises so readily even under the simple Cross learning process suggests that an intervention such as the previous section's Pigouvian pricing scheme is essential to avoid anti-competitive outcomes. Proposition 1 clearly does not account for this equilibrium selection, to which I plan to return theoretically in future work.

What would the optimal Pigouvian pricing scheme look like here? Let us assume that any excess profit over the socially optimal 0.2 is a direct transfer from consumers to firms, and that d_{ij} is the deadweight loss suffered under the pair (i, j) of pricing rules, with the function $\bar{d}(x, y)$ describing the expected deadweight loss under the strategy profile (x, y) . Then there exists some constant $W^0 > 0$ such that:

$$\bar{U}^0(x, y) = W^0 - (\bar{F}^R(x, y) - 0.2) - (\bar{F}^C(x, y) - 0.2) - \bar{d}(x, y).$$

In particular, let us assume that d_{ij} is half of aggregate excess profits under (i, j) (yielding a bilinear $\bar{F}^R + \bar{F}^C + \bar{U}^0$). The socially optimal tax scheme in Theorem 1 is then

$$T_i^R(x, y) = \sum_{j \in S^C} y_j (\pi_{ij}^R + d_{ij})$$

$$T_j^C(x, y) = \sum_{i \in S^R} x_i (\pi_{ij}^C + d_{ij}),$$

where π^r is the pricing rule game's per-period payoff matrix (given in the Appendix). Each firm is thus taxed the expected value of the sum of its own profit and the deadweight loss resulting from the pricing rules—i.e. the expected external cost of its chosen pricing rule. Re-running the above $\tau = 0.01$ simulations with this tax scheme yields the average strategy profile in Fig. 3, where weight is now focused largely on the 'always defect' pricing rule s^{15} , with a little weight on some other almost competitive outcomes and the chance of coordination on the 'grim-trigger' pricing rule reduced to 0.012.

4. Conclusion

In summary, this paper adds to Calvano et al. (2020) simulation evidence that reinforcement algorithms tend to collude in oligopoly pricing, showing this behavior to emerge in the canonical setting of the repeated prisoner's dilemma and under the highly simplified Cross learning process. It also makes theoretical progress towards accounting for this phenomenon by showing that the rest points of certain Cross learning processes are Bayes–Nash equilibria. Whilst this does not guarantee convergence of these processes, it is the first result tying stability of reinforcement learning to equilibrium in infinitely repeated games. There remains a gap between this theoretical result and a satisfying explanation of the collusive equilibrium selection that seems to emerge empirically. However, I do offer a policy avenue for correcting the resulting anti-competitive behavior; Sandholm (2002, 2005, 2007) evolutionary implementation prescribes taxes on algorithmic pricing rules that induce a competitive outcome without the demanding informational assumptions typically associated with Pigouvian taxation. In the simple prisoner's dilemma simulation conducted in the last section, this requires firms to be taxed the expected

profits and deadweight loss resulting from their chosen pricing rules; more generally, it forces firms to internalize the externalities that their pricing rules impose on other agents at any given moment. Whilst Cross learning is quite special, it shares properties with other reinforcement algorithms to which evolutionary dynamics can be linked. This seems likely to be a useful approach in confronting the policy questions raised by the advancing capabilities of artificial intelligence.

Declaration of Competing Interest

This research was partly funded by the economic consultancy Compass Lexecon. The views expressed in this paper are the sole responsibility of the author and cannot be attributed to Compass Lexecon or any other parties.

Appendix A

The sixteen possible memory-1 pricing rules for the simulations in Section 3 are as follows:

$$\begin{aligned}
 s^0 &= \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} & s^1 &= \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} & s^2 &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} & s^3 &= \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \\
 s^4 &= \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} & s^5 &= \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} & s^6 &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & s^7 &= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \\
 s^8 &= \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} & s^9 &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} & s^{10} &= \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} & s^{11} &= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \\
 s^{12} &= \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} & s^{13} &= \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} & s^{14} &= \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} & s^{15} &= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}
 \end{aligned}$$

where the matrix entries give the probability of action C following each of the four action profiles (C, C), (C, D), (D, C) and (D, D).²¹ The pricing rule game's per-period payoff matrix π^r —obtained by simulating the time-average payoffs until they converged—is then:

0.4	0.396	0.331	0.247	0.397	0.392	0.238	0.193	0.25	0.25	0.107	0.104	0.25	0.25	0.104	0.1
0.402	0.356	0.336	0.282	0.393	0.356	0.183	0.195	0.250	0.235	0.11	0.124	0.25	0.239	0.105	0.129
0.426	0.415	0.362	0.266	0.397	0.392	0.308	0.18	0.33	0.298	0.204	0.232	0.250	0.25	0.107	0.102
0.447	0.393	0.395	0.307	0.373	0.318	0.309	0.218	0.385	0.309	0.308	0.192	0.299	0.229	0.203	0.151
0.401	0.4	0.397	0.28	0.352	0.326	0.393	0.215	0.333	0.333	0.233	0.15	0.3	0.3	0.233	0.15
0.404	0.353	0.4	0.306	0.323	0.304	0.3	0.253	0.334	0.296	0.3	0.199	0.3	0.27	0.266	0.197
0.449	0.469	0.419	0.309	0.396	0.303	0.395	0.203	0.492	0.491	0.31	0.248	0.3	0.267	0.233	0.15
0.475	0.413	0.478	0.311	0.359	0.246	0.362	0.256	0.496	0.428	0.492	0.264	0.35	0.207	0.35	0.198
0.45	0.449	0.377	0.289	0.333	0.333	0.110	0.104	0.3	0.234	0.236	0.105	0.3	0.233	0.198	0.102
0.45	0.383	0.377	0.304	0.333	0.309	0.108	0.125	0.367	0.205	0.308	0.178	0.3	0.204	0.109	0.153
0.497	0.495	0.418	0.252	0.367	0.3	0.287	0.106	0.34	0.287	0.306	0.186	0.3	0.233	0.254	0.104
0.498	0.42	0.396	0.321	0.35	0.206	0.288	0.180	0.491	0.274	0.392	0.231	0.35	0.203	0.258	0.181
0.45	0.449	0.449	0.285	0.3	0.3	0.3	0.151	0.3	0.3	0.3	0.151	0.3	0.3	0.3	0.15
0.45	0.39	0.449	0.323	0.3	0.279	0.267	0.202	0.367	0.207	0.367	0.202	0.3	0.257	0.267	0.198
0.499	0.496	0.494	0.325	0.367	0.267	0.367	0.151	0.39	0.49	0.348	0.259	0.3	0.267	0.3	0.15
0.5	0.422	0.497	0.329	0.35	0.208	0.35	0.204	0.498	0.329	0.493	0.269	0.35	0.203	0.35	0.2

References

- Aumann, R.J., 1964. Mixed and behavior strategies in infinite extensive games. In: Dresher, M., Shapley, L.S., Tucker, A.W. (Eds.), *Annals of Mathematics Studies* 52. Princeton University Press, Princeton, NJ, pp. 627–650.
- Barlo, M., Carmona, G., Sabourian, H., 2016. Bounded memory folk theorem. *J. Econ. Theory* 163, 728–774.
- Beggs, A.W., 2005. On the convergence of reinforcement learning. *J. Econ. Theory* 122, 1–36.
- Bloembergen, D., Tuyls, K., Hennes, D., Kaisers, M., 2015. Evolutionary dynamics of multi-agent learning: a survey. *J. Artif. Intell. Res.* 53, 659–697.
- Börger, T., Morales, A.J., Sarin, R., 2004. Expedient and monotone learning rules. *Econometrica* 72, 383–405.
- Börger, T., Sarin, R., 1997. Learning through reinforcement and replicator dynamics. *J. Econ. Theory* 77, 1–14.
- Börger, T., Sarin, R., 2000. Naive reinforcement learning with endogenous aspirations. *Int. Econ. Rev. (Philadelphia)* 41, 921–9504.
- Bush, R.R., Mosteller, F., 1951. A mathematical model for simple learning. *Psychol. Rev.* 58, 313–323.
- Bush, R.R., Mosteller, F., 1955. *Stochastic Models for Learning*. Wiley, New York.
- Calvano, E., Calzolari, G., Denicolò, V., Pastorello, S., 2020. Artificial intelligence, algorithmic pricing, and collusion. *Am. Econ. Rev.* 110, 3267–3297.
- Calvano, E., Calzolari, G., Denicolò, V., Pastorello, S., 2021. Algorithmic collusion with imperfect monitoring. *Int. J. Ind. Organiz.* 79, 102712.
- Chen, L., Mislove, A., Wilson, C., 2012. An empirical analysis of algorithmic pricing on amazon marketplace. In: *WWW '16: Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Geneva, pp. 1339–1349.
- Cross, J.G., 1973. A stochastic learning model of economic behavior. *Q. J. Econ.* 87, 239–266.

²¹ The initial history is uniformly randomized.

- DiGiovanni, A., Zell, E.C., 2021. Survey of Self-Play in Reinforcement Learning. Working Paper.
- Dong, C., Wang, Y., Aldweesh, A., McCorry, P., van Moorsel, A., 2017. Betrayal, distrust, and rationality: smart counter-collusion contracts for verifiable cloud computing. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer Communications Security, Dallas, TX, 30 October–3 November.
- Hernandez-Leal, P., Kaisers, M., Baarslag, T., de Cote, E.M., 2017. A Survey of Learning in Multiagent Environments: Dealing with Non-Stationarity. Working Paper.
- Hopkins, E., 2002. Two competing models of how people learn in games. *Econometrica* 70, 2141–2166.
- Ianni, A., 2002. Reinforcement Learning and the Power Law of Practice: Some Analytical Results. 203. University of Southampton Discussion Papers in Economics and Econometrics.
- Jehiel, P., Samet, D., 2005. Learning to play games in extensive form by valuation. *J. Econ. Theory* 124, 129–148.
- Jehiel, P., Samet, D., 2007. Valuation equilibrium. *Theor. Econ.* 2, 163–185.
- Johnson, J., Rhodes, A., Wildenbeest, M.R., 2020. Platform Design When Sellers Use Pricing Algorithms. Working Paper, SSRN 3691621.
- Klein, T., 2021. Autonomous algorithmic collusion: Q-learning under sequential competition. *RAND J. Econ.* 52, 538–558.
- Kuhn, H.W., 1953. Extensive games and the problem of information. In: Kuhn, H.W., Tucker, A.W. (Eds.), Contributions to the Theory of Games II, *Annals of Mathematics Study* 28. Princeton University Press, Princeton, NJ, pp. 193–216.
- Maskin, E., Tirole, J., 1988. A theory of dynamic oligopoly. II: price competition, kinked demand curves, and edgeworth cycles. *Econometrica* 56, 571–599.
- McKelvey, R.D., Palfrey, T.R., 1995. Quantal response equilibria for normal form games. *Games Econ. Behav.* 10, 6–38.
- Mnih, V., Silver, K.K.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., Hassabis, D., 2015. Human-level control through deep reinforcement learning. *Nature* 518, 529–533.
- Monderer, D., Shapley, L., 1996. Potential games. *Games Econ. Behav.* 14, 124–143.
- Neyman, A., 1985. Bounded complexity justifies cooperation in the finitely repeated prisoners' dilemma. *Econ. Lett.* 19, 227–229.
- Nowé, A., Vrancx, P., De Hauwere, Y.-M., 2012. Game theory and multi-agent reinforcement learning. In: Wiering, M., van Otterlo, M. (Eds.), Reinforcement Learning: State-of-the-Art. Springer-Verlag, Berlin, pp. 441–467.
- Posch, M., 1997. Cycling in a stochastic learning algorithm for normal form games. *J. Evol. Econ.* 7, 193–207.
- Powers, R., Shoham, Y., 2004. New criteria and a new algorithm for learning in multi-agent systems. *Advances in Neural Information Processing Systems (NIPS)* 17.
- Salcedo, B., 2015. Pricing Algorithms and Tacit Collusion. Working Paper.
- Sandholm, T.W., Crites, R.H., 1996. Multiagent reinforcement learning in the iterated Prisoner's dilemma. *BioSystems* 37, 147–166.
- Sandholm, W.H., 2002. Evolutionary implementation and congestion pricing. *Rev. Econ. Stud.* 69, 667–689.
- Sandholm, W.H., 2005. Negative externalities and evolutionary implementation. *Rev. Econ. Stud.* 72, 885–915.
- Sandholm, W.H., 2007. Pigouvian pricing and stochastic evolutionary implementation. *J. Econ. Theory* 132, 367–382.
- Sato, Y., Crutchfield, J.P., 2003. Coupled replicator equations for the dynamics of learning in multiagent systems. *Phys. Rev. E* 67, 015206.
- Shoham, Y., Leyton-Brown, K., 2008. Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations. Cambridge University Press, Cambridge.
- 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., Hassabis, D., 2016. Mastering the game of go with deep neural networks and tree search. *Nature* 529, 484–489.
- Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., Hassabis, D., 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 1140–1144.
- Taylor, P.D., 1979. Evolutionary stable strategies with two types of players. *J. Appl. Probab.* 16, 76–83.
- Tuyls, K., Heytens, D., Nowe, A., Manderick, B., 2003. Extended replicator dynamics as a key to reinforcement learning in multi-agent systems. In: Proceedings of the 14th European Conference on Machine Learning (ECML). Springer.
- Tuyls, K., Lenaerts, T., Verbeeck, K., Maes, S., Manderick, B., 2002. Towards a relation between learning agents and evolutionary dynamics. In: Proceedings of the Belgian-Dutch Conference on Artificial Intelligence (BNAIC 2002). Cambridge University Press.
- Tuyls, K., T'Hoën, P.J., Vanschoenwinkel, B., 2006. An evolutionary dynamical analysis of multi-agent learning in iterated games. *Auton. Agent Multi Agent Syst.* 12, 115–153.
- Tuyls, K., Verbeeck, K., Lenaerts, T., 2003. A selection-mutation model for q-learning in multi-agent systems. In: Proceedings of the Third International Conference on Autonomous Agents and Multi-agent Systems (AAMAS).
- Waltman, L., Kaymak, U., 2008. Q-learning agents in a cournot duopoly model. *J. Econ. Dyn. Control* 32, 3275–3293.
- Watkins, C.J.C.H., Dayan, P., 1992. Q-learning. *Mach. Learn.* 8, 279–292.