Animal Behaviour 104 (2015) 1–28

Contents lists available at ScienceDirect

Animal Behaviour

journal homepage: www.elsevier.com/locate/anbehav

A model for the evolution of reinforcement learning in fluctuating games

Slimane Dridi*, Laurent Lehmann

Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 17 June 2014
Initial acceptance 18 June 2014
Final acceptance 16 January 2015
Available online xxx
MS. number: 14-00499R

Keywords:
evolution of cognition
evolutionarily stable learning rules
exploration–exploitation trade-off
repeated games
social interactions
trial-and-error learning

Many species are able to learn to associate behaviours with rewards as this gives fitness advantages in changing environments. Social interactions between population members may, however, require more cognitive abilities than simple trial-and-error learning, in particular the capacity to make accurate hypotheses about the material payoff consequences of alternative action combinations. It is unclear in this context whether natural selection necessarily favours individuals to use information about payoffs associated with nontried actions (hypothetical payoffs), as opposed to simple reinforcement of realized payoff. Here, we develop an evolutionary model in which individuals are genetically determined to use either trial-and-error learning or learning based on hypothetical reinforcements, and ask what is the evolutionarily stable learning rule under pairwise symmetric two-action stochastic repeated games played over the individual's lifetime. We analyse through stochastic approximation theory and simulations the learning dynamics on the behavioural timescale, and derive conditions where trial-and-error learning outcompetes hypothetical reinforcement learning on the evolutionary timescale. This occurs in particular under repeated cooperative interactions with the same partner. By contrast, we find that hypothetical reinforcement learners tend to be favoured under random interactions, but stable polymorphisms can also obtain where trial-and-error learners are maintained at a low frequency. We conclude that specific game structures can select for trial-and-error learning even in the absence of costs of cognition, which illustrates that cost-free increased cognition can be counterselected under social interactions.

Many species have a learning ability because this allows an individual to adapt, within its lifetime, to the currently fitness-relevant features of its environment (e.g. by tracking the location of food patches; Charnov, 1976; McNamara & Houston, 1985; Shettleworth, Krebs, Stephens, & Gibbon, 1988). Hence, learning is likely to provide a selective advantage (Dunlap & Stephens, 2009; Johnston, 1982; Mery & Kawecki, 2002; Stephens, 1991; Wakano, Aoki, & Feldman, 2004). One of the simplest ways of learning an action is through trial and error (Bush & Mostelller, 1951; Thorndike, 1911). This consists of trying different actions, experiencing the rewards associated with each action, and repeating more often the actions yielding higher rewards (or, equivalently, avoiding actions that yield negative payoffs, or punishments). For example, rats in the Skinner box learn that pressing a lever is associated with obtaining food, and various instances of reinforcement learning in other mammals, birds, fish and insects have been demonstrated (Dugatkin, 2010; Shettleworth, 2009).

Although trial and error is the main paradigm for describing the learning of actions in animals (Dickinson, 1980; Dugatkin, 2010; Shettleworth, 2009), it cannot solve all decision problems. With this behavioural rule, an individual has to physically try (or experience) an action to get the knowledge of the reward (or payoff) associated with it. In other words, information gathering and action choice cannot be dissociated. Inherent to this type of learning is thus the problem of balancing exploration and exploitation (Achbany, Fouss, Yen, Pirotte, & Saerens, 2006; Arnold, 1978; Krebs, Davies, & West, 1993; McNamara & Houston, 1985; Shettleworth et al., 1988; Sutton & Barto, 1998). The individual needs to try various actions in order to identify the good ones, but must also exploit at some point the information gathered during exploration. The balancing problem (or trade-off) comes in because an individual that does not explore enough risks misses highly rewarding actions. On the other hand, an individual that explores too much and disregards small rewards (always searching for the best options) risks not getting any payoff at all.

* Correspondence: S. Dridi, Department of Ecology and Evolution, Université de Lausanne, Biophore, Lausanne, CH-1015, Switzerland.
E-mail address: slimane.dridi@unil.ch (S. Dridi).

http://dx.doi.org/10.1016/j.anbehav.2015.01.037
0003-3472/

Faced with the exploration–exploitation dilemma, one is tempted to ask: in the course of learning, are there other ways than trial and error to get information about the payoff of an action? One can distinguish at least two non-mutually exclusive ways of obtaining information about the material consequences of actions without explicitly expressing them. First, an individual can use social information: it may observe conspecifics' actions and their consequences, and if an action tried by conspecifics is seen to be followed by positive consequences, the observer will subsequently have a greater probability of choosing that action (Kendal, Giraldeau, & Laland, 2009; Laland, 2004; Schlag, 1998). Second, an individual can use environmental cues to deduce information about the value of different actions. This may be achieved via belief-based learning, i.e. by representing in one's mind the outcome of alternative actions, which has been extensively studied as a model of human cognition (Camerer, 2003; Chmura, Goerg, & Selten, 2012; Feltovich, 2000). Further, it has been argued that chimpanzees, *Pan troglodytes*, and various large-brained bird species are capable of forming beliefs to solve cognitively challenging tasks (Emery & Clayton, 2004, 2009; Premack & Woodruff, 1978; Schloegl et al., 2009; Taylor, Miller, & Gray, 2012).

Two lines of evidence suggest that belief-based learning could give a selective advantage over trial-and-error learning and that this is relevant to animal learning. First, in the field of animal behaviour, it is often argued that natural selection should favour individuals that reason about their environment in a Bayesian fashion, because Bayesian learning (which is equivalent to belief-based learning, Fudenberg & Levine, 1998) leads to individuals having a correct representation (or belief) of the distribution of the states of the world (McNamara, Green, & Olsson, 2006; Trimmer et al., 2011). This has been extensively studied empirically in the context of individual decision problems, for example when an animal tries to learn about the quality of food patches (van Gils, Schenk, Bos, & Piersma, 2003; Lima, 1984; Luttbeg & Warner, 1999; for a review, see Valone, 2006). The second line of evidence suggesting that belief-based learning may perform better than trial-and-error comes from the theoretical literature on learning in games. Belief-based learning leads to the optimal solution (Nash equilibrium) in several types of social interactions (Hofbauer & Sandholm, 2002), while trial-and-error learning (studied under different specific forms) can lead to nonoptimal outcomes in the same social interactions (Izquierdo, Izquierdo, Gotts, & Polhill, 2007; Macy & Flache, 2002; Stephens & Clements, 1998). Since empirical evidence suggests that many social behaviours, such as cooperation, mate choice or conflict through the winner and loser effects, may involve learning (Dugatkin, 2010; Dugatkin & Reeve, 2000), it is relevant to understanding the conditions under which belief-based learning for social interactions can be favoured by natural selection.

While the evolution of both learning and social interactions has been extensively studied on its own (e.g. Maynard Smith, 1982; Boyd & Richerson, 1988; Rogers, 1988; Feldman, Aoki, & Kumm, 1996; Hofbauer & Sigmund, 1998; McElreath & Boyd, 2007; Borenstein, Feldman, & Aoki, 2008; Rendell et al., 2010; Kempe & Mesoudi, 2014) surprisingly few studies have examined the evolution of learning for social interaction dilemmas. For instance, many studies on the evolution of social learning have focused on individual decision problems. This is well exemplified by the social-learning tournament (Rendell et al., 2010), in which the tasks individuals need to learn to perform are individual decision problems, and not social interactions (so that individuals were not playing frequency-dependent games). Further, the studies that did investigate learning in games generally assumed that individuals face only a producer–scrounger game (Arbilly, Motro, Feldman, & Lotem, 2010; Dubois, Morand-Ferron, & Giraldeau, 2010; Hamblin & Giraldeau, 2009; Katsnelson, Motro, Feldman, & Lotem, 2011). For instance, Hamblin and Giraldeau (2009) showed that the relative-payoff sum (RPS), a simple variant of trial-and-error learning, can be the evolutionarily stable learning rule under the conditions of a producer–scrounger game. Arbilly et al. (2010, 2011) demonstrated that a simple learning rule can coexist with a more complex learning rule in a producer–scrounger environment. However, results from game theory suggest that the game faced by population members should change for learning to be really useful (Heller, 2004). This may explain why evolutionary ecologists have found it difficult for learning to evolve initially in the producer–scrounger game (Dubois et al., 2010; Katsnelson et al., 2011), and investigation of the evolution of learning rules when the game itself is changing appears to be lacking.

Previous results have also been divergent on whether trial-and-error learning or a more sophisticated learning rule should be favoured by selection. Interestingly, the models of both Hamblin and Giraldeau (2009) and Arbilly et al. (2010, 2011) suggest that simple learning rules can coexist with more complex learning rules. By contrast, Josephson (2008) modelled the competition between a continuum of rules from the linear operator to rules using hypothetical payoffs, and confirmed results from game theory that rules of the belief-based type, which put higher weight on hypothetical payoffs, are evolutionarily stable most of the time. It thus remains unclear under what ecological conditions one should expect to observe simple or complex learning, and more work is needed to understand the selection pressures on learning mechanisms in situations in which individuals can experience different games during their lifetime.

In this paper, we aim to relax previous assumptions and ask whether trial-and-error learning is sufficient in social interactions, or whether a more sophisticated belief-based learning rule will necessarily be selected for. To address this question, we studied the competition between two forms of reinforcement learning rules. The first is standard trial-and-error reinforcement learning (Amano, Ushiyama, Moriguchi, Fujita, & Higuchi, 2006; Bernstein, Kacelnik, & Krebs, 1988; Bush & Mostellller, 1951; Erev & Roth, 1998; Hamblin & Giraldeau, 2009; McNamara & Houston, 1987; Rescorla & Wagner, 1972; Stephens & Clements, 1998), while the second rule we call hypothetical reinforcement learning, a terminology borrowed from Camerer and Ho (1999) where individuals can use 'hypothetical reinforcements'. Here, individuals are assumed to have the ability to infer foregone payoffs given the actions of partners and the state of the environment (either via social observation of other interactions or active reasoning/mental simulation), and reinforce actions according to these hypothetical payoffs.

To assess whether learning based on hypothetical reinforcements provides a selective advantage over trial-and-error learning, we studied the evolutionary stability of trial-and-error and hypothetical reinforcement learning in the simplest possible social situation where the environment can change, i.e. in a situation of pairwise social interactions with only two actions. Our approach is very similar to that of Josephson (2008), because we use the framework of Camerer and Ho (1999) to capture learning rules that rely either on trial and error or on hypothetical reinforcements. In such a setting, genuine environmental fluctuations (where learning is necessary) correspond to the fact that the games faced by individuals change with time; in particular, the evolutionarily stable strategies (ESS) of these various games have to be different, and we studied exhaustively the cases where the environment switches between the Prisoner's Dilemma, the Hawk-Dove (a form of producer–scrounger game) and a Coordination game. These three games have been previously studied on their own to capture, respectively, cooperation (e.g. costly production of

a public good), conflict (e.g. fighting for territories) and coordination problems (e.g. collective hunting) in nature (Dugatkin & Reeve, 2000) and it is likely that animals of the same species may confront all of these games within their lifetime. We also analysed two ways of forming pairs of opponents from the population. In the first, each individual is randomly paired with only one partner with which it interacts for the entire interaction period. In the second matching scheme, there is random matching at each time step so the individuals learn to play against the whole population. Because the underlying model is stochastic, we used, when possible, a deterministic approximation to analyse the equations and then compared the analytical results to individual-based simulations.

The paper is organized as follows. In the next section ('Model'), we present the biological assumptions behind our model, we formalize trial-and-error and hypothetical reinforcement learning, and we describe how natural selection on these learning rules can be approximated analytically. Then, we present analytical and simulation results for the evolution of these learning rules under our two matching schemes, different types of games and different learning rates ('Results: one-shot matching' and 'Results: repeated matching'). Finally, we summarize and discuss the results ('Summary and discussion').

MODEL

Setting the Stage

Population

We consider a haploid population of constant size N. The main life cycle stages are the following. (1) Each individual interacts socially with others for T time periods. (2) Each individual reproduces according to its gains and losses incurred during social interactions. (3) All individuals of the parental generation die and N individuals from the offspring generation are randomly sampled to form the new adult generation.

Repeated game affected by environmental fluctuations

During stage (1), individuals play a game at each time period $t = 0,1,2,\ldots,T$, where the game that is played is determined by some environmental state ω, which is an element of the set Ω of environmental states. The environment may include any factor that alters the payoffs associated with actions taken by individuals. We assume that the environmental process follows an ergodic Markov chain (Karlin & Taylor, 1975), and denote by $\mu(\omega)$ the stationary probability that state ω obtains.

For example, one can consider an environment in which individuals play one of two games, e.g. a Prisoner's Dilemma and a Hawk–Dove game; the set of environmental states is then $\Omega =$ {Prisoner's Dilemma, Hawk–Dove}. In this example, one could set $\mu(\text{Prisoner's Dilemma}) = \mu(\text{Hawk–Dove}) = 1/2$, meaning that it is as if we toss a fair coin at each time t to determine the game to be played. If the current environmental state is $\omega =$ Prisoner's Dilemma, then all individuals in the population will have to choose between cooperating and defecting.

More generally, we consider that all the games in Ω consist of the same number of actions, say m (in our previous example we had $m = 2$). Hence, in every period of time, each organism in the population chooses its action from a fixed finite set of actions $A = \{1,\ldots,m\}$, and we denote by ω_t the game played at time t, which is a random variable. The action taken by individual i at time t is also a random variable denoted by $a_{i,t}$ (we allow individuals to use probabilistic action choice) and the action profile in the population is $\mathbf{a}_t = (a_{1,t},\ldots,a_{N,t})$ (this is the collection of the actions of all individuals in the population at time t). The payoff to individual i at time t when it takes action $a_{i,t}$ and the game is ω_t is denoted $\pi_i(a_{i,t},\mathbf{a}_{-i,t},\omega_t)$, where $\mathbf{a}_{-i,t}$ is the action profile of the remaining individuals in the population (all individuals excluding i).

With this, the mean payoff obtained by individual i during the whole sequence of interactions is

$$\Pi_i = \frac{1}{T}\sum_{t=1}^{T}\pi_i\left(a_{i,t},\mathbf{a}_{-i,t},\omega_t\right), \tag{1}$$

which is taken as its number of offspring produced during stage 2 of the life cycle (i.e. fecundity).

Learning Actions

To evaluate the fecundity, Π_i, of individual i, we need to predict the sequence of actions taken by individuals in the population. We assume that actions are learned and we now present a model of learning that takes both trial-and-error and hypothetical reinforcement learning into account.

Action choice

Our way of modelling learning is shared by many previous studies and relies upon two components: (1) dynamic preferences for action and (2) a rule for choosing an action given preferences (Arbilly et al., 2010, 2011; Camerer & Ho, 1999; Hamblin & Giraldeau, 2009; Harley, 1981; Ho, Camerer, & Chong, 2007; Leslie & Collins, 2005). Specifically, we let individuals have preferences or motivations for actions, which they update through repeated playing of the game. For each action a in its behavioural repertoire $\mathscr{A}$, individual i has an associated motivation $M_{i,t}(a)$ that represents how much action a is valued by i. Thus, the collection of motivations can be thought of as the state of the organism (Enquist & Ghirlanda, 2005; Niv, 2009) and we assume that action a is chosen at time t by individual i with probability

$$p_{i,t}(a) = \frac{\exp\left[\lambda M_{i,t}(a)\right]}{\sum_{k\in\mathscr{A}}\exp\left[\lambda M_{i,t}(k)\right]}. \tag{2}$$

This is a standard choice rule (Anderson, de Palma, & Thisse, 1992; Arbilly et al., 2010, 2011; Camerer & Ho, 1999; Fudenberg & Levine, 1998; Ho et al., 2007; McKelvey & Palfrey, 1995), where the action that has the highest motivation is chosen with the greatest probability. The parameter $\lambda \in [0,\infty)$ represents the sensitivity of an animal to its motivations. If λ is near 0, the animal is not very reactive to its motivations and has a tendency to explore (because $p_{i,t}(a)$ is close to $1/m$). If λ is high, we have a 'greedy' animal, almost certainly taking the action that has the highest motivation (if action a^* has the highest motivation, $p_{i,t}(a^*)$ is close to 1, see Dridi and Lehmann (2014) and references therein for more justifications underlying the use of equation (2)).

Motivations

The motivations of an individual are updated after each interaction stage. To achieve this updating, we use a special case of the model of Camerer and Ho (1999) and its application to stochastically varying environments (Dridi & Lehmann, 2014). Individual i starts off with some initial preferences over actions at time $t = 1$ given by the initial motivations $M_{i,1}(a)$ for all actions a. We assume that the motivation $M_{i,t+1}(a)$ for action a of individual i at time $t+1$ (for $t \geq 1$) is given by

$$M_{i,t+1}(a) = \frac{t}{t+1}\phi_{i,t}M_{i,t}(a) + \frac{1}{t+1}\left\{g_i + (1-g_i)1\left(a,a_{i,t}\right)\right\}\pi_i(a,\mathbf{a}_{-i,t},\omega_t). \tag{3}$$

Equation (3) can be seen as a weighted average of the previous motivation $M_{i,t}(a)$ and of the new payoff $\pi_i(a, \mathbf{a}_{-i,t}, \omega_t)$. In the first term, the motivation $M_{i,t}(a)$ is weighted by $\phi_{i,t} \geq 0$, a memory parameter, or learning rate, that indicates the relative importance of the last motivation as opposed to the current payoff (note that $\phi_{i,t}$ can change as a function of time and also has an initial value $\phi_{i,1}$, possibly genetically determined). This first term is also weighted by $t/(t+1)$, which entails that the previous motivation is weighted according to the number of interactions that have occurred up to time t.

The second term can be termed the increment, or reinforcement to the motivation. It has weight $1/(t+1)$ and depends on the payoff $\pi_i(a, \mathbf{a}_{-i,t}, \omega_t)$ of action a when all other individuals in the population play $\mathbf{a}_{-i,t}$ and the game is in state ω_t at time t. The expression $1(a, a_{i,t})$ is an indicator function, which is

$$1(a, a_{i,t}) = \begin{cases} 1, & \text{if } a_{i,t} = a, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

i.e. if individual i takes action a at time t, then $1(a, a_{i,t}) = 1$. We see that if $1(a, a_{i,t}) = 1$, the numerator of the second term of equation (3) reduces to $\pi_i(a, \mathbf{a}_{-i,t}, \omega_t)$. If, on the other hand, individual i does not play a at time t, then the numerator of the second term reduces to $g_i \pi_i(a, \mathbf{a}_{-i,t}, \omega_t)$.

The parameter g_i weights the ability of an individual to infer the payoffs of unchosen actions (nonrealized or foregone payoffs) and reinforce motivations accordingly. If $g_i = 0$, the individual is not able to infer nonrealized payoffs and thus only reinforces actions according to realized payoff. In this case we obtain a rule in the form of the linear operator model of Bush and Mostelller (1951) (see Appendix 1, equation (A1.3)). We will call this trial-and-error reinforcement learning (TR). By contrast, if $g_i = 1$, the payoffs associated with unchosen actions are always perfectly inferred. This is called hypothetical reinforcement learning (HR), where individuals have the capacity to access information about nonrealized payoffs. When a hypothetical reinforcement learner plays action a at time t, it reinforces not only action a but also all other actions according to the payoffs they would have yielded (see Appendix 1). In other words, the learner can observe a posteriori the payoff it would have obtained had it chosen another action, given its opponent's current action. This information about foregone payoffs may come from reasoning about the environment, or from observing interactions involving other individuals. A well-studied special case of HR obtained when $\phi_i = 1$ and $g_i = 1$ is belief-based learning, where the motivations represent expected payoffs given the history of play by individual i's opponents (Brown, 1951; Camerer, 2003; Chmura et al., 2012; Fudenberg & Levine, 1998; Feltovich, 2000; Hofbauer & Sandholm, 2002; Hopkins, 2002; see also Appendix 1).

Note that the motivations of all actions are updated by individual i after each time t. Consequently, a hypothetical reinforcement learner [with $g_i = 1$], which observes foregone payoffs, computes as many payoffs as there are available actions (m), at each time t. On the other hand, a trial-and-error learner [with $g_i = 0$] has a much lighter computational task because it computes only the payoff of the action it actually took. Thus, we postulate in the analysis below that the cost of additional computations will affect the fitness of HR by an amount k.

Evolutionary Analysis

Parameter space

Equations (2) and (3) define the learning dynamics of individual i for action a. Our aim was to investigate the coevolution of trial-and-error and hypothetical reinforcement learning under these dynamics. To that aim, we consider that the parameter g_i is the genotype of individual i, and that individuals in the population can use only two learning rules, TR with $g_i = 0$ or HR with $g_i = 1$. We investigated the following cases.

(1) We consider two special values of $\phi_{i,t}$, the learning rate, that differ in terms of the behavioural equilibrium they produce and the extent to which they relate to existing learning mechanisms. Standard belief-based learning (Camerer & Ho, 1999) is obtained in our model when $\phi_i = 1$, which entails that a constant amount of noise is included in its dynamics. We could not analyse mathematically the learning dynamics under this condition, so in order to obtain analytical insight, we studied an 'unperturbed' version of belief-based learning, which is obtained when $\phi_{i,t} = (1/t) + 1$. This particular value of $\phi_{i,t}$ does not result in a loss of generality: it is only a mathematical device used to make the analysis tractable (see Appendix 2 for details), and one can show that the unperturbed version of belief-based learning (when $\phi_{i,t} = (1/t) + 1$) can be recovered from the original version with noise (when $\phi_i = 1$), by making λ very large. One of our goals in this paper was to see whether the study of the unperturbed dynamics can help us understand the original perturbed dynamics. The two values of the learning rate are studied in the following order. First, we study the case $\phi_{i,t} = (1/t) + 1$, where analytical results can be derived about the learning behaviour (see next section). For this special value of the learning rate, we obtain a special version of TR, which we call pure trial-and-error reinforcement (PTR), and a special version of HR, which we call pure hypothetical reinforcement (PHR). These names follow from the fact that, at a behavioural equilibrium, individuals will express essentially only pure actions. Second, we use a constant value $\phi_i = 1$ for the learning rate, which entails that individuals are likely to be more exploratory at a behavioural equilibrium and are likely to express mixed actions. For this value of the learning rate, we obtain a special version of TR, which we call exploratory trial-and-error reinforcement (ETR), and a special version of HR, which we call exploratory hypothetical reinforcement learning (EHR), where the latter strategy corresponds to belief-based learning in the game theory literature, a learning procedure that relies on Bayesian updating of beliefs (Fudenberg & Levine, 1998). The ETR rule is related to the linear operator model, a simple learning algorithm that has a long tradition in the study of animal behaviour and has been used to describe the dynamics of instrumental learning in various species (Appendix 1, equation (A1.3); Amano et al., 2006; Bernstein et al., 1988; Bush & Mostelller, 1951; Hamblin & Giraldeau, 2009; McNamara & Houston, 1987; Rescorla & Wagner, 1972; Stephens & Clements, 1998).

In Appendix 1, we provide more details on the dynamics of the motivations (equation (3)) of these learning rules, and in our analysis we always consider competition between TR and HR under the same learning rate.

(2) We assume pairwise interactions and we analyse two matching rules. First, in what we call one-shot matching, individuals are randomly paired at the beginning of the game ($t = 1$) and each pair interacts for the whole duration of the game (until time T and reproduction). Second, we consider repeated matching, where individuals are rematched during each stage of the game, i.e. individuals meet different partners at each time t.

(3) All games consist of two actions ($m = 2$), that is, individuals play 2×2 symmetric games. This means that, at each time t, individuals play one of three types of games: a Prisoner's Dilemma (PD), a Hawk–Dove game (HD) or a (pure) Coordination game (CG); i.e. the set of games is $\Omega = \{\text{PD, HD, CG}\}$, and the stationary distribution of games thus satisfies $\mu(\text{PD}) + \mu(\text{CG}) + \mu(\text{HD}) = 1$. These three games are instances from the three possible categories of 2×2 games (the PD is a game with a dominant action, the HD is a type of 'anticoordination' game, and the CG is a special coordination

game, Weibull, 1997, Chapter 1). The four possible payoffs for game ω are real numbers and are written $\mathscr{R}_\omega$, $\mathscr{S}_\omega$, $\mathscr{T}_\omega$ and $\mathscr{P}_\omega$, where $\omega \in \{\text{PD, CG, HD}\}$. For instance, when the game is ω, player 1 plays action 2, and player 2 plays action 1, then player 1 obtains $\mathscr{T}_\omega$ and player 2 obtains $\mathscr{S}_\omega$. Table 1 describes the payoffs to each player for every possible combination of actions.

Stochastic approximation

Although equations (2) and (3) describe a bona fide learning process, this is a nonhomogeneous multidimensional Markov process that is very difficult to analyse (see Fig. 1). It is thus necessary to approximate it in order to obtain analytical results, which is useful to form an intuition about behavioural dynamics. We were able to perform such an analysis only under the one-shot matching scheme and when the learning rate $\phi_{i,t}$ takes the dynamic value $\phi_{i,t} = (1/t) + 1$.

Using the above assumptions (1–3) with $\phi_{i,t} = (1/t) + 1$ and stochastic approximation theory (Appendix 2; Benaim, 1999; Benveniste, Metivier, & Priouret, 1991), we can write a system of differential equations that describes the learning dynamics of a given pair of individuals (i,j) in the population for the one-shot matching model as

$$\begin{aligned}\dot{p}_i =& p_i(1-p_i)\lambda\Big[\Big\{p_j\mathscr{R} + \Big(1-p_j\Big)\mathscr{S}\Big\}\{p_i + g_i(1-p_i)\} \\ &- \Big\{p_j\mathscr{T} + \Big(1-p_j\Big)\mathscr{P}\Big\}\{g_i p_i + (1-p_i)\}\Big],\end{aligned} \tag{5}$$

$$\begin{aligned}\dot{p}_j =& p_j\Big(1-p_j\Big)\lambda\Big[\{p_i\mathscr{R} + (1-p_i)\mathscr{S}\}\Big\{p_j + g_j\Big(1-p_j\Big)\Big\} \\ &- \{p_i\mathscr{T} + (1-p_i)\mathscr{P}\}\Big\{g_j p_j + \Big(1-p_j\Big)\Big\}\Big],\end{aligned} \tag{6}$$

where a dot accent is used to represent a time derivative, i.e. $\dot{p}_i = \mathrm{d}p_i/\mathrm{d}t$, p_i is the probability that individual i plays action 1 and p_j is the probability that individual j (the opponent of individual i) plays action 1. Since individuals cannot detect the state ω of the game, the parameters $\mathscr{R}$, $\mathscr{S}$, $\mathscr{T}$ and $\mathscr{P}$ are actually the payoffs of the average game faced by the individuals in the fluctuating environment (Table 4), that is, the average over the distribution $\mu(\omega)$ of games. For example, when a focal individual chooses action 2 and its opponent chooses action 1, the former obtains the average payoff $\mathscr{T} = \mu(\text{PD})\mathscr{T}_{\text{PD}} + \mu(\text{HD})\mathscr{T}_{\text{HD}} + \mu(\text{CG})\mathscr{T}_{\text{CG}}$, and the three other payoffs $\mathscr{R}$, $\mathscr{S}$ and $\mathscr{P}$, are similarly computed. Equations (5) and (6) show that, asymptotically, the probability of playing actions is driven by the average payoff to actions, which itself can be thought to determine a game (Appendix 2, equation (A2.3)), which we call the average game (see Dridi & Lehmann, 2014; for more details).

To evaluate fitness, we assume that the learning dynamic (equations (5) and (6)) has reached an equilibrium during an individual's lifetime, and that the game is played a long enough time after the equilibrium action choice has been reached. With this, the payoffs obtained before the equilibrium is reached can be ignored. This is equivalent to saying that we let the time horizon, T, of the game become very large (ideally infinite) so that the equilibrium of equations (5) and (6) determines the fecundity of individuals i and j. Then, the fecundity of individual i (equation (1)) is taken as the average (or expected) payoff obtained at the equilibrium of learning (when $T \to \infty$). That is,

$$\Pi_{ij} = \widehat{p}_i\Big(\widehat{p}_j\mathscr{R} + \Big(1-\widehat{p}_j\Big)\mathscr{S}\Big) + (1-\widehat{p}_i)\Big(\widehat{p}_j\mathscr{T} + \Big(1-\widehat{p}_j\Big)\mathscr{P}\Big), \tag{7}$$

where we use the subscript j in Π_{ij} to emphasize that the fecundity of individual i depends on its single opponent (j), $\widehat{p}_i$ is the equilibrium probability that individual i plays action 1, and $\widehat{p}_j$ is the equilibrium probability that individual j plays action 1 (i.e. $\widehat{p}_i$ and $\widehat{p}_j$ are the solutions of the system of equations $\dot{p}_i = 0, \dot{p}_j = 0$).

Because there are two learning rules in the population (TR or HR), it is also convenient to use the subscripts in equation (7) to identify learning rules of interacting individuals, whereby $i,j \in \{\text{TR, HR}\}$. To use these payoffs to investigate evolutionary dynamics, we make the customary assumption that the population size is infinitely large so that we can use a deterministic evolutionary model. Calling q_i the frequency of learning rule $i \in \{\text{TR, HR}\}$ in the population, the expected reproductive output of an individual of learning rule $i \in \{\text{TR, HR}\}$ is then defined as

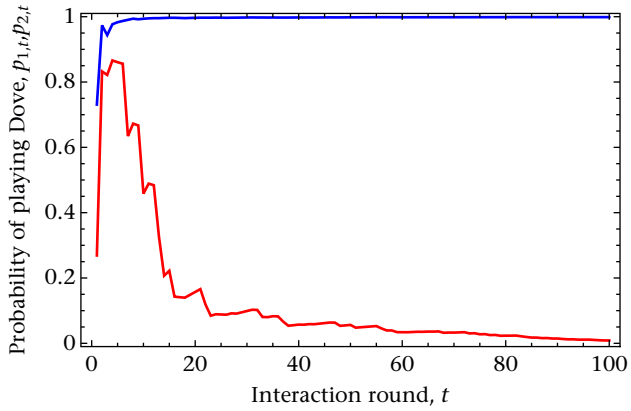

Figure 1. Example of learning dynamics for two interacting individuals (1 and 2) in a 2×2 Hawk–Dove game with $\pi(1,1) = B/2$, $\pi(1,2) = 0$, $\pi(2,1) = B$, $\pi(2,2) = B/2 - C$, where $B = 5$ and $C = 3$. The blue line represents the probability $p1,t$ of playing Dove for individual 1 and the red line the probability $p2,t$ of playing Dove for individual 2 when the learning rule is characterized by $\varphi i,t = 1 + 1/t$ and $\lambda i = 1$, for both players (rule called pure trial-and-error reinforcement learning, PTR). Parameter values for player 1 are $g1 = 0$, $M1,1$ (Dove) $= 1$ and $M1,1$ (Hawk) $= 0$ (hence $p1,1 \approx 0.73$), while for player 2 they are $g2 = 0$, $M2,1$ (Dove) $= 0$ and $M2,1$ (Hawk) $= 1$ (hence $p2,1 \approx 0.27$).

Table 1
Payoff matrix for a typical stage game ω, where $\omega \in \{\text{PD, CG, HD}\}$

	Action 1	Action 2
Action 1	$\mathscr{R}_\omega$	$\mathscr{S}_\omega$
Action 2	$\mathscr{T}_\omega$	$\mathscr{P}_\omega$

One player chooses a row and its opponent chooses a column. Payoffs are to row player. In order to numerically implement the three possible sub-games (PD, CG, HD) we used the following constraints on the payoffs. In the Prisoner's Dilemma game (PD): $\mathscr{T}_{\text{PD}} > \mathscr{R}_{\text{PD}} > \mathscr{P}_{\text{PD}} > \mathscr{S}_{\text{PD}}$ and $(\mathscr{T}_{\text{PD}} + \mathscr{S}_{\text{PD}})/2 < \mathscr{R}_{\text{PD}}$. In the Hawk-Dove game (HD): $\mathscr{T}_{\text{HD}} > \mathscr{R}_{\text{HD}}, \mathscr{S}_{\text{HD}} > \mathscr{P}_{\text{HD}}, \mathscr{P}_{\text{HD}} > \mathscr{R}_{\text{HD}}$. In the Coordination Game (CG): $\mathscr{R}_{\text{CG}} > \mathscr{S}_{\text{CG}}, \mathscr{R}_{\text{CG}} = \mathscr{P}_{\text{CG}}, \mathscr{S}_{\text{CG}} = \mathscr{T}_{\text{CG}}$.

Table 2
Local Stability analysis of the equilibria in the PTR vs. PTR case when the average game is the Prisoner's Dilemma

Equilibrium	Associated Eigenvalues	Eigenvalues' sign
(0,0)	(0,0)	(0,0)
(0,1)	$(-B,C)$	$(-,+)$
(1,0)	$(-B,C)$	$(-,+)$
(1,1)	$(C-B,C-B)$	$(-,-)$
$\left(1, \frac{B}{2B-C}\right)$	$\left(C + \frac{B^2}{C-2B}, B + \frac{B^2}{C-2B}\right)$	$(-,+)$
$\left(\frac{B}{2B-C}, 1\right)$	$\left(C + \frac{B^2}{C-2B}, B + \frac{B^2}{C-2B}\right)$	$(-,+)$
$\left(\frac{B+C}{2B}, \frac{B+C}{2B}\right)$	$\left(\frac{(B-C)^2(B+C)}{4B^2}, \frac{(B-C)(B+C)^2}{4B^2}\right)$	$(+,+)$

Table 3
Local Stability analysis of the equilibria for the PTR vs. PTR interaction in the Hawk-Dove game

Equilibrium	Associated Eigenvalues	Eigenvalues' sign
(0,0)	$\left(-\frac{B}{2},-\frac{B}{2}\right)$	(−,−)
(0,1)	$(-B,0)$	(−,0)
(1,0)	$(-B,0)$	(−,0)
(1,1)	$\left(-\frac{B}{2}+C,-\frac{B}{2}+C\right)$	(+,+)
(0,1/3)	$\left(-\frac{B}{3},\frac{B}{3}\right)$	(−,+)
$\left(\frac{1}{3},0\right)$	$\left(-\frac{B}{3},\frac{B}{3}\right)$	(−,+)
$\left(\frac{B}{2B+\sqrt{2B(B-C)}},\frac{B}{2B+\sqrt{2B(B-C)}}\right)$		(+,+)
$\left(\frac{B}{2B-\sqrt{2B(B-C)}},\frac{B}{2B-\sqrt{2B(B-C)}}\right)$		(−,+)

Expressions of the eigenvalues associated to the interior equilibria are too long to fit in the table.

$$f_i = \alpha + q_i\Pi_{ii} + q_j\Pi_{ij}, \tag{8}$$

where α is a baseline fecundity. In the following, we call $q_{TR} \equiv q$ the frequency of TR in the population (so that $q_{HR} = 1 - q$ is the frequency of HR). Hence, the change in frequency Δq of TR in the one-shot matching model over one iteration of the life-cycle is given by the discrete-time replicator dynamics

$$\Delta q = q(1-q)\left(\frac{f_{TR} - f_{HR}}{\bar{f}}\right), \tag{9}$$

where $\bar{f} = qf_{TR} + (1-q)f_{HR}$ is the mean reproductive output in the population. In the next section, we evaluate the replicator dynamics (equation (9)) for three different average games: the Prisoner's Dilemma (PD), the Hawk–Dove game (HD) and the Coordination game (CG) (Table 4). The details of the analysis are provided in Appendix 4. Note that we use p_i and p_j to denote the probability of playing action 1 when both individuals have the same learning rule (i.e. for the interactions PTR versus PTR and PHR versus PHR) but we use p_{TR} and p_{HR} for the interaction between a PTR and a PHR learner.

RESULTS: ONE-SHOT MATCHING

In this section, we present both analytical and simulation results for the three average games (Prisoner's Dilemma, Hawk–Dove and Coordination game). For each game, we first analyse the learning dynamics for $\phi_{i,t} = (1/t) + 1$ (section 'Equilibrium behaviour'), and use this to characterize analytically the evolutionarily stable learning rule (section 'ESS analysis'). For each game, we also compare our analytical results to individual-based simulations (section 'Simulations: dynamic learning rate'), and then extend the simulation analysis to the case where $\phi_i = 1$ (section 'Simulations: constant learning rate').

Table 4
Payoff matrices for the average games studied

$\overline{G}$	Action 1	Action 2
Action 1	$\mathscr{R}$	$\mathscr{S}$
Action 2	$\mathscr{T}$	$\mathscr{P}$
$\overline{PD}$	Cooperate	Defect
Cooperate	$B-C$	$-C$
Defect	B	0
$\overline{HD}$	Dove	Hawk
Dove	$B/2$	0
Hawk	B	$B/2-C$
$\overline{CG}$	Left	Right
Left	B	0
Right	0	B

The rows correspond to the actions of player 1 and the columns correspond to the actions of player 2. Payoffs are to row player. The matrix at the top shows the generic payoffs used in the paper. In the Prisoner's Dilemma game ($\overline{PD}$, second sub-table), we assume $B > C > 0$. In the Hawk-Dove game ($\overline{HD}$, third sub-table), we have $B > C > B/2 > 0$. In the Coordination game ($\overline{CG}$, fourth sub-table), $B > 0$.

Prisoner's Dilemma

Equilibrium behaviour

The learning dynamics for the Prisoner's Dilemma is obtained by substituting into equations (5) and (6) the payoffs $(\mathscr{R},\mathscr{S},\mathscr{T},\mathscr{P})$ defined in Table 4 for the corresponding average game. Hence, action 1 can now be thought of as 'Cooperate' and action 2 as 'Defect'. To determine the fate of PTR (equation (9)), we need to evaluate the payoffs for each possible interaction between types of learners, which will depend on the equilibrium points of the learning dynamics. In the population, three types of pairwise interaction can occur: (1) PTR versus PTR; (2) PHR versus PHR; (3) PTR versus PHR.

When a PTR is paired with another PTR we find that the learning dynamics can end in two possible states, depending on the initial preferences $(p_{i,1},p_{j,1})$ individuals have for each action. If individuals have a high enough initial probability of playing Cooperate, then both PTR will learn to cooperate $(\widehat{p}_i = 1, \widehat{p}_j = 1)$ at the equilibrium of learning (Fig. 2a and Appenidx 4). For other initial conditions, both PTR learn to defect $(\widehat{p}_i = 0, \widehat{p}_j = 0)$. The fact that cooperation is a possible endpoint of the learning dynamics can be intuitively understood by realizing that the payoff for mutual cooperation is positive $B - C > 0$, where B is the benefit received when the opponent cooperates and C the cost of cooperation (Table 4). Hence, if both players initially cooperate on some rounds, cooperation will be rewarded, and ultimately players will stick with Cooperation. Defection is also a stable equilibrium because an individual that experiences the high payoff $B > 0$ if it defects and its opponent cooperates will reinforce Defection. This will in turn lead the other player to Defect because cooperating against a defector is punished by the payoff $-C < 0$. The interaction between two PHR gives a different result: irrespective of initial conditions, both PHR will learn to defect (Fig. 2b). Finally, when a PTR meets a PHR, both individuals learn to defect regardless of initial conditions, which means that PTR does not get exploited by PHR (Fig. 2c). The behaviour of PHR stems from its ability to observe that defection is better than cooperation ($B > B - C$ and $0 > -C$), whatever the opponent chooses. A PTR individual does not get exploited by PHR because the payoff $-C < 0$ is punishing, and thus will be avoided by a trial-and-error learner.

ESS analysis

Using the above results on equilibrium action play, we can compute the fitness of both learning rules (equation (8)) by using Table 4 and calling k the cost of PHR for cognitive complexity. For the Prisoner's Dilemma, this gives

$$\begin{cases} f_{HR} = \alpha - k, f_{TR} = \alpha + q(B-C) & \text{if IC is in the basin of } (1,1), \\ f_{HR} = \alpha - k, f_{TR} = \alpha & \text{otherwise,} \end{cases} \tag{10}$$

where IC refers to the initial conditions of learning in the PTR versus PTR interaction. Under the replicator dynamics (equation (9)), the frequency of the PTR learning rule increases when $f_{TR} > f_{HR}$. In the

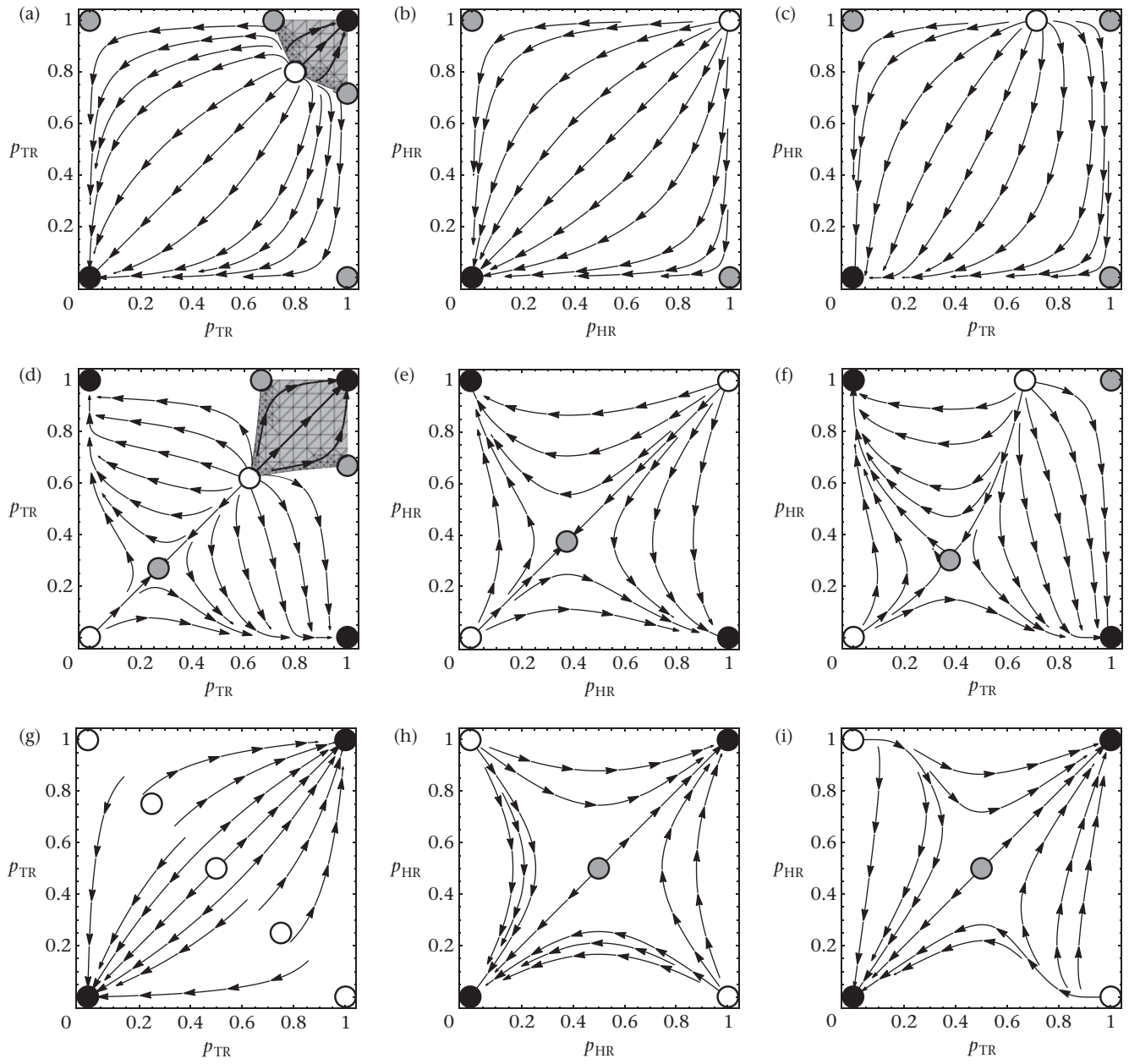

Figure 2. Solution orbits of the learning dynamics in the three average games for the one-shot matching model. (a, b, c) Prisoner's Dilemma ($B = 5$ and $C = 3$). (d, e, f) Hawk–Dove ($B = 5$ and $C = 3$). (g, h, i) Coordination game ($B = 5$). (a, d, g) The interaction between two TRs. (b, e, h) The interaction between two HRs. (c, f, i) The interaction between a TR and a HR. A white-filled circle denotes an unstable node (both associated eigenvalues are positive), a grey-filled circle is a saddle (one positive and one negative eigenvalue) and a black circle is a locally stable equilibrium. In (a) the grey shaded area represents the initial conditions for which all trajectories go to the (Cooperate, Cooperate) (1, 1) equilibrium. In (d) the grey shaded area represents the initial conditions for which all trajectories go to the (Dove, Dove) (1, 1) equilibrium.

first case of equation (10) (i.e. when we start close enough to the equilibrium (1,1)), this means that PTR invades PHR if

$$q(B - C) + k > 0. \tag{11}$$

Because the left-hand side is always positive, PTR is the evolutionarily stable learning rule (ESLR); that is, it cannot be invaded by PHR and can invade PHR. Note that, even when $k = 0$, PTR is still the ESLR.

In the second case of equation (10), i.e. when the learning dynamics start in the basin of (Defect, Defect) of the PTR versus PTR interaction, both PTR and PHR learn to defect. Here, $f_{TR} > f_{HR}$ always holds when $k > 0$, hence PTR is also the ESLR. For $k = 0$, we have neutrality ($f_{TR} = f_{HR}$ for all q).

Why does PTR outcompete PHR? The reason is that trial-and-error learners reinforce positive payoffs and so a pair of PTR individuals learn to Cooperate when they have an initial tendency to do so. In that case a PTR can avoid being exploited by a defector PHR and learns to defect, since playing cooperate then leads to very low payoffs. Hence, if $k > 0$ and there is initially a small frequency of PTR individuals ($q > 0$), these can always invade a population of PHR individuals and establish cooperation if the initial conditions of the learning dynamics favour it; otherwise they will maintain defection. By contrast, a small frequency of PHR individuals cannot invade a population of PTR individuals since they cannot exploit them. Note that the advantage of PTR over PHR when individuals initially prefer cooperation still holds even when there is no cost of cognition ($k = 0$).

While we consider the initial preferences as parameters of the model, and not as evolving characteristics of the individuals, these results suggest that it is unlikely that a mutant PTR with an initial genetic tendency for defection invades a population of PTR that initially prefer to cooperate. Indeed, pairs of PTR defectors will obtain a lower payoff than pairs of PTR cooperators, so they will not be able to displace cooperators. This illustrates that in our one-shot matching setting there is a selection pressure in favour of cooperation.

Individual-based simulations

To see whether the above analytical approximation reflects accurately the underlying stochastic model of learning and evolution, we performed individual-based simulations. We ask two questions in the simulations. (1) What are the parameter values for which the above approximation works? (2) To what extent can the results where an approximation works be generalized to the (more realistic) case where the learning rate is constant? As to question (1), we find that an exploration parameter of $\lambda = 10$ and a number $T = 500$ of interactions in the game lead to similar results in the simulations and the analysis. Given the values we used for the payoffs, a bigger value of λ would induce almost deterministic behaviour after the first few steps of exploration, which would lead to an irreversible escape of the basins of attractions of predicted equilibria. A smaller value of λ would induce very slow convergence to the equilibria and would thus imply that a higher T is needed to observe correspondence between analysis and simulations. Regarding question (2), we generally find a good correspondence between the results for the two studied learning rates in this one-shot matching model (Table 5). In Appendix 5, we provide a detailed description of these simulations.

Simulations: dynamic learning rate

Since the analytical prediction in the Prisoner's Dilemma depends on whether individuals have an initial preference for Cooperation or Defection (section 'ESS analysis'), we ran a set of simulations for each type of initial preference (see Appendix 6 for a detailed description of results for both dynamic learning rate and constant learning rate for all studied games).

Both players initially prefer cooperation. For this case, simulation results are similar to the analytical results (section 'ESS analysis'). In particular, two PTR players that are paired with one another can learn to cooperate, while a PTR player that is paired with PHR always learns to defect (Fig. 3a, b, c). This leads to the fixation of PTR in the population. However, it is noteworthy that the simulations differ slightly from the analysis: pairs of PTR can sometimes learn to defect (Fig. 3a), which is a consequence of the possibility of escaping the basin of attraction of cooperation due to stochastic fluctuations. This does not affect our evolutionary prediction (Table 5): PTR players will obtain a higher fitness than PHR because some pairs of PTR learn to cooperate while PHR individuals defect and only meet defectors. This is true for any probability P that pairs of PTR learn to cooperate. Indeed, in our large population, P will also be the proportion of PTR pairs that learn to cooperate, so the fitness of PTR reads

$$f_{\mathrm{TR}} = \alpha + q[P(B - C) + (1 - P)0], \tag{12}$$

while the fitness of PHR is only $f_{\mathrm{HR}} = \alpha - k$, so for any positive P and any q, we have $f_{\mathrm{TR}} > f_{\mathrm{HR}}$.

Both players initially prefer defection. For these initial preferences, the simulation results differ from the analytical prediction regarding equilibrium action play (section 'Equilibrium behaviour') and also regarding the evolutionary outcome (section 'ESS analysis'), because pairs of PTR individuals sometimes learn to cooperate despite their initial tendency to defect. Hence, certain pairs of PTR individuals learn to cooperate and some other pairs learn to defect, while the interactions involving PHR always lead to Defection (Fig. 3d, e, f). We then are in the situation described by equation (12) so that the fitness of PTR is higher than that of PHR, even if the value of P is here much smaller than when individuals initially preferred cooperation (compare Fig. 3a and d). As a consequence, we indeed observe that PTR fixes in the population in our evolutionary simulations (Table 5).

Simulations: constant learning rate

Both players initially prefer cooperation. In this scenario, the same qualitative results as in the dynamic learning rate situation are obtained. Namely, pairs of ETR can learn to cooperate (Fig. A1a, b, c) so the ETR learning rule fixes in the population (Table 5).

Both players initially prefer defection. In this situation there is only one (but important) difference compared to the dynamic learning rate case. Namely, when ETR individuals are paired with EHR, we observe that ETR do not learn full defection and converge to a positive probability of cooperating, while EHR always learn to defect. This gives an evolutionary advantage to the latter when present in high frequency in the population (Fig. A1d, e, f). This implies that in our simulations of evolution we observe that ETR fixes when the population initially consists only of ETR, but EHR fixes when the population initially consists only of EHR (Table 5). In other words, mutants cannot invade in this case.

Table 5
Summary of results in the one-shot matching model

Average Game	Initial condition of learning	Predicted q^*	Learning rate	Simulated q^*
Prisoner's Dilemma	Basin of (Cooperate, Cooperate) of TRvsTR	1	Dynamic	0.99
			Constant	0.98
	Basin of (Defect, Defect) of TRvsTR	1/2	Dynamic	0.98
			Constant	0.5[†]
Hawk-Dove Game	Basin of (Hawk, Dove) of TRvsHR	1	Dynamic	0.97
			Constant	0.98
	Basin of (Dove, Hawk) of TRvsHR	0	Dynamic	0.01
			Constant	0.01
Coordination Game	Basin of (Left, Left)	1/2	Dynamic	0.27
			Constant	0.56

The column "Predicted q^*" shows the frequency of TR expected at evolutionary equilibrium under the deterministic approximation. The column "Simulated q^*" gives the approximate equilibrium frequency of TR obtained in the corresponding evolutionary simulation. The simulation results represent the average over simulation runs with different initial compositions of the population (see Appendix 5 for details).

[†] In this case, the different simulation runs give disparate results with either learning rule getting fixed depending on the initial conditions (see main text).

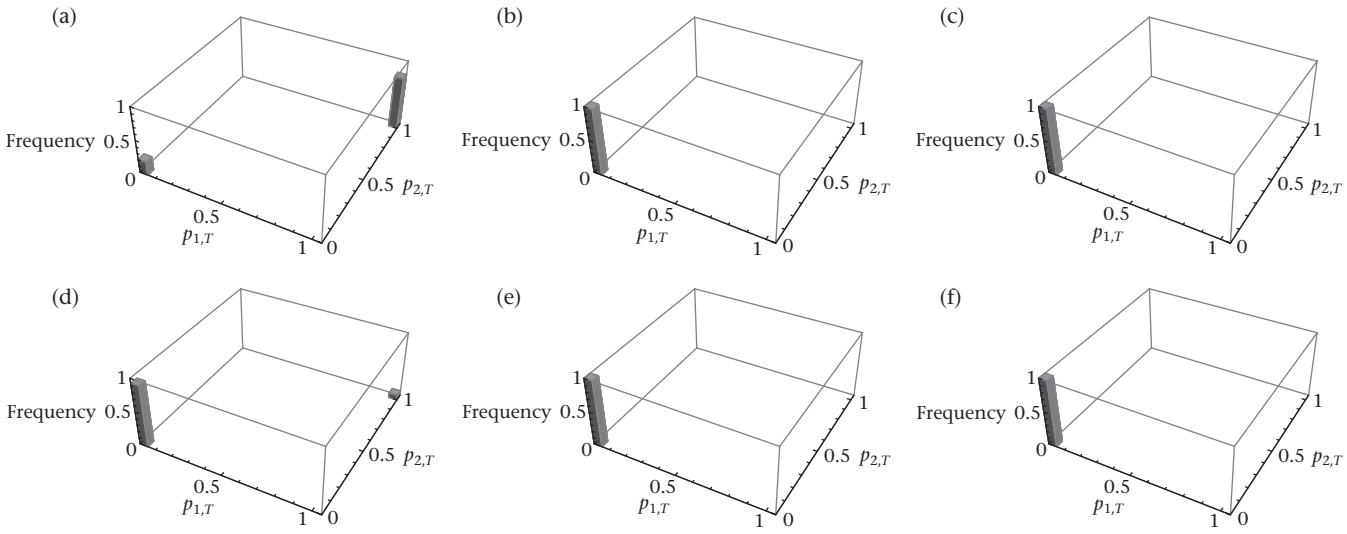

Figure 3. Behavioural equilibrium of learning in the average Prisoner's Dilemma for the one-shot matching model with dynamic learning rate for pairs of opponents (simulation test of the analytical predictions of the section 'Equilibrium behaviour'). This represents the frequency of pairs having reached given probabilities of playing action 1 ($p_{1,T}$, $p_{2,T}$) at the end of the individual's lifetime, T. We used a total of 1000 individuals of each learning rule in each simulation. (a, b, c) Initial preference for Cooperation ($p_{i,1} = 0.85$). (d, e, f) Initial preference for Defection ($p_{i,1} = 0.15$). (a, d) Interaction between two TRs. (b, e) Interaction between two HRs. (c, f) Interaction between TR (player 1) and HR (player 2).

Hawk–Dove Game

Equilibrium behaviour

To analyse the dynamics of learning (equations (5) and (6)) for the HD game, we use the corresponding payoffs given in Table 4. Now, action 1 corresponds to 'Dove' and action 2 to 'Hawk'.

In the PTR versus PTR interaction, the learning dynamics have three pure stable equilibria: (Hawk, Dove), (Dove, Hawk) or (Dove, Dove). In other words, depending on the initial conditions, individuals will reach either a Nash equilibrium or the 'cooperative' outcome where both individuals choose Dove, a result similar to that one obtained in the average PD game (Fig. 2d, Appendix 4). This can be explained in terms of the signs of the payoffs and in terms of reward and punishment. The (Dove, Dove) outcome is rewarded (it yields $B/2 > 0$) when players try the Dove action, so it is a possible endpoint of the learning dynamics. The eventually stick with Dove (because $0 > B/2 - C$). Finally, when a PTR meets a PHR, there are two possible endpoints: the equilibrium where PTR plays Hawk and PHR plays Dove; or the reverse situation where PTR learns to play Dove and PHR learns to play Hawk (Fig. 2f). Hence, in this heterogeneous interaction, depending on the initial conditions, either PTR or PHR will get exploited by its opponent.

ESS analysis

The homogeneous interactions (PTR versus PTR and PHR versus PHR) always lead to a payoff of $B/2$, irrespective of the initial preferences for actions, because both individuals have equal chances of learning to become a Hawk or a Dove. However, the payoffs in the heterogeneous interaction between PTR and PHR depend on whether the initial condition is in the basin of attraction of (0,1) or (1,0). Thus, the reproductive output is

$$\begin{cases} f_{\mathrm{HR}} = \alpha + qB + (1-q)\dfrac{B}{2} - k, f_{\mathrm{TR}} = \alpha + q\dfrac{B}{2} & \text{if IC. in basin of } (0,1) \\ f_{\mathrm{HR}} = \alpha + (1-q)\dfrac{B}{2} - k, f_{\mathrm{TR}} = \alpha + q\dfrac{B}{2} + (1-q)B & \text{if IC. in basin of } (1,0), \end{cases} \tag{13}$$

Nash equilibria are also endpoints because if one player plays Hawk against a player that plays Dove, it will also reinforce Hawk. Its opponent cannot reinforce Hawk because this would yield a punishment ($B/2 - C < 0$), hence (Hawk, Dove) is a stable equilibrium. When two PHR interact (Fig. 2e), they end up playing one of the two Nash equilibria (Hawk, Dove) or (Dove, Hawk). The reason why (Dove, Dove) is not an equilibrium for pairs of PHR is that this outcome induces a 'regret' in both players of not having played the Hawk action (because they observe that they could have obtained $B > B/2$). But if both players switch to Hawk, this is also an unsatisfactory outcome and one of the players will

where IC is the initial condition of the PTR versus PHR interaction.

In the first case of equation (13), PTR increases in frequency when rare ($f_{\mathrm{TR}} > f_{\mathrm{HR}}$) if

$$k - \frac{B}{2} > 0. \tag{14}$$

This implies that when $k > B/2$, PTR is the ESLR; when $k = B/2$ evolution is neutral; when $k < B/2$, PHR is the ESLR.

In the second case of equation (13), PTR increases in frequency if

$$\frac{B}{2}+k>0, \tag{15}$$

which is always true because $B>0$ and $k\geq 0$. In this case, PTR is the only ESLR.

In other words, in the HD game the learning rule that is the ESLR is the one that has the greatest initial preference for playing Hawk (in the absence of cost, $k=0$) because this allows it to learn the Hawk action against the other learning rule that learns to play Dove. The learner playing Hawk will thus obtain a larger payoff. Indeed, either learning rule with any initial preference for actions in a monomorphic population obtains an average payoff of $B/2$ (as explained above). Hence, to determine whether a mutant learning rule can invade, one should determine whether the mutant obtains a higher payoff than $B/2$ when paired with the resident. The learning rule that prefers Hawk can invade when the resident prefers Dove because it obtains the payoff B of the (Hawk, Dove) outcome. On the other hand, the mutant preferring Dove is not able to invade when the resident prefers Hawk because the mutant will obtain the payoff 0 of the (Dove, Hawk) outcome. Hence the learning rule that prefers Hawk can invade and is stable against invasion, and is consequently the ESLR in the Hawk–Dove game in the absence of cost.

Simulations: dynamic learning rate

PTR initially prefers to play Hawk and PHR prefers Dove. In this case, the simulation results agree qualitatively well with the analytical results (section 'ESS analysis'). The only difference is that in the simulations, we observe some pairs of PTR that learn the outcome (Dove, Dove) (Fig. 4a, b, c), while we expect them to learn (Hawk, Dove) or (Dove, Hawk) under the above analysis. This is due to the possibility of escaping a basin of attraction in the stochastic learning model (as occurred above for the PD). However, this small difference does not affect the evolutionary outcome because the average payoff for pairs of PTR is the same at the equilibrium (Dove, Dove) as it is in the equilibria (Hawk, Dove) or (Dove, Hawk). Consequently, the analytical prediction (section 'ESS analysis') regarding evolution applies and we indeed observe that PTR fixes in the population in the long run (Table 5).

PTR initially prefers to play Dove and PHR prefers Hawk. As explained in the ESS analysis above (section 'ESS analysis'), the equilibrium behaviour of homogeneous pairs (PTR versus PTR and PHR versus PHR) does not depend on initial conditions (one member of the pair learns Hawk and the other learns Dove). However, the equilibrium behaviour in the heterogeneous interaction (PTR versus PHR) does depend on the initial preferences for actions. In the current scenario, according to the above ESS analysis, PHR should learn to play Hawk and PTR should learn to play Dove. All of these predictions are observed in the simulations of learning (Fig. 4d, e, f). Thus the analytical prediction regarding which learning rule will fix in the evolutionary long run is also verified in the evolutionary simulations: PHR individuals fix in the population (Table 5).

Simulations: constant learning rate

ETR initially prefers to play Hawk and EHR prefers Dove. Here the results are the same as under the dynamic learning rate and ETR fixes in the population in the long run (Fig. A2a, b, c, Table 5).

ETR initially prefers to play Dove and EHR prefers Hawk. This case is similar to the situation with the dynamic learning rate, where EHR individuals outcompete ETR and fix to a frequency close to 1 at the equilibrium of evolution (Fig. A2d, e, f, Table 5).

Coordination Game

Equilibrium behaviour

If we use the payoffs of the CG game (Table 4) in equations (5) and (6), we obtain the learning dynamics of a pair of opponents in the coordination game. In the labelling of Table 4, action 1 corresponds to Left and action 2 corresponds to Right. In this game, all three types of pairs succeed in learning to coordinate in the long run (Fig. 2g, h, i), and depending on the initial preferences for Right or Left, the equilibrium reached will either be (Right, Right) or (Left, Left).

ESS analysis

Under the three types of interactions, the players coordinate on a single action and get a payoff at equilibrium of B. The fitness of the

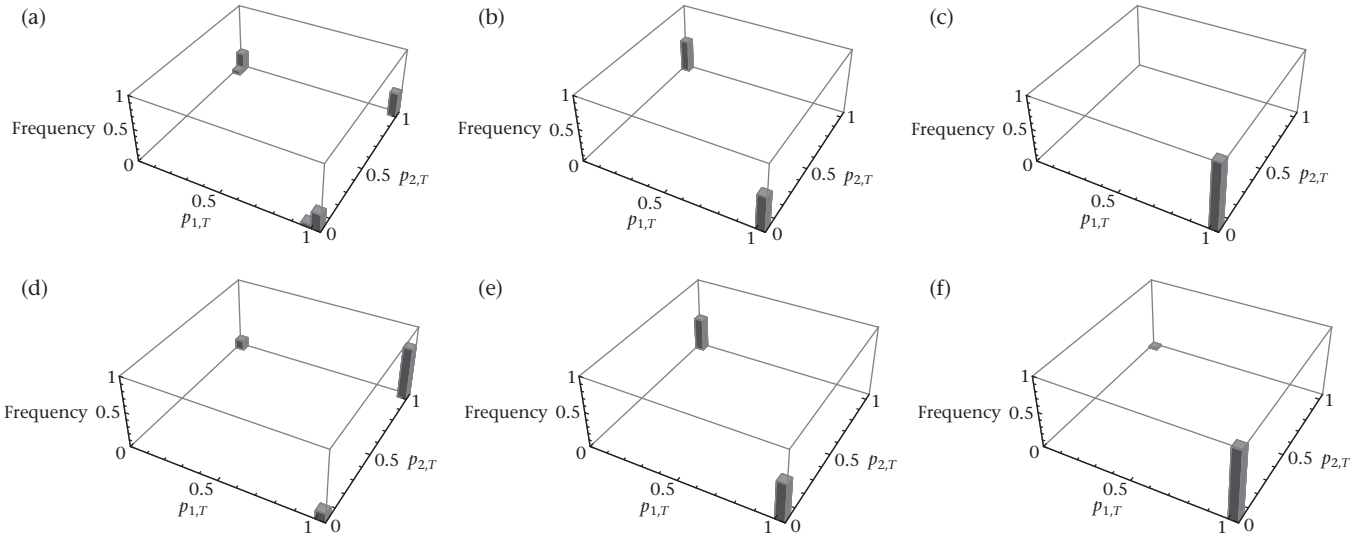

Figure 4. Same as Fig. 3 describing the results for the behavioural equilibrium under the average Prisoner's Dilemma but now for the average Hawk–Dove game. (a, b, c) TR initially prefers Hawk ($p_{TR,1}=0.15$) and HR prefers Dove ($p_{HR,1}=0.85$). (d, e, f) TR initially prefers Dove ($p_{TR,1}=0.85$) and HR prefers Hawk ($p_{HR,1}=0.15$). (a, d) Interaction between two TRs. (b, e). Interaction between two HRs. (c, f). Interaction between TR (player 1) and HR (player 2).

PTR learning rule is then $f_{TR} = \alpha + B$ and the fitness of PHR is $f_{HR} = \alpha + B - k$. Trivially, PTR is the ESLR for all positive k; for $k = 0$, evolution is neutral. This is because both learning rules get the average payoff B in all types of interactions, so no learning rule is able to invade a population consisting of the other learning rule. Hence, under the recurrent inflow of symmetric mutations, we expect, in the long run, to observe a stationary state of the population where the average frequency of both learning rules is 1/2.

Simulations: dynamic learning rate

In this game, our analytical results predict that evolution does not favour one learning rule over the other for all initial conditions because both learning rules always induce coordination on a single action (section 'ESS analysis'). Because our evolutionary simulations feature recurrent mutations, this situation should lead to neutral evolutionary dynamics where the expected frequency of both types under the stationary distribution of the process is equal ($q = 1/2$). In the simulations, we arbitrarily set initial preferences of individuals close to the outcome (Left, Left) and we indeed observed that all types of pairs of individuals succeed in coordinating on a given action (but the particular action learned depends on the pair, Fig. 5). This learning behaviour corresponds to our analytical prediction (section 'Equilibrium behaviour') but simulations of evolution do not match this, and we observe a mixed evolutionary equilibrium dominated by PHR (Table 5). This result can be explained by game stochasticity. Indeed, when we simulate evolution by letting the Coordination game be the only game played by individuals (this gives a situation with a constant game instead of fluctuating games), we observe that both learning rules coexist in near-equal frequency at evolutionary equilibrium (Fig. A6).

Simulations: constant learning rate

Learning in this case also leads to coordination of all pairs (Fig. A3). Evolutionary simulations give a result close to what is expected, with a frequency of ETR of $q \approx 0.56$ at the equilibrium of evolution (Table 5).

RESULTS: REPEATED MATCHING

We now assume that individuals in the population are randomly paired at every time t of their lifetime, but otherwise keep all previous assumptions. An individual will now meet different partners during its lifetime. Its learning dynamics may then depend on the distribution of behaviour of all individuals in the population because anybody can be met for a one-shot interaction.

Unfortunately, we could not find any simple analytic approximation to the evolutionary dynamics for this matching model, so we used exclusively individual based simulations to investigate the evolutionary stability of TR and HR.

The simulations for the repeated matching model follow the same two steps applied for the one-shot matching model. First, we simulated only learning dynamics in a large population (this step can be thought of as simulating the behavioural dynamics of a single generation). These simulations of learning were performed for various values of q, the frequency of TR in the population. To understand how the frequency of the learning rules affected the learning dynamics, we computed the behavioural equilibrium of TR and HR for all frequencies q (Fig. 6). Second, we simulated the full evolutionary process, namely, learning and reproduction, over many generations, in order to track the frequency of TR and HR over evolutionary time (Table 6). In these two types of simulations (learning and evolutionary), we used the same initial conditions for the learning dynamics that we used in the one-shot matching model, so that we can compare the results of the two matching schemes (for more details on the methodology of the simulations, see Appendix 5). The results of the evolutionary simulations can be interpreted in light of the simulations of learning dynamics.

Prisoner's Dilemma

Dynamic learning rate

All individuals initially prefer cooperation. For this case, when we simulate learning in the population, we obtain that PTR individuals can learn to cooperate when very common in the population (precisely, when $q \geq 0.8$), while PHR individuals always learn to defect for all compositions of the population (Fig. 6a). Consequently, the fitness of PHR and PTR are the same for $q < 0.8$ (because everybody defects) but PTR has a smaller fitness when $q \geq 0.8$, because it cooperates a positive proportion of the time, which leads PHR to exploit PTR. This implies that, on the evolutionary timescale, the population will move neutrally through all the states such that $q < 0.8$, but is repelled from the states where $q \geq 0.8$. As a consequence, we observe in our evolutionary simulations that the equilibrium frequency of PTR is small but positive (Table 6).

All individuals prefer defection. With these initial preferences, the simulations of learning show that both learning rules always converge to Defection, so no one has an advantage (Fig. 6b). When we simulate evolution, the result is qualitatively in agreement with this: the frequency of PHR at evolutionary equilibrium is slightly above 0.5 (Table 6). When both learning rules have the same fitness, we expect that mutation in the population will maintain both learning rules at a frequency of 0.5. The slight deviation from 0.5 that we observed can be explained by the variance in convergence time of PTR individuals. Some of them might converge more slowly to full Defection and will be exploited on some interaction rounds when they meet the defector PHR (Fig. A4).

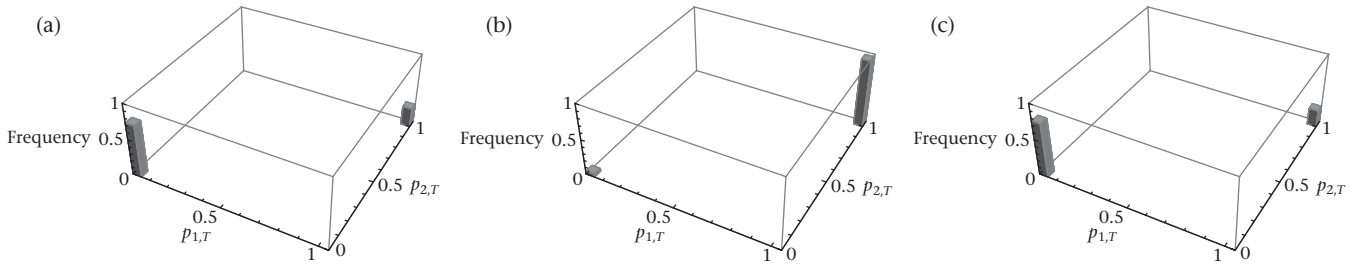

Figure 5. Same as Fig. 3 describing the results for the behavioural equilibrium under the average Prisoner's Dilemma but for the behavioural equilibrium under the average Coordination game. (a) Interaction between two TRs. (b) Interaction between two HRs. (c) Interaction between TR (player 1) and HR (player 2).

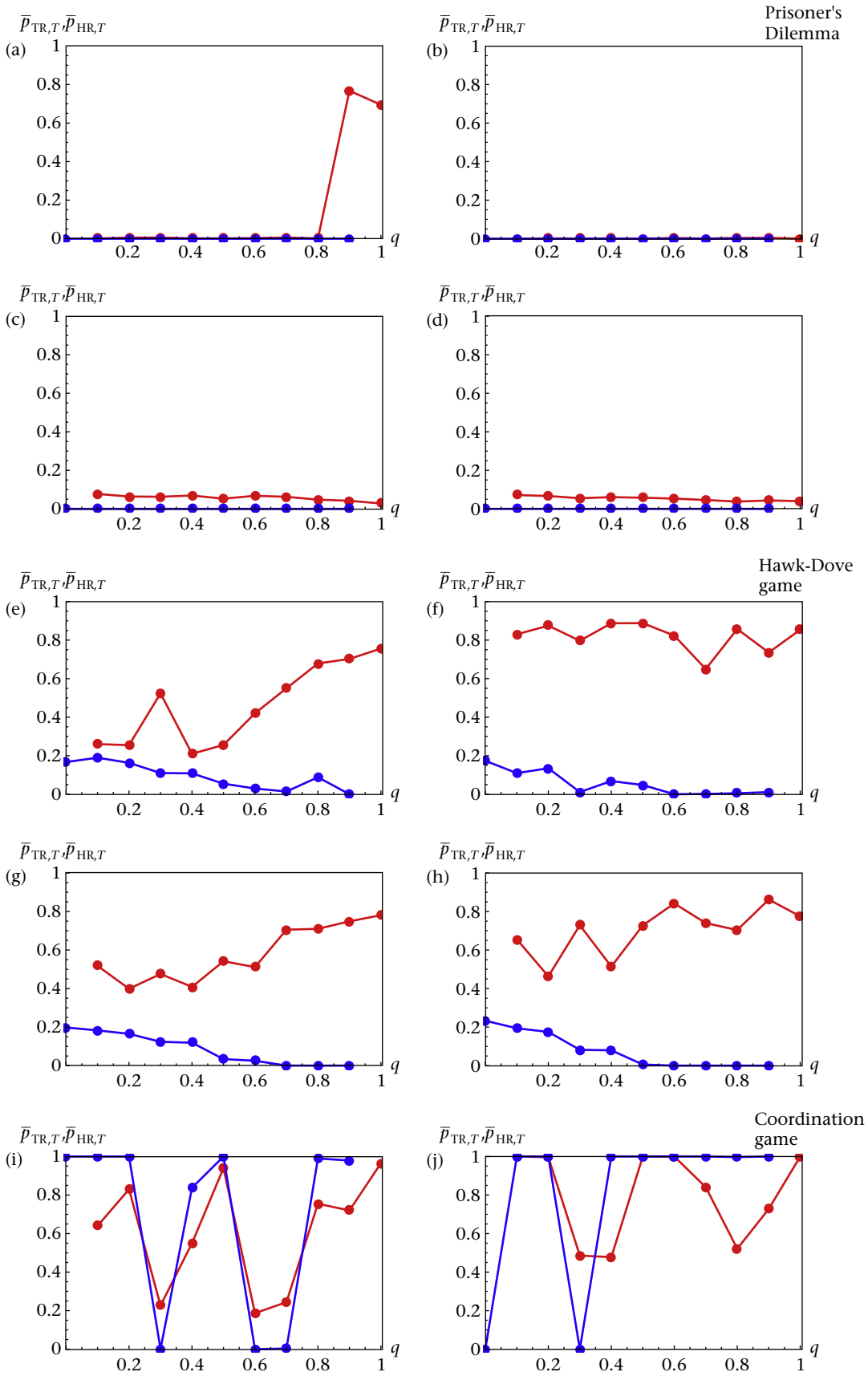

Figure 6. Average probability of choosing action 1 at the behavioural equilibrium for different frequencies q of TR in the population in the repeated matching model. Red line: TR; blue line: HR. (a, b, c, d) Prisoner's Dilemma. (e, f, g, h) Hawk–Dove game. (i, j) Coordination game. For the Prisoner's Dilemma: (a, b) dynamic learning rate; (c, d) constant learning rate. Individuals start with an initial preference for (a, c) Cooperation ($p_{i,1} = 0.85$) or (b, d) Defection ($p_{i,1} = 0.15$). For the Hawk–Dove game: (e, f) dynamic learning rate; (g, h) constant learning rate. (e, g) TR initially prefers Hawk ($p_{TR,1} = 0.15$) and HR prefers Dove ($p_{HR,1} = 0.85$). (f, h) TR initially prefers Dove ($p_{TR,1} = 0.85$) and HR prefers Hawk ($p_{HR,1} = 0.15$). For the Coordination game: (i) dynamic learning rate; (j) constant learning rate.

Table 6
Summary of results in the repeated matching model

Average Game	Initial condition of learning	Learning rate	Simulated q^*
Prisoner's Dilemma	All individuals prefer Cooperation	Dynamic	0.2
		Constant	0.05
	All individuals prefer Defection	Dynamic	0.48
		Constant	0.04
Hawk-Dove Game	TR prefers Hawk, HR prefers Dove	Dynamic	0.18
		Constant	0.11
	TR prefers Dove, HR prefers Hawk	Dynamic	0.07
		Constant	0.05
Coordination Game	All individuals prefer Left	Dynamic	0.01
		Constant	0.06

This is the same table as for the one-shot matching model (Table 5) but without analytic prediction.

Constant learning rate

All individuals initially prefer cooperation. Here ETR players learn to cooperate with a small probability for all frequencies, while EHR learn to defect as usual (Fig. 6c). As a consequence, the frequency of EHR in evolutionary simulations is close to 1 at equilibrium (Table 6).

All individuals initially prefer defection. The result is here the same as in the previous section: ETR individuals learn to cooperate with a positive probability (Fig. 6d) which leads to the evolutionary success of EHR observed in our evolutionary simulations (Table 6).

Hawk–Dove Game

Dynamic learning rate

PTR initially prefers to play Hawk and PHR prefers Dove. Here, we obtain that individuals using the PTR rule learn to play Dove with a higher average probability than PHR does. As the frequency of PTR increases, their probability of playing Dove also increases, while PHR plays Hawk more and more often (Fig. 6e). Consequently, when we perform simulations of evolution, we observe a stable polymorphism with a clear domination of PHR (Table 6). The reason is that playing Hawk with a higher probability than the opponent is beneficial in one-to-one interactions (which favours PHR in interactions against PTR) but playing Hawk too often renders the population susceptible to invasion by individuals that play the nicer Dove action (which favours PTR).

PTR initially prefers to play Dove and PHR prefers Hawk. The learning behaviour of the learning rules is similar to the previous case but now the average probability that PTR plays Dove is less affected by its frequency, and is always relatively high. Individuals using PHR still have a tendency to increase their probability of playing Hawk as q increases (Fig. 6f), which gives them an advantage. This is confirmed by our evolutionary simulations where PHR dominates the population at a polymorphic evolutionary equilibrium (Table 6).

Constant learning rate

ETR initially prefers to play Hawk and EHR prefers Dove. Here the results are similar to the case with a dynamic learning rate so ETR learns to play Dove more often than PHR (Fig. 6g), such that the simulations of evolution lead to a polymorphic equilibrium, where PHR constitutes almost all the population (Table 6).

ETR initially prefers to play Dove and EHR prefers Hawk. The results for the learning behaviour are not different from the case with a dynamic learning rate (Fig. 6h) so we also observe in our simulations of evolution that EHR constitutes almost all the population at an evolutionary endpoint (Table 6).

Coordination Game

Dynamic learning rate

The results for the learning behaviour show here that PTR individuals fail to all coordinate on a given action. The PHR individuals are efficient in doing this (Fig. 6i) and thus fix in the population (Table 6).

Constant learning rate

The learning behaviour of both learning rules is qualitatively similar to the dynamic learning rate case (Fig. 6j) and this explains why EHR almost fix in the population (Table 6).

Robustness to Changes in λ *and* T

In order to check whether the above results are robust to changes in parameter values, we ran simulations for other values of λ (the sensitivity to payoffs) and T (the length of the lifespan). The results of these simulations are reported in Appendix 8 (Tables A1–A10). We find that for small λ and T values, the number of situations where trial-and-error learning (TR) subsists in the population is generally reduced. In the PD game, TR is no longer evolutionarily stable, but our qualitative results continue to hold for the other two games (e.g. Table A5). Lower values of λ and T generate an evolutionary pressure for learning speed, in which case hypothetical reinforcement learning (HR) is favoured in the PD game because an HR individual starts defecting when paired with a TR. By contrast, TR individuals switch to defection against HR only after having experienced a sequence of punishments with the payoff $-C$ (Fig. A4). Hence, when λ and T are small, TR individuals do not have the time to switch to Defection against HR but pairs of TR still learn to cooperate (Fig. A5). Although lower values of λ generally induce slower learning for any learning rule, our additional results suggest that HR are less affected than TR by a reduction in the learning speed (i.e. HR individuals are faster than TR). Higher values of λ induce faster learning and the results are essentially identical to the case with an intermediate value of λ (regardless of the value of T). However, high λ seems to slightly favour more TR than under intermediate λ (Tables A2, A4).

In summary, only small λ affect our qualitative results (i.e. TR is not an ESLR in the PD game), and this effect is increased when T is small. Otherwise varying these parameter values does not change the qualitative results, because we still observe situations where the simple TR outcompete the more complex HR, especially in the PD game (e.g. Tables A6–A7).

SUMMARY AND DISCUSSION

To determine whether selection favours cognitively more sophisticated individuals than simple trial-and-error learning in social interactions, we analysed an evolutionary model of the competition between trial-and-error and hypothetical reinforcement learning. Trial-and-error learners only use information about realized payoffs, which is a common form of learning in animals (e.g. Mery & Kawecki, 2002; Staddon & Cerutti, 2003), while hypothetical reinforcement learners also use information about the foregone payoffs of playing alternative actions (where this information can come from either social copying or active reasoning). This learning mode can be thought of as one step up in the cognitive hierarchy and is a special case of standard belief-based learning (see equation (A1.9) in Appendix 1), which is a form of learning that has been suggested to occur in animals (van Gils et al., 2003; Lima,

1984; Luttbeg & Warner, 1999; Valone, 2006) and humans (Camerer, 2003; Chmura et al., 2012; Feltovich, 2000).

To analyse the coevolution of trial-and-error and hypothetical reinforcement learning, we assumed that individuals in a large population are genetically programmed to be either trial-and-error learners or hypothetical reinforcement learners and interact repeatedly in a two-player, two-action stochastically fluctuating game. We defined two matching schemes, that is, two ways in which individuals meet to play the games. In the one-shot matching model, individuals are paired at the beginning of the fluctuating game and each pair interacts for the rest of the game. In the second model, we used repeated matching: here, a random matching is realized at each period of the game so that an individual has a negligible probability of playing twice against the same partner (since we also consider that the population is very large). Payoffs are evaluated at the equilibrium of the learning process, and this defines the number of offspring produced (fecundity) by an individual.

We applied stochastic approximation theory to analyse learning during an individual's lifetime, and obtained that the equilibrium behaviour of the learners could be characterized in terms of an average game of the fluctuating game (i.e. a game whose payoffs are averages of the subgames' payoffs of the original fluctuating game). We thus analysed three standard cases of the average game: the Prisoner's Dilemma, the Hawk–Dove game and the Coordination game, and checked our analytic approximations by running simulations of the exact process.

Evolutionary Outcomes

Overall, the presupposed domination of hypothetical reinforcement learning over trial-and-error described in the introductory section is not complete in our results. In other words, the ability to obtain information about payoff outcomes of unchosen actions does not necessarily give a selective advantage in social interactions. We do not claim that a selective advantage of hypothetical reinforcement is unlikely, but we have found a set of social situations and learning dynamics where this ability does not provide a selective advantage. In general, we observed three main types of results, which hold for the two learning rates we studied (constant or dynamic).

(1) In simple social interactions, where the average game faced by individuals only requires that two partners coordinate on the same action or 'anticoordinate' on two different actions, trial-and-error and hypothetical reinforcement learning do not produce different behaviours at the behavioural equilibrium. In this case, natural selection does not favour one learning rule over the other.

(2) In the class of Prisoner's Dilemma games parametrized by a benefit and a cost variable, the ability of two trial-and-error learners to generate cooperative pairwise interactions by reinforcement of actual rewards, rather than to play a Nash equilibrium, can give them a selective advantage over hypothetical reinforcement learners. The main explanation of this result is that trial-and-error learners can learn any outcome that yields positive payoffs (which is the case when both members of a pair cooperate), while hypothetical reinforcement learners reach the Nash equilibrium of the one-shot game (which is to defect), but miss the important information that the game is repeated (because they are not forward looking). However, in the repeated Prisoner's Dilemma, when individuals with the same learning rule cooperate with each other, they are favoured over full defectors (if they avoid being exploited by defectors). Importantly, we obtained this result even when hypothetical reinforcement learning incurred no cost for cognitive complexity.

(3) We observed many examples where hypothetical reinforcement learning dominates the population at an evolutionary endpoint (i.e. it either gets to fixation or the population reaches a polymorphism where hypothetical reinforcement learning is at a high frequency), especially when two individuals cannot interact more than once (i.e. in the repeated matching model). Since hypothetical reinforcement learning produces a Nash equilibrium at the level of the one-shot average game, this makes perfect sense, because it is the type of behaviour that is selected when individuals cannot interact more than once.

Across all these results, the interaction between the learning rule and the initial preferences (which can be interpreted as an innate predisposition for a certain type of action) plays an important role for the outcome. We observed that this predisposition can be overcome or reversed after many learning rounds, but this is much constrained by the dynamic properties of the interacting learning rules (e.g. the size of the basin of attraction of the behavioural equilibria). This should be kept in mind in the following discussion about the more specific effect on evolutionary outcomes of the two different matching models.

One-shot Matching

The one-shot matching model allows one to capture situations where the same pair of animals interact many times together, so that each animal has the opportunity to adapt its behaviour during its lifetime to the actions of its partner. Examples of such situations may include communal breeding species that live in relatively small stable groups like meerkats, *Suricata suricatta*, cooperative breeding cichlids, or chestnut-crowned babblers, *Pomatostomus ruficeps*. But one-shot matching also captures interactions between males and females in monogamous species (Black, 1996; Kleiman, 1977), where unrelated partners interact for very long periods of time, possibly under conditions of environmental change.

In the one-shot matching model, we either observed results of type (1) or type (2) but we never observed results of type (3). Results of type (1) were obtained when the average game was either a Hawk–Dove game or a Coordination game. In these two games, the two learners did not differ from one another at a behavioural equilibrium, which leads to the absence of an advantage of one learning rule over the other. In the Coordination game, we found that individuals of both learning rules generally succeeded in coordinating on a single action, because it is only necessary to repeat the action that leads to positive payoffs, and both the trial-and-error and hypothetical reinforcement learning are capable of this.

The Hawk–Dove game favours one learning rule or the other depending on the initial conditions, so no learning rule is favoured under all conditions, because both learning rules are able to reach the optimal (Nash equilibria) outcomes (Hawk, Dove) and (Dove, Hawk). This game actually illustrates well the interaction effect between genetic predisposition and learning rule, because we find that it is the learning rule that has the biggest predisposition for aggression (big initial probability of choosing 'Hawk') that is evolutionarily stable. Besides, it is noteworthy that the behaviour of reinforcement learners is slightly more cooperative in the sense that pairs of this type could also learn to play the socially peaceful outcome (Dove, Dove). This does not give an advantage or a disadvantage compared to hypothetical reinforcement learners in this particular game, but it is interesting to note that trial-and-error learners are capable of achieving the equilibrium with the highest mutual payoff (Dove, Dove), where nobody is playing the aggressive 'Hawk' strategy, while hypothetical reinforcement learners cannot. The reason for this is that the equilibrium (Dove, Dove) provides a positive payoff (i.e. a reward) to both players in the Hawk–Dove game, which suffices to make it a stable equilibrium for the trial-and-error learning dynamics.

In the average Prisoner's Dilemma, we also observed that trial-and-error learners could reach the socially beneficial outcome where both partners cooperate (i.e. they 'solve' the dilemma), but this time this ability made trial-and-error learning evolutionarily stable. Indeed, pairs of trial-and-error learners are able to learn to cooperate together because the payoff for mutual cooperation is positive. On the other hand, hypothetical reinforcement learning always leads to defection (because if they cooperate, they have the regret of not having defected in the previous round). Interestingly, trial-and-error learners do not get exploited by hypothetical reinforcement learners as they succeed in learning to defect against them. The long-run behaviour of trial-and-error learning is actually reminiscent of that of the Tit-for-tat (TFT) strategy (Axelrod, 1980; Axelrod & Hamilton, 1981; Rapoport & Chammah, 1965): trial-and-error learning cooperates with itself a high proportion of the time (but not always) and defects against the defector hypothetical reinforcement learning. This result is striking when we realize that the underlying principle of trial-and-error is closer to win-stay, lose-shift (WSLS, Axelrod, 1980; Axelrod & Hamilton, 1981; Nowak & Sigmund, 1993; Rapoport & Chammah, 1965) than to TFT: a trial-and-error learner is essentially repeating actions followed by positive consequences and avoiding actions followed by negative consequences (in Appendix 7, we indeed show that impulsive trial-and-error learners derived from our learning process (equation (3)) with a memory of only one time step behave as WSLS).

The repeated Prisoner's Dilemma has been the topic of many studies aiming at understanding the evolution of cooperation, but most of the time no learning rules are used in evolutionary analysis, but qualitative strategies consisting of finite state automata such as TFT (or WSLS, Grim, etc.). Learning rules may actually represent a more appropriate way of conceptualizing animal behaviour because it describes several realistic features of animals, such as incremental adaptation, forgetting or, habituation. Moreover, learning rules provide a quantitative approach (in our case through the motivations for actions) that can potentially be linked to neuronal decision making (Dayan & Abbott, 2005; Enquist & Ghirlanda, 2005; Niv, 2009). It is thus interesting to see that a fairly simple learning rule based on the widely accepted principle of trial-and-error produces qualitatively similar behaviour as TFT, although trial-and-error is much less domain specific than TFT. It is noteworthy that this behaviour of trial-and-error learning results from its ignorance of foregone payoffs, not from a better ability to understand the specificities of repeated games. This suggests that, while trial-and-error learning performs well in the Prisoner's Dilemma, it is unlikely do so in general-domain learning tasks where the ability to form simple stimulus-action association is not sufficient.

Repeated Matching

In the repeated matching setting, individuals meet different partners at each interaction round. The learning task is here more complicated because an individual must adapt to an entire population composed of individuals with different learning rules which themselves adapt their behaviour. This type of matching model is widely used in evolutionary biology as a baseline model of interactions (Maynard Smith, 1982), and refers to cases where no particular assumption regarding population or social structure can be applied. This may be the case when interference competition for resources occurs periodically during a lifetime (Begon, Harper, & Townsend, 1996), but at random between a large number of population members during each period, as for instance during fishing by marine birds.

Here, we mainly observed results of type (3), i.e. hypothetical reinforcement learning generally dominated the population at an evolutionary equilibrium. The most representative example is the average Hawk–Dove game, where we found that hypothetical reinforcement learning was able to exploit the tendency of trial-and-error learners to play the 'Dove' action too often, especially when trial-and-error learners are at a high frequency in the population. Besides, we also observed in this repeated matching model an interaction between genetic predisposition for actions and learning rule in the average Prisoner's Dilemma. When individuals initially prefer 'Defection', we obtained a neutral situation where both learning rules learn full defection. However, when individuals had an initial preference for 'Cooperation', trial-and-error learning could lead to cooperation when at high frequency in the population, and get exploited by hypothetical reinforcement learning which always leads to defect.

Evolutionary Paths to Increased Cognition

Our primary goal in this paper was to provide some intuition as to when a learning rule located one step up above trial-and-error in the cognitive hierarchy can be favoured by selection in simple situations of social interactions. In comparative cognition (Shettleworth, 2009), trial-and-error learning is often the null hypothesis against which one tests hypotheses regarding more advanced cognition, and we adopted the same approach here by letting trial-and-error compete with hypothetical reinforcement learning. Because hypothetical reinforcement learning is related to the class of belief-based learning rules, which have been shown to lead to Nash equilibrium in the simple social interactions that we studied here (Hofbauer & Sandholm, 2002), one would expect hypothetical reinforcement learning to generally outcompete trial-and-error learning. Our findings do not support this view. Rather, they illustrate that learning rules using more information do not necessarily outcompete trial-and-error learning in social interactions, so the advantage of 'complex cognition' is not automatic and depends on the type of games played by the individuals in a population. Since Bayesian decision making (which is closely related to belief-based learning, i.e. our EHR rule) has been previously proposed to be likely to occur in animals facing individual decision problems (McNamara et al., 2006; Trimmer et al., 2011; Valone, 2006), our results suggest that behavioural rules that perform well for social interactions are not necessarily the same as those performing well under individual decisions, as long as individuals discount strongly the future when making decisions (see Molleman, van den Berg, & Weissing, 2014 for empirical evidence along these lines in humans). Hence, despite its apparent simplicity, trial-and-error or instrumental learning is not easy to displace, which may in part explain why it is ubiquitous across taxa (Staddon & Cerutti, 2003).

That a simple learning mechanism can be favoured in social interactions is not a surprising result when we look at previous work in evolutionary ecology (Arbilly et al., 2010, 2011; Hamblin & Giraldeau, 2009). However, the reasons why the simple mechanism of trial-and-error learning outcompeted the sophisticated hypothetical reinforcement learning are different from previous work. In Hamblin and Giraldeau (2009), the relative payoff sum rule (another implementation of trial-and-error learning) outcompeted a rule with perfect memory mainly because it had less inertia, which was most adapted to the changing environment considered by the authors. In our setting, trial-and-error and hypothetical reinforcement learning have the same level of inertia (same learning rate) and we generally compare the long-run behaviour of learners, ignoring transient behaviour. In Arbilly et al. (2010, 2011), an allele coding for simple learning can be genetically associated with an allele coding for a scrounger strategy in a producer–scrounger game, while complex learning is associated with a

producer strategy. Such genetic associations cannot happen in our setting because we studied evolution on only one locus, but note that in the Prisoner's Dilemma, there was a phenotypic association between trial-and-error learning and cooperation, which strongly influenced the result under both matching schemes. We also note that our results are also consistent with the findings of Josephson (2008). His model of matching corresponds to our repeated matching scheme, and even though we did not use exactly the same rules, we find like him that a learning rule with no observation of hypothetical payoffs (i.e. the trial-and-error learning rule) cannot be evolutionarily stable in this case. Hence, adding a more realistic feature such as environmental fluctuations does not help in favouring trial-and-error learning, suggesting that random interactions constitute a strong factor in favour of belief-based learning.

Our results also echo findings in the field of repeated games. For instance in Axelrod's tournament, the dominating strategy (Tit-for-tat) was not the most complex of all the proposed strategies (Axelrod, 1980). Further, research in economics also indicate that repeated games and social interactions do not always select for more complex strategies, and these results were obtained either by taking into account explicitly the costs of cognitive complexity (Binmore & Samuelson, 1992) or, more recently, in the absence of such costs (Duersch, Oechssler, & Schipper, 2014; Horváth, Kovářík, & Mengel, 2012; Mohlin, 2012). All this suggests that it is relevant to try to characterize the type of games species are playing in nature (or the social problems they face), in order to characterize the real demands on social cognition.

It might be difficult to assess the payoff structure of the games played by individuals, but our results show that when individuals cannot condition their actions on the type of games they are playing (as we assumed here), the behavioural outcomes only depend on the average game, which can thus be used to produce predictions about the psychological capacities of that species. Indeed, we only had to analyse the behaviour in the average game (Table 4) in order to predict equilibrium behaviour for the learning dynamics, and learning behaviour in the average game explained well the results of evolutionary simulations. These complications regarding the fluctuating aspects of the games can be ignored in potential experimental tests of our model because our results showing that trial-and-error and hypothetical reinforcement learning produce different behaviours in social interactions still hold if we interpret the average game as only a constant, fixed game. Accordingly, in order to test which of these learning rules is used by a particular species, we think that experiments in which individuals socially interact according to a Prisoner's Dilemma are the best suited because our two learning rules lead to distinct outcomes in this game.

By contrast, a coordination problem is not a good situation to perform these tests because it is solved by both trial-and-error and hypothetical reinforcement learning and consequently does not allow one to distinguish between these rules. In previous empirical studies, Prisoner's Dilemma games have been implemented to test the cooperativeness of species (e.g. Schneeberger, Dietz, & Taborsky, 2012) by giving individuals the choice between selfish and social options, e.g. when deciding whether or not to make an effort to provide food to a conspecific. Repeating such tasks multiple times between pairs of individuals can determine whether only simple trial and error is used, in which case one would expect to see a proportion of pairs learning to exchange food in this task, by reinforcement of actual reward. On the other hand, observing only defection would suggest that the species of interest is using a rule other than trial-and-error learning, and the use of hypothetical reinforcement learning could be further tested using classical tests of Bayesian learning in an individual decision task. It must be noted that in such empirical tests, one does not need to consider that the animal is able to condition its behaviour on the type of game it is facing. Being able to condition behaviour on the type of game requires the individual to detect features of the environment and combine them correctly in order to compute game payoffs accordingly, which is cognitively more demanding and requires more inference than we have assumed in this paper. Here, we only assumed that individuals had access to limited information on opportunities they did not try. This information can be reached through explicit reasoning (e.g. noticing that a food patch has been exploited by others is a cue that nearby patches have probably also been exploited) or via social observation.

In this work, we assumed that the information obtained by hypothetical reinforcement learners was always correct. In other words, nonrealized payoffs are always perfectly observed by HR, but this assumption does not always hold in nature. Social observation of nonrealized payoffs is dependent on the frequency of actions taken in the population, so an individual is generally unable to evaluate socially the relative advantages of all possible behaviours, i.e. if nobody in the population takes an optimal action, a social learner will be unable to act optimally. Moreover, social observation is error-prone (e.g. when an individual is trying to assess how much food a conspecific has) and reasoning is even more so. Future research could produce different results by modelling a situation in which the information acquisition process involves errors, e.g. when the genotype g in equation (3) can take values lower than one (payoffs are underestimated) or higher than one (overestimated payoffs), or even when g is a random variable whose support, mean and variance are genetically determined (representing a situation where payoffs are sometimes under- and sometimes overestimated).

Our paper also contributes to modelling in behavioural ecology, where researchers acknowledge the need to describe animal behaviour in terms of general rules that are used to face several decision problems an individual may face (Dijker, 2011; Fawcett, Hamblin, & Giraldeau, 2013; Hammerstein and Stevens, 2012; McNamara & Houston, 2009). By modelling an environment in which individuals face different social games and partners, but use the same behavioural rule (either trial-and-error or hypothetical reinforcement learning), a learning rule can only work well on average. We concentrated here on social behaviours only under the simplest games and mainly studied long-term behavioural dynamics, but this approach can also be applied for studying rules of behaviour that can serve under a variety of ecological contexts, under more complex social structures, or to investigate the effects of learning speed. This may help to better delineate the possible evolutionary paths from simple to more sophisticated decision-making processes.

We thank the reviewers for extensive comments on the manuscript. This work was supported by Swiss NSF grant PP00P3-123344.

References

Achbany, Y., Fouss, F., Yen, L., Pirotte, A., & Saerens, M. (2006). Optimal tuning of continual online exploration in reinforcement learning. In S. Kollias, A. Stafylopatis, W. Duch, & E. Oja (Eds.), *Artificial neural networks – ICANN 2006, volume 4131 of lecture notes in computer science* (pp. 790–800). Berlin, Germany: Springer.

Amano, T., Ushiyama, K., Moriguchi, S., Fujita, G., & Higuchi, H. (2006). Decision-making in group foragers with incomplete information: test of individual-based model in geese. *Ecological Monographs, 76*(4), 601–616.

Anderson, S. P., de Palma, A., & Thisse, J.-F. (1992). *Discrete choice theory of product differentiation*. Cambridge, MA: MIT Press.

Arbilly, M., Motro, U., Feldman, M. W., & Lotem, A. (2010). Co-evolution of learning complexity and social foraging strategies. *Journal of Theoretical Biology, 267*(4), 573–581.

Arbilly, M., Motro, U., Feldman, M. W., & Lotem, A. (2011). Recombination and the evolution of coordinated phenotypic expression in a frequency-dependent game. *Theoretical Population Biology, 80*(4), 244–255.

Arnold, S. J. (1978). The evolution of a special class of modifiable behaviors in relation to environmental pattern. *The American Naturalist, 112*(984), 415–427.

Axelrod, R. (1980). Effective choice in the prisoner's dilemma. *The Journal of Conflict Resolution, 24*(1), 3–25.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*, 1390–1396.

Begon, M., Harper, J. L., & Townsend, C. R. (1996). *Ecology: Individuals, populations and communities* (6th ed.). Boston, MA: Blackwell Science.

Benaim, M. (1999). Dynamics of stochastic approximation algorithms. In *Séminaire de Probabilités XXXIII* (vol. 1709, pp. 1–68).

Benveniste, A., Metivier, M., & Priouret, P. (1991). *Adaptive algorithms and stochastic approximations*. Berlin, Germany: Springer-Verlag.

Bernstein, C., Kacelnik, A., & Krebs, J. R. (1988). Individual decisions and the distribution of predators in a patchy environment. *Journal of Animal Ecology, 57*(3), 1007–1026.

Binmore, K. G., & Samuelson, L. (1992). Evolutionary stability in repeated games played by finite automata. *Journal of Economic Theory, 57*, 278–305.

Black, J. M. (1996). *Partnerships in birds: The study of monogamy*. New York, NY: Oxford University Press.

Borenstein, E., Feldman, M. W., & Aoki, K. (2008). Evolution of learning in fluctuating environments: when selection favors both social and exploratory individual learning. *Evolution, 62*(3), 586–602.

Borkar, V. S. (2008). *Stochastic approximation: A dynamical systems viewpoint*. Cambridge, U.K.: Cambridge University Press.

Boyd, R., & Richerson, P. J. (1988). *Culture and the evolutionary process*. Chicago, IL: University of Chicago Press.

Brown, G. W. (1951). Iterative solution of games by fictitious play. In T. Koopmans (Ed.), *Activity analysis of production and allocation* (pp. 374–376). New York, NY: Wiley.

Bush, R. R., & Mostellller, F. (1951). A mathematical model for simple learning. *Psychological Review, 58*(5), 313–323.

Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

Camerer, C., & Ho, T. H. (1999). Experienced-weighted attraction learning in normal form games. *Econometrica, 67*(4), 827–874.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology, 9*(2), 129–136.

Chmura, T., Goerg, S. J., & Selten, R. (2012). Learning in experimental games. *Games and Economic Behavior, 76*(1), 44–73.

Dayan, P., & Abbott, L. F. (2005). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.

Dickinson, A. (1980). *Contemporary animal learning theory*. Cambridge, U.K.: Cambridge University Press.

Dijker, A. (2011). Physical constraints on the evolution of cooperation. *Evolutionary Biology, 38*(2), 124–143.

Dridi, S., & Lehmann, L. (2014). On learning dynamics underlying the evolution of learning rules. *Theoretical Population Biology, 91*, 20–36.

Dubois, F., Morand-Ferron, J., & Giraldeau, L.-A. (2010). Learning in a game context: strategy choice by some keeps learning from evolving in others. *Proceedings of the Royal Society B: Biological Sciences, 277*(1700), 3609–3616.

Duersch, P., Oechssler, J., & Schipper, B. C. (2014). When is tit-for-tat unbeatable? *International Journal of Game Theory, 43*(1), 25–36.

Dugatkin, L. A. (2010). *Principles of animal behavior* (2nd ed.). New York, NY: WW Norton.

Dugatkin, L. A., & Reeve, H. K. (2000). *Game theory and animal behavior*. New York, NY: Oxford University Press.

Dunlap, A. S., & Stephens, D. W. (2009). Components of change in the evolution of learning and unlearned preference. *Proceedings of the Royal Society B: Biological Sciences, 276*(1670), 3201–3208.

Emery, N. J., & Clayton, N. S. (2004). The mentality of crows: convergent evolution of intelligence in corvids and apes. *Science, 306*(5703), 1903–1907.

Emery, N. J., & Clayton, N. S. (2009). Comparative social cognition. *Annual Review of Psychology, 60*(1), 87–113.

Enquist, M. E., & Ghirlanda, S. (2005). *Neural networks and animal behavior*. Princeton, NJ: Princeton University Press.

Erev, I., & Roth, A. E. (1998). Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review, 88*(4), 848–881.

Fawcett, T. W., Hamblin, S., & Giraldeau, L.-A. (2013). Exposing the behavioral gambit: the evolution of learning and decision rules. *Behavioral Ecology, 24*(1), 2–11.

Feldman, M., Aoki, K., & Kumm, J. (1996). Individual versus social learning: evolutionary analysis in a fluctuating environment. *Anthropological Science, 104*, 209–232.

Feltovich, N. (2000). Reinforcement-based versus belief-based learning models in experimental asymmetric-information games. *Econometrica, 68*(3), 605–641.

Fudenberg, D., & Levine, D. K. (1998). *The theory of learning in games*. Cambridge, MA: MIT Press.

van Gils, J. A., Schenk, I. W., Bos, O., & Piersma, T. (2003). Incompletely informed shorebirds that face a digestive constraint maximize net energy gain when exploiting patches. *The American Naturalist, 161*(5), 777–793.

Hamblin, S., & Giraldeau, L.-A. (2009). Finding the evolutionarily stable learning rule for frequency-dependent foraging. *Animal Behaviour, 78*(6), 1343–1350.

Hammerstein, P., & Stevens, J. R. (Eds.). (2012). *Evolution and the mechanisms of decision making*. Cambridge, MA: MIT Press.

Harley, C. B. (1981). Learning the evolutionarily stable strategy. *Journal of Theoretical Biology, 89*(4), 611–633.

Heller, D. (2004). An evolutionary approach to learning in a changing environment. *Journal of Economic Theory, 114*(1), 31–55.

Hirsch, M. W., Smale, S., & Devaney, R. L. (2004). *Differential equations, dynamical systems, and an introduction to chaos*. San Diego, CA: Academic Press.

Ho, T. H., Camerer, C. F., & Chong, J.-K. (2007). Self-tuning experience weighted attraction learning in games. *Journal of Economic Theory, 133*(1), 177–198.

Hofbauer, J., & Sandholm, W. H. (2002). On the global convergence of stochastic fictitious play. *Econometrica, 70*(6), 2265–2294.

Hofbauer, J., & Sigmund, K. (1998). *Evolutionary games and population dynamics*. Cambridge, U.K.: Cambridge University Press.

Hopkins, E. (2002). Two competing models of how people learn in games. *Econometrica, 70*(6), 2141–2166.

Horváth, G., Kovárík, J., & Mengel, F. (2012). Limited memory can be beneficial for the evolution of cooperation. *Journal of Theoretical Biology, 300*, 193–205.

Izquierdo, L. R., Izquierdo, S. S., Gotts, N. M., & Polhill, J. G. (2007). Transient and asymptotic dynamics of reinforcement learning in games. *Games and Economic Behavior, 61*(2), 259–276.

Johnston, T. D. (1982). Selective costs and benefits in the evolution of learning. In S. Jay, & R. A. H. Rosenblatt (Eds.), *Advances in the study of behavior* (Vol. 12, pp. 65–106). San Diego, CA: Academic Press.

Josephson, J. (2008). A numerical analysis of the evolutionary stability of learning rules. *Journal of Economic Dynamics and Control, 32*(5), 1569–1599.

Karlin, S., & Taylor, H. E. (1975). *A first course in stochastic processes*. San Diego, CA: Academic Press.

Katsnelson, E., Motro, U., Feldman, M. W., & Lotem, A. (2011). Evolution of learned strategy choice in a frequency-dependent game. *Proceedings of the Royal Society B: Biological Sciences*. rspb20111734.

Kempe, M., & Mesoudi, A. (2014). Experimental and theoretical models of human cultural evolution. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(3), 317–326.

Kendal, J., Giraldeau, L.-A., & Laland, K. (2009). The evolution of social learning rules: payoff-biased and frequency-dependent biased transmission. *Journal of Theoretical Biology, 260*(2), 210–219.

Kleiman, D. G. (1977). Monogamy in mammals. *The Quarterly Review of Biology, 52*(1), 39–69.

Krebs, J. R., Davies, N. B., & West, S. A. (1993). *An introduction to behavioural ecology* (3rd ed.). Chichester, U.K.: Wiley-Blackwell.

Laland, K. N. (2004). Social learning strategies. *Learning & Behavior, 32*(1), 4–14.

Leslie, D. S., & Collins, E. J. (2005). Individual Q-learning in normal form games. *SIAM Journal on Control and Optimization, 44*(2), 495–514.

Lima, S. L. (1984). Downy woodpecker foraging behavior: efficient sampling in simple stochastic environments. *Ecology, 65*(1), 166–174.

Luttbeg, B., & Warner, R. R. (1999). Reproductive decision-making by female peacock wrasses: flexible versus fixed behavioral rules in variable environments. *Behavioral Ecology, 10*(6), 666–674.

Macy, M. W., & Flache, A. (2002). Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences of the United States of America, 99*(90003), 7229–7236.

Maynard Smith, J. (1982). *Evolution and the theory of games*. Cambridge, U.K.: Cambridge University Press.

McElreath, R., & Boyd, R. (2007). *Mathematical models of social evolution: A guide for the perplexed*. Chicago, IL: University of Chicago Press.

McKelvey, R. D., & Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior, 10*(1), 6–38.

McNamara, J. M., Green, R. F., & Olsson, O. (2006). Bayes' theorem and its applications in animal behaviour. *Oikos, 112*(2), 243–251.

McNamara, J. M., & Houston, A. I. (1985). Optimal foraging and learning. *Journal of Theoretical Biology, 117*(2), 231–249.

McNamara, J. M., & Houston, A. I. (1987). Memory and the efficient use of information. *Journal of Theoretical Biology, 125*(4), 385–395.

McNamara, J. M., & Houston, A. I. (2009). Integrating function and mechanism. *Trends in Ecology & Evolution, 24*(12), 670–675.

Mery, F., & Kawecki, T. J. (2002). Experimental evolution of learning ability in fruit flies. *Proceedings of the National Academy of Sciences of the United States of America, 99*(22), 14274–14279.

Mohlin, E. (2012). Evolution of theories of mind. *Games and Economic Behavior, 75*(1), 299–318.

Molleman, L., van den Berg, P., & Weissing, F. J. (2014). Consistent individual differences in human social learning strategies. *Nature Communications*, 5.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology, 53*(3), 139–154.

Nowak, M. A., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature, 364*, 56–58.

Pemantle, R. (1990). Nonconvergence to unstable points in urn models and stochastic approximations. *The Annals of Probability, 18*(2), 698–712.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(04), 515–526.

Rapoport, A., & Chammah, A. M. (1965). *Prisoner's dilemma*. Ann Arbor, MI: University of Michigan Press.

Rendell, L., Boyd, R., Cownden, D., Enquist, M., Eriksson, K., Feldman, M. W., et al. (2010). Why copy others? insights from the social learning strategies tournament. *Science, 328*(5975), 208–213.
Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.
Rogers, A. R. (1988). Does biology constrain culture? *American Anthropologist, 90*(4), 819–831.
Schlag, K. H. (1998). Why imitate, and if so, how? *Journal of Economic Theory, 78*, 130–156.
Schloegl, C., Dierks, A., Gajdon, G. K., Huber, L., Kotrschal, K., & Bugnyar, T. (2009). What you see is what you get? exclusion performances in ravens and keas. *PLoS One, 4*(8), e6368.
Schneeberger, K., Dietz, M., & Taborsky, M. (2012). Reciprocal cooperation between unrelated rats depends on cost to donor and benefit to recipient. *BMC Evolutionary Biology, 12*(1), 41.
Shettleworth, S. J. (2009). *Cognition, evolution, and behavior*. New York, NY: Oxford University Press.
Shettleworth, S. J., Krebs, J. R., Stephens, D. W., & Gibbon, J. (1988). Tracking a fluctuating environment: a study of sampling. *Animal Behaviour, 36*(1), 87–105.
Staddon, J. E. R., & Cerutti, D. T. (2003). Operant conditioning. *Annual Review of Psychology, 54*(1), 115–144.
Stephens, D. W. (1991). Change, regularity, and value in the evolution of animal learning. *Behavioral Ecology, 2*(1), 77–89.
Stephens, D. W., & Clements, K. C. (1998). Game theory and learning. In L. A. Dugatkin, & H. K. Reeve (Eds.), *Game theory and animal behavior* (pp. 239–260). New York, NY: Oxford University Press.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
Taylor, A. H., Miller, R., & Gray, R. D. (2012). New Caledonian crows reason about hidden causal agents. *Proceedings of the National Academy of Sciences of the United States of America, 109*(40), 16389–16391.
Thorndike, E. L. (1911). *Animal intelligence*. Darien, CT: Hafner.
Trimmer, P. C., Houston, A. I., Marshall, J. A. R., Mendl, M. T., Paul, E. S., & McNamara, J. M. (2011). Decision-making under uncertainty: biases and bayesians. *Animal Cognition, 14*(4), 465–476.
Valone, T. J. (2006). Are animals capable of bayesian updating? An empirical review. *Oikos, 112*(2), 252–259.
Wakano, J. Y., Aoki, K., & Feldman, M. W. (2004). Evolution of social learning: a mathematical analysis. *Theoretical Population Biology, 66*(3), 249–258.
Weibull, J. W. (1997). *Evolutionary game theory*. Cambridge, MA: MIT Press.

APPENDIX 1. TRIAL-AND-ERROR AND HYPOTHETICAL REINFORCEMENT LEARNING

In this appendix, we give the stochastic recursions for the motivations for the four learning rules studied in this paper. We have defined two forms of trial-and-error learning, PTR and ETR, and two forms of hypothetical reinforcement learning, PHR and EHR.

Trial-and-Error

Dynamic learning rate

Substituting $g_i = 0$, $\phi_{i,t} = (1/t) + 1$ into equation (3), we obtain PTR with the updating rule

$$M_{i,t+1}(a) = M_{i,t}(a) + \frac{1}{t+1} 1(a, a_{i,t}) \pi_i(a, \mathbf{a}_{-i,t}, \omega_t). \tag{A1.1}$$

Every new experienced payoff is thus divided by the total number of interactions and added to the previous motivation. In the long run, the effect of new payoffs on motivations goes to zero. Note that when action a is not played, the motivation is not updated. Moreover, the learner does not forget information from the past. It is the payoffs obtained in the first rounds of interaction that have the biggest effect on the motivations at time t.

Constant learning rate

Substituting $g_i = 0$ and $\phi_{i,t} = 1$ into equation (3) gives ETR, which has the updating rule

$$M_{i,t+1}(a) = \frac{t}{t+1} M_{i,t}(a) + \frac{1}{t+1} 1(a, a_{i,t}) \pi_i(a, \mathbf{a}_{-i,t}, \omega_t). \tag{A1.2}$$

This rule looks like a time average of the payoffs obtained when playing action a but it is actually a biased average. Indeed, in a nonbiased average, the motivation of action a should not be updated when action a is not played. However, here when action a is not played at time t, the motivation is still updated but it is as if the payoff obtained for action a at time t was zero. Hence, depending on the signs of the payoffs in the game, the nonplayed actions have a tendency to lose weight (e.g. when all payoffs in the game are positive) or gain weight (e.g. when all payoffs in the game are negative).

While equation (A1.2) may seem unfamiliar to students of animal learning, especially because of the time average principle, one can translate this equation into the well-known form

$$M_{i,t+1}(a) = \begin{cases} (1-\gamma_t) M_{i,t}(a) + \gamma_t \pi_i(a, \mathbf{a}_{-i,t}, \omega_t), & \text{if } a = a_{i,t}, \\ (1-\gamma_t) M_{i,t}(a), & \text{otherwise,} \end{cases} \tag{A1.3}$$

where $\gamma_t = 1/(t+1)$ can be seen as a learning rate or discount factor. This type of equation is common in the literature on animal learning and is often termed the linear operator model (Amano et al., 2006; Bernstein et al., 1988; Bush & Mostelller, 1951; Hamblin & Giraldeau, 2009; McNamara & Houston, 1987; Rescorla & Wagner, 1972; Stephens & Clements, 1998). However in general the learning rate γ is considered to be a constant, and not a function of time as is the case here. The effect of having a dynamic rather than constant learning rate is discussed in Sutton and Barto (1998). In our model we required that $\gamma_t = 1/(t+1)$, but our results hold as long as γ_t is a decreasing step size (in the sense of stochastic approximation theory, Benaim, 1999; Dridi & Lehmann, 2014).

Hypothetical Reinforcement Learning

Dynamic learning rate

Substituting $g_i = 1$ and $\phi_{i,t} = (1/t) + 1$ into equation (3) yields the PHR updating rule

$$M_{i,t+1}(a) = M_{i,t}(a) + \frac{1}{t+1} \pi_i(a, \mathbf{a}_{-i,t}, \omega_t). \tag{A1.4}$$

This equation is the counterpart of equation (A1.1) but without the factor $1(a, a_{i,t})$ in front of the payoffs. The dynamics will essentially obey the same principles as equation (A1.1), with early payoffs (small t) having a bigger effect than late payoffs (big t).

Constant learning rate

Substituting $g_i = 1$ and $\phi_{i,t} = 1$ into equation (3) gives EHR, which is the standard belief-based learning rule (Camerer & Ho, 1999; Fudenberg & Levine, 1998), with motivation updating given by

$$M_{i,t+1}(a) = \frac{t}{t+1} M_{i,t}(a) + \frac{1}{t+1} \pi_i(a, \mathbf{a}_{-i,t}, \omega_t). \tag{A1.5}$$

For every action a, this equation represents the average payoff that a player would have obtained if he were constantly playing action a, given the history of his opponent's actions and environmental states $\{\mathbf{a}_{-i,\tau}, \omega_\tau\}_{\tau=1}^{t}$.

Belief-based learning

We now show that equation (A1.5) can also be interpreted in terms of updating average payoffs given beliefs over the action play probabilities of partners as in Camerer and Ho (1999). Note that belief-based learning requires observing at the same time the

action of the opponent and the entire payoff matrix of the game after each interaction round, which relies on much more information than just knowing the payoffs of unchosen actions as in equation (A1.5). In this section, we also show that extending belief-based learning to a situation of environmental fluctuations is not trivial. In particular, the individual cannot just ignore the effect of the environment on its payoffs. Indeed, whereas the individual cannot observe environmental states ω_t, we will see that it must realize that another factor than just the behaviour of its opponent is influencing its payoff. The individual must average out the effect of this additional factor explicitly, as is detailed below.

For ease of presentation, but without loss of generality, we consider that individual i interacts only with one other individual in the population (one-shot matching model), which plays action $a_{-i,t} \in \mathscr{A}$ at time t. We then write for $t \geq 1$

$$\mathscr{B}_{-i,t+1}(a) = \frac{t}{t+1}\mathscr{B}_{-i,t}(a) + \frac{1}{t+1}1(a, a_{-i,t}), \tag{A1.6}$$

where $\mathscr{B}_{-i,t+1}(a) \in [0,1]$ is the frequency of times the partner of individual i has played action a up to time t, and which is the belief of individual i that its partner plays a at $t+1$ given the initial belief $\mathscr{B}_{-i,1}$ at $t=1$. We also define

$$\tilde{\pi}_{i,t+1}(a,k) = \frac{\sum_{\tau=1}^{t} 1(k, a_{-i,\tau})\pi_i(a,k,\omega_\tau)}{n_{-i,t+1}(k)}, \tag{A1.7}$$

to be the average payoff at time $t+1$ individual i would have obtained if he constantly played action a on all the interaction rounds where his opponent played k. This allows the individual to average out the effect of environmental fluctuations on its payoffs. In equation (A1.7), $n_{-i,t+1}(k) = (t+1)\mathscr{B}_{-i,t+1}(k)$ is the number of times the opponent of i played action k up to time t. In the following, it will be useful to write equation (A1.7) as the recursion

$$\begin{aligned}\tilde{\pi}_{i,t+1}(a,k) = {} & \frac{t\mathscr{B}_{-i,t}(k)}{(t+1)\mathscr{B}_{-i,t+1}(k)}\tilde{\pi}_{i,t}(a,k) \\ & + \frac{1}{(t+1)\mathscr{B}_{-i,t+1}(k)}1(k, a_{-i,t})\pi_i(a,k,\omega_t).\end{aligned} \tag{A1.8}$$

Let us now write

$$M_{i,t+1}(a) = \sum_{k\in\mathscr{A}} \tilde{\pi}_{i,t+1}(a,k)\mathscr{B}_{-i,t+1}(k), \tag{A1.9}$$

which can be interpreted as the expected payoff to individual i given its beliefs over the action distribution of its partner and given the history of environmental states. Substituting equation (A1.6) and equation (A1.8) into equation (A1.9) shows that the motivation dynamics still satisfy equation (A1.5) for the case of two players. Hence, when a hypothetical reinforcement learner expresses action by using the logit choice rule (equation (2)), it behaves as if it tries to maximize its expected current reward given its beliefs.

APPENDIX 2. STOCHASTIC APPROXIMATION

Here, we show the main steps to derive equations (5) and (6) from equations (2) and (3). First, an application of stochastic approximation theory (e.g. Benaim, 1999) shows that equations (2) and (3) can be approximated by the set of differential equations

$$\begin{aligned}\dot{p}_i(a) = {} & p_i(a)\left[\varepsilon_i \sum_{k\in\mathscr{A}} \log\left(\frac{p_i(k)}{p_i(a)}\right)p_i(k)\right. \\ & \left. + \lambda\left(\overline{R}_i(a) - \sum_{k\in\mathscr{A}} \overline{R}_i(k)p_i(k)\right)\right],\end{aligned} \tag{A2.1}$$

where

$$\overline{R}_i(a) = [p_i(a) + g_i(1-p_i(a))] \sum_{\mathbf{a}_{-i}\in\mathscr{A}^{N-1}} p_{-i}(\mathbf{a}_{-i})\overline{\pi}_i(a, \mathbf{a}_{-i}) \tag{A2.2}$$

and

$$\overline{\pi}_i(a, \mathbf{a}_{-i}) = \sum_{\omega\in\Omega} \mu(\omega)\pi_i(a, \mathbf{a}_{-i}, \omega) \tag{A2.3}$$

(Dridi & Lehmann, 2014, equations (11)–(13)). Here, $\overline{R}_{i,t}(a, \mathbf{M}_t)$ is the expectation of the reinforcement to the motivation of action a, i.e. the expectation of the numerator of the second term of equation (3) over the distribution of environmental states and the distribution of choice probabilities, $p_{-i}(\mathbf{a}_{-i})$, is the probability of the joint action profile of individuals other than i, and $\overline{\pi}_i(a, \mathbf{a}_{-i})$ represents the payoff of the average game in which individual i is involved.

In our context, the parameter ε_i in equation (A2.1) takes the value $\varepsilon_i = 1 + t(1-\phi_{i,t})$ (Dridi & Lehmann, 2014, equation (8) with $n_{i,t} = t$ and $\rho_i = 1$). Since we assumed that $\phi_{i,t} = (1/t) + 1$, this gives $\varepsilon_i = 0$ and the exploration term (the first term in square brackets in equation (A2.1)) cancels. Moreover, we are interested in 2×2 games so there are only two actions ($\mathscr{A} = \{1,2\}$) and two players. Let the two players be denoted i and j, p_i the probability that individual i takes action 1, and p_j the probability that individual j takes action 1. With this, we can write the differential equation for the probability that individual i takes action 1 using equation (A2.1) as

$$\dot{p}_i = p_i\lambda(\overline{R}_i(1) - [\overline{R}_i(1)p_i + \overline{R}_i(2)(1-p_i)]), \tag{A2.4}$$

where

$$\overline{R}_i(1) = [p_i + g_i(1-p_i)] \sum_{\mathbf{a}_{-i}\in\mathscr{A}^{N-1}} p_{-i}(\mathbf{a}_{-i})\overline{\pi}_i(1, \mathbf{a}_{-i}) \tag{A2.5}$$

$$\overline{R}_i(2) = [(1-p_i) + g_i p_i] \sum_{\mathbf{a}_{-i}\in\mathscr{A}^{N-1}} p_{-i}(\mathbf{a}_{-i})\overline{\pi}_i(2, \mathbf{a}_{-i}). \tag{A2.6}$$

Further, in the 2×2 games that we study here (one-shot matching model), the single opponent of individual i is individual j so $\mathbf{a}_{-i} \in \{1,2\}$ and $p_{-i}(1) = p_j$. Replacing these in equation (A2.5), we can write equation (A2.4) as

$$\begin{aligned}\dot{p}_i = {} & p_i(1-p_i)\lambda\left[\left\{p_j\overline{\pi}_i(1,1) + \left(1-p_j\right)\overline{\pi}_i(1,2)\right\}\{p_i + g_i(1-p_i)\}\right. \\ & \left. - \left\{p_j\overline{\pi}_i(2,1) + \left(1-p_j\right)\overline{\pi}_i(2,2)\right\}\{g_i p_i + (1-p_i)\}\right].\end{aligned} \tag{A2.7}$$

Using the definition of the payoffs of the average game in Table 4, we have $\overline{\pi}_i(1,1) = \mathscr{R}, \overline{\pi}_i(1,2) = \mathscr{S}, \overline{\pi}_i(2,1) = \mathscr{T}, \overline{\pi}_i(2,2) = \mathscr{P}$, to equation (2.7) yields

$$\begin{aligned}\dot{p}_i = {} & p_i(1-p_i)\lambda\left[\left\{p_j\mathscr{R} + \left(1-p_j\right)\mathscr{S}\right\}\{p_i + g_i(1-p_i)\}\right. \\ & \left. - \left\{p_j\mathscr{T} + \left(1-p_j\right)\mathscr{P}\right\}\{g_i p_i + (1-p_i)\}\right].\end{aligned} \tag{A2.8}$$

For player j, the differential equation for its probability to take

action 1 is the exact symmetric (because i is the single opponent of j), whereby

$$\begin{aligned}\dot{p}_j =& p_j\left(1-p_j\right)\lambda\Big[\{p_i\mathscr{R}+(1-p_i)\mathscr{S}\}\left\{p_j+g_j\left(1-p_j\right)\right\}\\ &-\{p_i\mathscr{T}+(1-p_i)\mathscr{P}\}\left\{g_jp_j+\left(1-p_j\right)\right\}\Big].\end{aligned} \tag{A2.9}$$

Perturbed versus Unperturbed Learning Dynamics

In this section we discuss how the case with dynamic learning rate approximates the case with constant learning rate. When the learning rate is dynamic ($\phi_{i,t}=(1/t)+1$), the exploration term in equation (A2.1) disappears (because in this case $\varepsilon_i=0$) and we call this the unperturbed dynamics. When the learning rate is constant $\phi_i=1$, the exploration term is weighted by $\varepsilon_i=1$, and we call this case the perturbed dynamics. The role of the exploration term is essentially to move the dynamics from pure states (i.e. the corners of the state space) in the perturbed dynamics. However, when λ is very high the second term in brackets of equation (A2.1) dominates the exploration term so that the equilibria of the perturbed dynamics approach the equilibria of the unperturbed dynamics (see Dridi & Lehmann, 2014 for an illustration of this statement).

APPENDIX 3. FECUNDITY AT BEHAVIOURAL EQUILIBRIUM

In this section, we derive an expression for fecundity (equation (7)) under the assumption that the learning process (equations (5) and (6)) has reached an equilibrium during the individual's lifetime. Indeed, if the individuals interact for a long enough time, the action probabilities $p_{i,t}(a)$ may reach an equilibrium for all i and a, and the fecundity of player i will be its average payoff at equilibrium. Then, the fecundity of individual i is

$$\Pi_i=\sum_{\omega\in\Omega}\sum_{a\in\mathscr{A}}\sum_{\mathbf{a}_{-i}\in\mathscr{A}^{N-1}}\mu(\omega)\widehat{p}_i(a)\widehat{p}_{-i}(\mathbf{a}_{-i})\pi_i(a,\mathbf{a}_{-i},\omega), \tag{A3.1}$$

where $\widehat{p}_i(a)$ denotes the equilibrium probability with which individual i chooses action a, while $\widehat{p}_{-i}(\mathbf{a}_{-i})$ is the equilibrium probability with which the opponents of individual i choose action profile $\mathbf{a}_{-i}$. This equilibrium is obtained by setting $\dot{p}_j(a)=0$ in equation (A2.1) for all j and a.

Equation (A3.1) should be understood as a long-run average payoff in the game taken over three distributions. The first distribution gives the probability $\mu(\omega)$ of playing game ω; the second distribution gives the equilibrium probability $\widehat{p}_i(a)$ that player i takes action a; the third distribution gives the probability $\widehat{p}_{-i}(\mathbf{a}_{-i})$ that the opponents of individual i take action profile $\mathbf{a}_{-i}$. The distribution $\mu(\omega)$ is already provided as a parameter of the model. The other tow distributions have to be computed by studying the equilibria of the choice probabilities $\widehat{p}_i(a)$ for all i in the population. Using equation (A2.3), the average payoff can be simplified to

$$\Pi_i=\sum_{a\in\mathscr{A}}\sum_{\mathbf{a}_{-i}\in\mathscr{A}^{N-1}}\widehat{p}_i(a)\widehat{p}_{-i}(\mathbf{a}_{-i})\overline{\pi}_i(a,\mathbf{a}_{-i}). \tag{A3.2}$$

Since we are concerned with 2×2 games, and since the single opponent of individual i is individual j (equations (5) and (6)), we have $a\in\{1,2\}$ and $\mathbf{a}_{-i}\in\{1,2\}$. If we further call $\widehat{p}_i$ the probability that individual i plays action 1 at a behavioural equilibrium and $\widehat{p}_j$ the corresponding probability for individual j, equation (A3.2) can be developed as

$$\begin{aligned}\Pi_i=&\widehat{p}_i\widehat{p}_j\overline{\pi}_i(1,1)+\widehat{p}_i\left(1-\widehat{p}_j\right)\overline{\pi}_i(1,2)+(1-\widehat{p}_i)\widehat{p}_j\overline{\pi}_i(2,1)\\ &+(1-\widehat{p}_i)\left(1-\widehat{p}_j\right)\overline{\pi}_i(2,2).\end{aligned} \tag{A3.3}$$

Factoring out and replacing the average payoffs, $\overline{\pi}_i(\bullet,\bullet)$, by their values in Table 4, we finally obtain

$$\Pi_i=\widehat{p}_i\left(\widehat{p}_j\mathscr{R}+\left(1-\widehat{p}_j\right)\mathscr{S}\right)+(1-\widehat{p}_i)\left(\widehat{p}_j\mathscr{T}+\left(1-\widehat{p}_j\right)\mathscr{P}\right), \tag{A3.4}$$

where in the main text we used $\Pi_{ij}=\Pi_i$ in order to emphasize that in the one-shot matching model, the payoff of individual i depends only on its single opponent (j).

APPENDIX 4. QUALITATIVE ANALYSIS FOR THE ONE-SHOT MATCHING MODEL

Here, we carry out the stability analysis of the equilibrium points of the learning dynamics presented in the main text (equations (5) and (6)). Before starting the analysis, let us make a technical remark. The dynamical systems we will analyse can display hyperbolic equilibria that admits stable manifolds (with one positive and one negative eigenvalue). We will completely discard the possibility that an initial condition is on a stable manifold because doing so leads to a locally unstable equilibrium, which is not robust to small perturbations under the original stochastic process (Pemantle, 1990). Equilibria and eigenvalues were calculated using the Mathematica software. The payoffs of the average games are defined in Table 4.

Prisoner's Dilemma

PTR versus PTR. The dynamical system obtained by setting $g_i=0$, $g_j=0$, $\mathscr{R}=B-C$, $\mathscr{S}=-C$, $\mathscr{T}=B$, and $\mathscr{P}=0$ into equations (5) and (6) admits seven equilibria (Table 2). Four are at the corners of the state space, two are on the edges and one is completely interior. The first equilibrium on the edge is situated on the line $p_i=1$ and the other one is symmetric with respect to the line $p_j=p_i$. The interior equilibrium is $\left(\widehat{p}_i=\frac{B+C}{2B},\widehat{p}_j=\frac{B+C}{2B}\right)$.

Evaluating the Jacobian matrix at each equilibrium and computing the eigenvalues (Table 2) reveals that all equilibria are characterized by at least one positive eigenvalue, except the two equilibria (0,0) (both players defect) and (1,1) (both players cooperate). These two latter equilibria are thus the only possible endpoints of the learning dynamics (Hirsch, Smale, & Devaney, 2004), for all possible solution orbits (Fig. 2a). There, we can see that the interior equilibrium ($B+C/2B$, $B+C/2B$) is an unstable node, i.e. both its eigenvalues are positive. Moreover, the two equilibria on the edges admit a stable and unstable manifold because they have one positive eigenvalue and one negative eigenvalue. All this implies that the equilibria (0,0) and (1,1) have a basin of attraction that is delimited by these stable manifolds. Since solutions along the nullcline defined by $\dot{p}_i=0$ verify $\dot{p}_j>0$ and solutions along the other nullcline ($\dot{p}_j=0$) verify $\dot{p}_i>0$, solving the inequalities $\dot{p}_i>0,\dot{p}_j>0$ for p_i and p_j (the region above the nullclines) gives a subset of the basin of attraction of (1,1). In other words, trajectories are increasing in this region and cannot escape it. Using Mathematica, we find that these inequalities are satisfied in two cases:

$$\begin{cases} \left(\dfrac{B}{2B-C} < p_{j,1} \leq \dfrac{B+C}{2B} \quad \text{and} \quad \dfrac{Cp_{j,1}}{-B+2Bp_{j,1}} < p_{i,1} < 1\right) \quad \text{or} \\ \left(\dfrac{B+C}{2B} < p_{j,1} < 1 \quad \text{and} \quad \dfrac{Bp_{j,1}}{-C+2Bp_{j,1}} < p_{i,1} < 1\right). \end{cases} \tag{A4.1}$$

PHR versus PHR. The dynamical system obtained by setting $g_i = 1$, $g_j = 1$, and the PD game payoffs into equations (5) and (6) admits the four corners of the state space [(0,1), (1,1), (1,0), (1,1)] as equilibria. The only locally stable equilibrium is the point (0,0) because it has negative eigenvalues $(-C,-C)$. Hence, two players using belief learning will end up always defecting (Fig. 2b).

PTR versus PHR. Setting $g_i = 0$, $g_j = 1$, and the PD game payoff into equations (5) and (6), we find five equilibria. The four corner equilibria and one on the edge $p_{\mathrm{HR}} = 1$ situated at $\left(\widehat{p}_{\mathrm{TR}} = \frac{B}{2B-C}, \widehat{p}_{\mathrm{HR}} = 1\right)$. The linearization shows that all equilibria are characterized by at least one positive eigenvalue, except the equilibrium (0,0) which has eigenvalues $(-C,0)$, implying, by elimination, that it is the only stable equilibrium. Both players will tend to defect in the long run (Fig. 2c).

Hawk–Dove Game

PTR versus PTR. In this case ($g_i = 0$, $g_j = 0$, $\mathscr{R} = B/2 - C$, $\mathscr{S} = B$, $\mathscr{T} = 0$, and $\mathscr{P} = B/2$), equations (5) and (6) have eight different equilibria. In addition to the four at the corners, we have two interior equilibria and two symmetric (with respect to the line $p_i = p_j$) equilibria on the edges $p_i = 0$ and $p_j = 0$ (Table 3). The vector field can be divided into three regions, each one being the basin of attraction of an asymptotically stable equilibrium. The first is the region where all trajectories tend to the equilibrium (0,0). This equilibrium has negative eigenvalues. Its basin of attraction is delimited by the stable manifolds of the equilibria situated on the edges, precisely situated at (0,1/3) and (1/3,0). The nullclines give a good approximation of the limits of this basin (Fig. 2d) and delimit a region where all the points verify that $\dot{p}_i < 0$, $\dot{p}_j < 0$, $p_{j,1} < \frac{B}{2B-\sqrt{2B(B-C)}}$, $p_{j,1} < \frac{B}{2B-\sqrt{2B(B-C)}}$. These are the points below the equilibrium $\left(\frac{B}{2B-\sqrt{2B(B-C)}}, \frac{B}{2B-\sqrt{2B(B-C)}}\right)$ and where the vector field points southwest. Excluding this specific region, all points below the diagonal line $p_j = p_i$ are in the basin of (0,1) and all points above this line pertain to the basin of (1,0). The points on this line $p_i = p_j$ (again excluding the points that are in the basin of (0,0)) are on the stable manifold of the interior equilibrium $\left(\frac{B}{2B-\sqrt{2B(B-C)}}, \frac{B}{2B-\sqrt{2B(B-C)}}\right)$.

PHR versus PHR. Here, equations (5) and (6) admit only two equilibria (0,1) and (1,0), which are asymptotically stable in the region below the diagonal line $p_i = p_j$ and above this line, respectively. This represents the stable manifold of the equilibrium ($B/2C$, $B/2C$). On $p_i = p_j$, we have the single population replicator dynamics (Fig. 2e), hence the stable point on this line is the ESS of the Hawk-Dove game, ($B/2C$, $B/2C$) (Weibull, 1997).

PTR versus PHR. We have six equilibria and three of them have at least one positive eigenvalue. We are left with (0,1), (1,0) and one interior at ($B/2C$,$3B - 2C/2B$). The latter equilibrium has eigenvalues $\left(-B + \frac{3B^2}{8C} + \frac{C}{2}, -\frac{B(B-2C)}{4C}\right)$, where the first one is always negative and the second one always positive. This equilibrium thus admits a stable manifold that splits the vector field into two regions: above the stable manifold, this is the basin of attraction of (1,0) and below it trajectories go to (0,1) (Fig. 2f). This stable manifold is a curve passing through the equilibria (0,1/3), ($B/2C$,$3B - 2C/2B$), and (1,1).

Coordination Game

This game provides the simplest dynamics, where the equilibria (0,0) and (1,1) are always the only two asymptotically stable states.

PTR versus PTR. Here, we set $g_i = 0$, $g_j = 0$, $\mathscr{R} = B$, $\mathscr{S} = 0$, $\mathscr{T} = 0$, and $\mathscr{P} = B$ in equations (5) and (6), which then admits four trivial corner equilibria plus all the points on the line $p_j = 1 - p_i$. The equilibria (0,0) and (1,1) both have negative eigenvalues and are thus locally stable. The two other equilibria in the corners ((0,1) and (1,0)) have eigenvalues (0,0). The equilibria on the line $p_j = 1 - p_j$ have eigenvalues $(0, 2Bp_i(1 - p_i))$, where the second eigenvalue is 0 when $p_i = 0$ or $p_i = 1$ and positive otherwise. This all implies that the equilibrium (0,0) is asymptotically stable in the region below the line $p_j = 1 - p_i$ and the equilibrium (1,1) is asymptotically stable above this line (Fig. 2g). Note that the unstable line $p_j = 1 - p_i$ is not an interesting set of initial conditions.

PHR versus PHR. The system in equations (5) and (6) admits five equilibria in this situation: the four corners and one interior at (½,½). The equilibria (0,0) and (1,1) are both asymptotically stable because they both have eigenvalues $(-B,-B)$ while the equilibria (0,1) and (1,0) have eigenvalues (B,B). The interior equilibrium (½,½) is a saddle with eigenvalues $(-B/2,B/2)$ and consequently admits a stable and an unstable manifold. It is easy to see that the stable manifold is the diagonal line $p_j = 1 - p_i$ while the unstable manifold is the other diagonal $p_j = p_i$ (Fig. 2h). Every trajectory starting above $p_j = 1 - p_i$ will tend to (1,1) while if it starts below this line it will tend to (0,0).

PTR versus PHR. Here, equations (5) and (6) have again five equilibria: the four corners plus one interior at (½,½). The points (0,0) and (1,1) are both asymptotically stable having eigenvalues $(-B,-B)$. The interior equilibrium is a saddle with eigenvalues $(-B/4,B/2)$. Hence a stable manifold passing through this saddle splits the vector field in two regions, which correspond respectively to the basin of attraction of (0,0) and (1,1). Here the stable manifold is no longer situated on the diagonal because we have lost the symmetry property of the PHR versus PHR case (compare Fig. 2i with Fig. 2h).

APPENDIX 5. SIMULATIONS

Individual-Based Simulations

Here, we present the algorithm of our individual-based simulations. Each individual $i \in \{1,2,\ldots,N\}$ takes a genotypic value $g_i \in \{0,1\}$. In every generation, each individual i obtains fecundity $\Pi_i = \sum_{t=1}^{T} \pi_i(a_{i,t}, \mathbf{a}_{-i,t}, \omega_t)$ (equation (1), but we do not use the normalization factor $1/T$ in the simulations), where the actions $(a_{i,t}, \mathbf{a}_{-i,t})$, which are random variables, are calculated by implementing equations (2) and (3). The environmental state in each period (ω_t) is drawn from a uniform distribution, which entails that $\mu(\mathrm{PD}) = \mu(\mathrm{HD}) = \mu(\mathrm{CG}) = 1/3$. The average game is parametrized according to the B and C parameters as in Table 4 (we always used $B = 5$ and $C = 3$), but the payoffs of the three subgames (PD, HD and CG) are randomly generated at the beginning of each generation so that they average to the desired average game and satisfy the inequalities described in Table 4. This implies that there are between-generation fluctuations, and we used them to represent the conditions where a learning ability gives an advantage over innate behaviour. Under one-shot matching, individuals were paired only at the beginning of the generation ($t = 1$) and each pair played

together until T, while under repeated matching individuals were rematched at each time period $t = 1,2,\ldots,T$.

The next generation is sampled with replacement according to the relative fecundity of individuals (i.e. $\Pi_i/\sum_{i=1}^{N}\Pi_i$, a Wright-Fisher process). An offspring inherits the genotype of its parent with probability $1-\eta$ or mutates with probability η to the other genotype. The mutation rate was set to $\eta = 10^{-3}$. We ran simulations with $N = 1000$ and we used a value of T bigger than the average time needed for learning to converge to a stable value (where this average time was computed separately for each game and initial condition, and we did not use values smaller than $T = 500$). Each case described in the main text (i.e. each game and each type of initial preferences of the learners) was run in three different replicates: one with an initial population (at the first generation) composed of half PTRs and half PHRs; one with an initial population of only PTRs; one with an initial population of only PHRs. This was to check that our simulation results are independent of initial conditions. Moreover, we waited for each run that the dynamic mean frequency of learning rules converge to a stable value, at which point we stopped the simulation. Our convergence criterion was met when the time average of the learning rules' frequencies did not change by more than 10^{-6} for 100 successive generations. We used $k = 0$ and varied λ from 10^{-1} to 10^{3} (see below).

We also carried out simulations only of the learning phase. To that aim, we implemented, in the one-shot matching model, pairs of individuals playing outside a population setting using the same parameters as in the full evolutionary simulations. We shall remark at this point that, contrary to the evolutionary simulations (which implement a Markov Chain admitting a stationary distribution), the learning dynamic is not ergodic so it does not admit a stationary distribution; stochastic approximation theory rather shows that the dynamics of learning will converge to one of the equilibrium points under the corresponding deterministic differential equation (see for example Borkar, 2008, Chapter 2, Corollary 4). In our simulations we observed convergence only to linearly stable equilibrium points (but note that, in general, stochastic approximation algorithms may not necessarily converge to a linearly stable equilibrium point, Pemantle, 1990). To understand the simulations, just note that the learning dynamics should converge to one of the equilibria but that it need not converge to the same equilibrium for two different replicates of a simulation, hence the necessity to run many replicates of the same parameter set.

For repeated matching, learning simulations consisted of running only one generation but setting the frequencies of PTR and PHR manually. We took the same cases (i.e. the three different average games (PD, HD and CG) and the different initial preferences over actions of the learners) as in the one-shot matching model and simulated learning behaviour for 11 different values of the frequency q of PTR in the population ranging from 0 to 1 by steps of 0.1.

Tuning Parameters

To find parameters that reproduce our analytical results, the process of running simulations consisted of two steps. First, we ran simulations between pairs of learners (PTR versus PTR, PTR versus PHR, and PHR versus PHR) for each game, in order to establish the accuracy of the approximation of the equilibrium action play probabilities. It has been previously found that the time t needed for the simulated learning process to converge to the predicted equilibrium critically depends on the sensitivity to motivations, λ, and on the initial difference between motivations of action 1 and 2, $\Delta M_{i,1} \equiv M_{i,1}(1) - M_{i,1}(2)$ at time $t = 1$ (Dridi & Lehmann, 2014). Since our analytic prediction is asymptotic, it tells nothing about the values of λ, $\Delta M_{i,1}$, and T. Hence, we first ran several simulations of learning with different values of λ and $\Delta M_{i,1}$, and waited for the learning process to 'converge' (based on a numerical convergence criterion). This gives us the parameter T needed for convergence to happen during the individual's lifetime. We then compared the equilibrium behaviour in these simulations with the predicted equilibrium, and chose the values of $\Delta M_{i,1}$ and λ that give the best match to prediction. To reduce our search in parameter space, we fixed the value of $p_{i,1}(a)$ to 0.85 for the initially preferred action a (see below) and only varied λ. This automatically changes the initial value of motivations $\Delta M_{i,1}$, such that we do not require to vary them explicitly. We used five different values of λ of the form 10^{α} for $\alpha = -1,0,1,2,3$. In the second step, we simulated the evolutionary process of selection on PTR versus PHR. To this end, we used the values of λ, ΔM, and T found in step 1 which give the best match between the stochastic learning process and deterministic approximation. The idea is that, since we chose parameters of learning where the approximation works well, the evolutionary simulations should also match the evolutionary predictions based on our approximation of learning. To further investigate the model under the alternative condition that ϕ_i is constant, we also performed the same set of simulations but with $\phi_i = 1$. This changes the type of the learning rules, which are now the counterpart to PTR and PHR when $\phi_i = 1$; that is, exploratory trial-and-error reinforcement (ETR) and hypothetical reinforcement learning (EHR). We used the parameter values for λ and $\Delta M_{i,1}$ applied in the case with dynamic learning rate.

We also explored parameter values that did not necessarily reproduce our analytical results. We report the results of such exploration in Appendix 8.

APPENDIX 6. DETAILED SIMULATION RESULTS

In this appendix, we describe in greater detail the results of our individual-based simulations.

One-shot Matching

Prisoner's Dilemma: dynamic learning rate

Both players initially prefer cooperation. Here, the simulations give results close to what is expected under our approximation when $\lambda = 10$. Regarding learning, we find the following results for the three possible interactions between learning rules. When PTR plays against PTR, most pairs learn to cooperate, but some pairs learn to defect (Fig. 3a): this is not surprising because under the stochastic learning process, individuals can escape basins of attraction and reach the locally stable equilibrium where both players defect. In the interaction between PHRs, all pairs learn to defect (Fig. 3b), as predicted by the analysis. In the interaction between PTR and PHR, both learning rules learn to defect: PTR does not get exploited by PHR (Fig. 3c). Overall this gives an advantage to PTR because this learning rule cooperates with itself but defects with the defector PHR. As a consequence, in the evolutionary simulations, PTR fixes in the population irrespective of the initial composition of the population (Table 5).

Both players initially prefer defection. In this case, we also find that $\lambda = 10$ gives the best fit to deterministic analysis. Surprisingly at first sight, we also observe that PTR individuals sometimes learn to cooperate when paired with themselves (Fig. 3d; while the analysis predicts that they will always defect in this case). This is actually perfectly possible, and is explained (as above) by the fact that initial conditions do not constrain absolutely the equilibrium behaviour: individuals with initial preference for defection can still learn to cooperate because this is also a stable equilibrium for the dynamics. PHR individuals do not deviate from perfect defection (Fig. 3e). For

the interaction between PTR and PHR, we find again that both learning rules learn to defect (Fig. 3f). Since learning behaviour is somehow similar to the case where individuals prefer cooperation, we find as expected that PTR fixes in the population when we simulate natural selection. This is due to the tendency of PTR to sometimes cooperate with itself (Table 5).

Prisoner's Dilemma: constant learning rate

Both players initially prefer Cooperation. In this situation, the results are very similar to the case with a dynamic learning rate. Namely, ETR is able to learn to cooperate against itself, a behaviour that EHR cannot express, and this gives a fitness advantage to ETR because ETR learns to defect against EHR (Fig. A1a, b, c). As a consequence, the frequency of ETR at an evolutionary equilibrium is very close to 1 (Table 5).

Hawk–Dove game: dynamic learning rate

PTR initially prefers to play Hawk and PHR prefers Dove. In the analysis, we predict that with these initial preferences, PTR individuals will learn to play half of the time Hawk and half of the time Dove when paired against themselves. However, in the simulations, we observe that a high proportion of PTR learned to play Dove (Fig. 4a; the best value found for sensitivity is here $\lambda = 10$). As before, this can be explained by the possibility of escaping a basin of attraction: the outcome (Dove, Dove) is also an equilibrium for the dynamics and some pairs of individuals converge to this equilibrium. When a PHR plays against a PHR, we find as predicted that approximatively half PHRs learn Hawk and half learn Dove (Fig. 4b). Finally, for PTR versus PHR, things go as predicted with a vast majority of PTR learning Hawk and a vast majority of PHR learning Dove: PHR gets 'exploited' by PTR here (Fig. 4c).

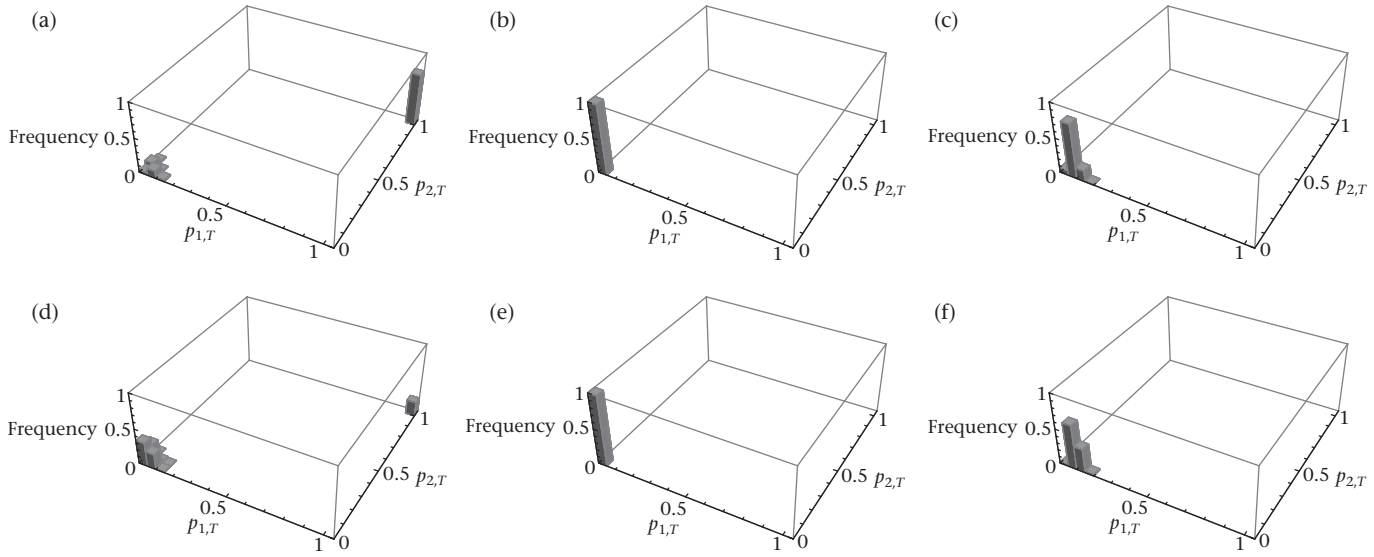

Figure A1. Distribution of behaviour at equilibrium of learning in the average Prisoner's Dilemma for the one-shot matching model with constant learning rate for pairs of opponents. This represents the frequency of pairs having reached the probability of playing action 1 ($p_{1,T}$, $p_{2,T}$) at the end of the individual's lifetime, T. We used a total of 1000 individuals of each learning rule in each simulation. (a, b, c) Initial preference for Cooperation ($p_{i,1} = 0.85$). (d, e, f) Initial preference for Defection ($p_{i,1} = 0.15$). (a, d) Interaction between two TRs. (b, e) Interaction between two HRs. (c, f) Interaction between TR (player 1) and HR (player 2).

Both players initially prefer defection. Here ETR individuals display the same behaviour as with a dynamic learning rate when paired together: namely some pairs learn to defect and others learn to cooperate (Fig. A1d). EHR individuals on the other hand, still learn to defect whatever their opponent is (Fig. A1e). The interactions between ETR and EHR display a different outcome than previously. Here we observe that ETR individuals converge to a state where they have a positive probability of cooperating, and hence get exploited on some interaction rounds (Fig. A1f).

As a consequence of this learning behaviour, the evolutionary simulations show that EHR fixes in the population in the long run when they are initially at high frequency but otherwise this is ETR that invades (Table 5). Such a result is possible if there is an interior unstable equilibrium in the evolutionary dynamics: when ETR are common in the population, they have a tendency to increase in frequency; when they are at low frequency they have a tendency to further decrease in frequency. This situation corresponds to observed learning behaviour: while ETR individuals have the advantage of cooperating with themselves, this does not seem to compensate the for fitness loss due to sometimes cooperating against the defector EHR when EHR constitutes most of the population.

When we run simulations of natural selection in a population of PTR and PHR we obtain that PTR fixes for all initial compositions of the population (Table 5). Since learning behaviour is in conformity with our analytic prediction, this is not a surprise. The important interaction between PTR and PHR turns to the advantage of PTR. The latter is more prompt to learn the Hawk strategy and PHR is penalized by its initial preference for Dove. One unpredicted outcome of learning, namely the fact that PTR will learn to play Dove against itself even if it initially prefers Hawk, gives no special advantage to PTR with the payoff structure of our Hawk–Dove game.

PTR initially prefers to play Dove and PHR prefers Hawk. This initial condition is mirroring the previous case, and the analysis thus predicts that PHR should be the one that learns to play Hawk against PTR (the sensitivity that gives the best match to prediction is $\lambda = 100$ for this case). This is indeed what we observe (Fig. 4f). For the interactions between the same learning rules (PTR versus PTR and PHR versus PHR), we have the same behaviour as before: most PTR learn to play Dove (Fig. 4d), and PHR learn half of the time to play Hawk and half of the time to play Dove (Fig. 4e). With this learning behaviour, PHR invades and fixes in the population for all

initial compositions of the population, and this happens very fast (in the first generations; Table 5).

Hawk–Dove game: constant learning rate

PTR initially prefers to play Hawk and PHR prefers Dove. In this situation, we observe qualitatively the same learning behaviours as with a dynamic learning rate. In particular, ETR learns Hawk against EHR and the latter learns Dove (Fig. 2a, b, c). As a consequence, ETR fixes to a frequency of 1 at an evolutionary equilibrium (Table 5).

PTR initially prefers to play Dove and PHR prefers Hawk. The result is again very similar to the case with dynamic learning rate. Namely, in ETR versus HR interactions, ETR learn Dove and EHR learn Hawk (Fig. A2d, e, f) and this implies that EHR rapidly take over and fixes in the population under the evolutionary simulations (Table 5).

mostly on the Right action (Fig. 5c). This result is difficult to explain because individuals had an initial preference for 'Left' but since the analysis demonstrates that (Right, Right) is also a stable equilibrium of the associated deterministic system, this result does not contradict our qualitative analysis.

Interestingly, the evolutionary simulations give an outcome different than what we expected. While the above learning behaviour suggests that both learning rules should coexist in equal frequency in the long-run, we find that the population converges to a mixed state with domination of PHR individuals (Table 5). Again, it is difficult to know why this happened, but a possible reason for this might be that some PTR individuals converge more slowly to the equilibrium. Even if we chose T big enough, our criterion was based on the average time needed for all individuals to converge in the population. Some individuals might converge more slowly than in T time steps, and fail to coordinate at this time, giving an advantage to PHR.

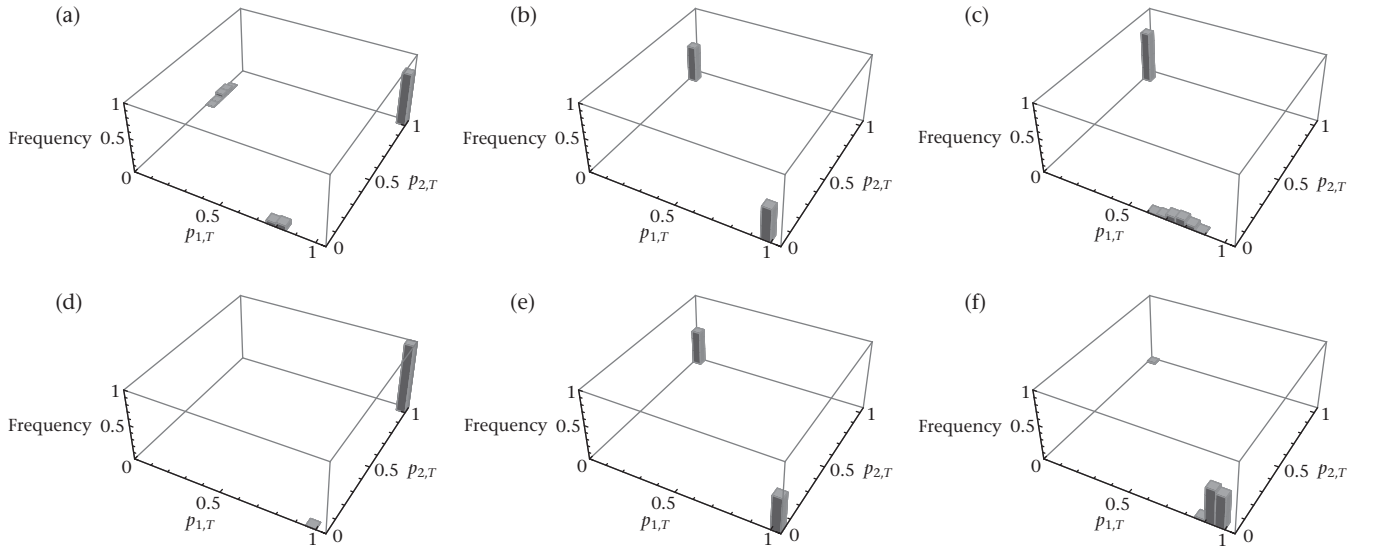

Figure A2. Same as Fig. A1 but for the average Hawk–Dove game in the one-shot matching model with constant learning rate. (a, b, c) TR initially prefers Hawk ($p_{TR,1} = 0.15$) and HR prefers Dove ($p_{HR,1} = 0.85$). (d, e, f) TR initially prefers Dove ($p_{TR,1} = 0.85$) and HR prefers Hawk ($p_{HR,1} = 0.15$). (a, d) Interaction between two TRs. (b, e) Interaction between two HRs. (c, f) Interaction between TR (player 1) and HR (player 2).

Coordination game: dynamic learning rate

In this average game, we give to all individuals a preference for the 'Left' action and find that all individuals succeed in coordinating at time T. Pairs of PTR individuals coordinate mostly on action 2 ('Right') while pairs of PHR coordinate mostly on action 1 ('Left') (Fig. 5a, b). The heterogeneous pairs (PTR versus PHR) coordinate

Coordination game: constant learning rate

The results of learning dynamics under constant learning rate are very similar to the above, with ETR pairs coordinating on Right, EHR pairs coordinating on Left and heterogeneous pairs coordinating on Right (Fig. A3a, b, c).

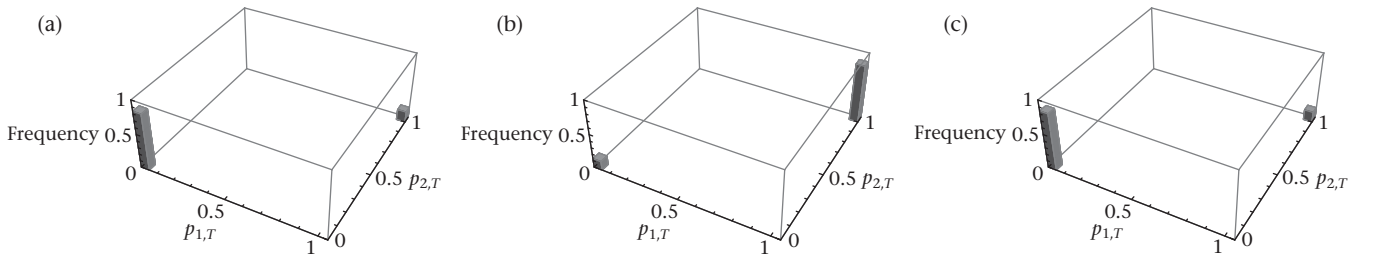

Figure A3. Distribution of behaviour at equilibrium of learning in the average Coordination game in the one-shot matching model with constant learning rate. (a) Interaction between two TRs. (b) Interaction between two HRs. (c) Interaction between TR (player 1) and HR (player 2).

Regarding evolution, the result is in conformity to our analysis since the population converges to a mixed state where the frequency of both learning rules is close to 0.5 but with a slight domination of ETR individuals ($q \approx 0.56$; Table 5).

Repeated Matching

Prisoner's Dilemma: dynamic learning rate

All individuals initially prefer cooperation. In the average PD game, we use the value $\lambda = 10$ that gives the best correspondence between analysis and simulation in the one-shot matching (OM) model. We find that PHR learns to defect irrespective of the frequency of the learning rules in the population. However, PTR individuals can learn to cooperate when at sufficiently high frequency in the population (precisely, the average probability of cooperating at equilibrium of learning of PTR is above 0 for $q \geq 0.8$; Fig. 6a). This result is not surprising given the OM results. Indeed, when at high frequency, there is a positive probability that a PTR individual meets only other PTRs and the dynamics of those individuals will likely be similar to the dynamics of pairs of PTR in the OM model.

This learning behaviour implies that evolution leads to an interior equilibrium with the coexistence of PTR and PHR. Indeed, when PTR are at lower frequency than 0.8, evolution is neutral because everybody defects. However, when the population reaches a state where the frequency of PTR is above 0.8, PTR starts to cooperate so the latter has a disadvantage compared to PHR and thus has a tendency to decrease in frequency. The population thus visits all the states such that the frequency of PTR is less than 0.8 equally often, and in our evolutionary simulations we obtain a stable average frequency of PTR around $q \approx 0.19$ (Table 6).

All individuals initially prefer defection. In this game and with these initial preferences, individuals of both learning rules learn to defect for all frequencies of PTR in the population (Fig. 6b).

This leads us to think that evolution is neutral here and that the evolutionary simulations should converge to a state where $q = 0.5$. Our simulations give a result close to this prediction but we also find that PHR slightly dominates the population at an evolutionary equilibrium (Table 6).

Prisoner's Dilemma: constant learning rate

All individuals initially prefer cooperation. At all frequencies, ETR individuals converge to a state where their average probability of cooperation is above 0, which gives a complete advantage to the defector EHR in this repeated matching model (Fig. 6c). As a consequence, the evolutionary dynamics display a state where the domination of EHR is almost total as they nearly fix in the population at an evolutionary equilibrium (Table 6).

All individuals initially prefer Defection.

In the same vein as when individuals initially prefer Cooperation, ETR players also converge to an average probability of Cooperation that is positive for all values of q, while EHR always learn to defect (Fig. 6d). This situation makes EHR almost fix in the population in an evolutionary long run (Table 6).

Hawk–Dove game: dynamic learning rate

ETR initially prefers to play Hawk and EHR prefers Dove. Even though we implement an initial preference for Hawk to PTR individuals, we find here that, in their learning behaviour, PTR converge to a state where their probability of playing Dove is always higher than that of PHR. Moreover, PTR have a tendency to increase their probability of playing Dove as they increase in frequency while we observe the inverse tendency among PHR individuals, which decrease their learned probability of playing Dove as the frequency of PTR increases (Fig. 6e).

As a result, the evolutionary simulations of natural selection give the result that both learning rules coexist in the long run with a domination of PHR individuals. Indeed, we observe that at low frequencies q, PHR are playing Hawk a little more often than prescribed by the ESS (which is here $1 - B/2C \approx 0.17$ because we use $B = 5$ and $C = 3$), and PTR are playing Dove with a high average probability. This gives an advantage to PTR at low frequencies since they perform better against PHR than PHR perform against themselves (because they play Hawk too often). However, when the frequency of PTR increases, the latter gets exploited more often by PHR because PTR plays more and more Dove while PHR plays more and more Hawk, which gives an advantage to PHR. Hence at high enough frequencies of PTR, PHR has a higher fitness than PTR. This explains the interior equilibrium with a domination of PHR in the evolutionary simulations (Table 6).

ETR initially prefers to play Dove and EHR prefers Hawk. This initial condition was favouring PHR in the one-shot matching model, and we have the same situation here. We observe that PTR learns to play Dove with high average probability (not smaller than 0.6) for all q, while PHR learns to play Hawk with a high average probability, and this probability even decreases as PTR increases in frequency (Fig. 6f).

Hence, PHR obtains a higher payoff against PTR than PTR against PTR which gives it an advantage for high q. However, for low q, PTR against PHR cannot obtain a much better payoff than PHR against PHR since the latter plays close to the ESS at low q. This is why we observe in the evolutionary simulations that PHR dominates the population in an interior equilibrium (Table 6).

Hawk–Dove game: constant learning rate

ETR initially prefers to play Hawk and EHR prefers Dove. The results for this case resemble the case with dynamic learning rate above. Namely, ETR individuals have for all q a learned probability of playing Dove above that of PHR. Moreover, the probability of playing Dove of ETR increases as q increases but this probability decreases for EHR (Fig. 6g).

For the same reason as in the case with dynamic learning rate, natural selection leads to an interior equilibrium with a large domination of PHR (Table 6).

ETR initially prefers to play Dove and EHR prefers Hawk. Again this situation is very similar to the one with dynamic learning rate above. ETR always has an average probability of playing Dove higher than PHR, while the latter plays close to the ESS when frequent in the population and plays almost always Hawk when rare (Fig. 6h).

This learning behaviour favours EHR over ETR and simulations of evolution confirm this by showing that the frequency of ETR at evolutionary equilibrium is around $q \approx 0.05$ (Table 6).

Coordination game: dynamic learning rate

In this game, PTR has difficulties in coordinating on the same equilibrium as PHR for all frequencies q, while PHR always succeeds in learning to coordinate on a single action (that can be Left or Right depending on stochastic events in the simulations; Fig. 6i).

This learning behaviour implies that PTR has lower fitness for all q. Indeed simulations of natural selection show that PHR fix in the population in the long run (Table 6).

Coordination game: constant learning rate

In this case, the learning behaviour of EHR is similar to the situation with dynamic learning rate, namely, all EHR learn to coordinate on a single action. On the other hand, ETR still have

difficulties coordinating for sufficiently high q, but coordinate efficiently for low q (Fig. 6j).

This learning behaviour implies that evolution should be neutral for low q but should favour EHR for higher q. This is indeed consistent with what we obtain when we simulate natural selection, where we observe an interior equilibrium with a large domination of EHR ($q \approx 0.06$; Table 6).

APPENDIX 7. WIN-STAY, LOSE-SHIFT FROM EWA

Here, we show that the automaton strategy win-stay, lose-shift (WSLS) can be obtained from EWA (Camerer & Ho, 1999), which confirms the verbal explanation given by Nowak and Sigmund (1993) arguing that WSLS is a payoff-based strategy in the spirit of trial-and-error reinforcement learning (but see Stephens and Clements (1998) on the misuse of the name 'Pavlov' to designate WSLS). Essentially, WSLS repeats its last action if the payoff received is above a certain aspiration level, but WSLS switches action if the payoff is below that aspiration level. Let L be such an aspiration level, which must satisfy $\mathscr{T} > \mathscr{R} > L > \mathscr{P} > \mathscr{S}$. We will show that EWA with parameters $g = 0$, $\phi = \rho = 0$, $\lambda = \infty$, and L subtracted from motivations can indeed produce the WSLS behaviour. The updating rule for motivations with these parameters in a pairwise interaction is

$$M_{i,t+1}(a) = 1(a, a_{i,t})\left(\pi_i(a, a_{-i,t}, \omega_t) - L\right). \tag{A7.1}$$

Consider two players engaged in the repeated play of the Prisoner's Dilemma with payoff matrix

$$\begin{pmatrix} \mathscr{R} & \mathscr{S} \\ \mathscr{T} & \mathscr{P} \end{pmatrix}.$$

Let us focus on player 1 which will be called player i.

Player i plays according to WSLS if

$$a_{i,t+1} = \begin{cases} \mathrm{C}, & \text{if } \mathbf{a}_t = (\mathrm{C},\mathrm{C}), \\ \mathrm{C}, & \text{if } \mathbf{a}_t = (\mathrm{D},\mathrm{D}), \\ \mathrm{D}, & \text{if } \mathbf{a}_t = (\mathrm{C},\mathrm{D}), \\ \mathrm{D}, & \text{if } \mathbf{a}_t = (\mathrm{D},\mathrm{C}), \end{cases} \tag{A7.2}$$

where $\mathbf{a}_t = (a_{i,t}, a_{-i,t})$. In terms of motivations under EWA, this means that we must have, when $\lambda = \infty$,

$$\begin{cases} M_{i,t+1}(\mathrm{C}) > M_{i,t+1}(\mathrm{D}), & \text{if } \mathbf{a}_t = (\mathrm{C},\mathrm{C}), \\ M_{i,t+1}(\mathrm{C}) > M_{i,t+1}(\mathrm{D}), & \text{if } \mathbf{a}_t = (\mathrm{D},\mathrm{D}), \\ M_{i,t+1}(\mathrm{C}) < M_{i,t+1}(\mathrm{D}), & \text{if } \mathbf{a}_t = (\mathrm{C},\mathrm{D}), \\ M_{i,t+1}(\mathrm{C}) < M_{i,t+1}(\mathrm{D}), & \text{if } \mathbf{a}_t = (\mathrm{D},\mathrm{C}). \end{cases} \tag{A7.3}$$

Substituting the definition of the motivation (equation (A7.1)), we obtain

$$\begin{cases} \mathscr{R} - L > 0, & \text{if } \mathbf{a}_t = (\mathrm{C},\mathrm{C}), \\ 0 > \mathscr{P} - L, & \text{if } \mathbf{a}_t = (\mathrm{D},\mathrm{D}), \\ \mathscr{S} - L < 0, & \text{if } \mathbf{a}_t = (\mathrm{C},\mathrm{D}), \\ 0 < \mathscr{T} - L, & \text{if } \mathbf{a}_t = (\mathrm{D},\mathrm{C}), \end{cases} \tag{A7.4}$$

which effectively gives

$$\mathscr{T} > \mathscr{R} > L > \mathscr{P} > \mathscr{S}, \tag{A7.5}$$

as required by the definition of WSLS given above.

APPENDIX 8. ROBUSTNESS TO CHANGES IN λ AND T

In this appendix, we display results of simulations for other values of λ and T, in order to test the robustness of our results (Tables A1–A10) and Figs. A4–A6.

Table A1
Summary of results in the one-shot matching model for $\lambda = 1$ (otherwise identical to Table 5)

Average Game	Initial condition of learning	Predicted q^*	Learning rate	Simulated q^*
Prisoner's Dilemma	Basin of (Cooperate, Cooperate) of TRvsTR	1	Dynamic	0.97
			Constant	0.02
	Basin of (Defect, Defect) of TRvsTR	1/2	Dynamic	0.09
			Constant	0.02
Hawk-Dove Game	Basin of (Hawk, Dove) of TRvsHR	1	Dynamic	0.99
			Constant	0.99
	Basin of (Dove, Hawk) of TRvsHR	0	Dynamic	0
			Constant	0.03
Coordination Game	Basin of (Left, Left)	1/2	Dynamic	0.52
			Constant	0.61

Table A2
Summary of results in the one-shot matching model for $\lambda = 100$ (otherwise identical to Table 5)

Average Game	Initial condition of learning	Predicted q^*	Learning rate	Simulated q^*
Prisoner's Dilemma	Basin of (Cooperate, Cooperate) of TRvsTR	1	Dynamic	0.99
			Constant	0.99
	Basin of (Defect, Defect) of TRvsTR	1/2	Dynamic	0.98
			Constant	0.98
Hawk-Dove Game	Basin of (Hawk, Dove) of TRvsHR	1	Dynamic	0.97
			Constant	0.22
	Basin of (Dove, Hawk) of TRvsHR	0	Dynamic	0
			Constant	0
Coordination Game	Basin of (Left, Left)	1/2	Dynamic	0.21
			Constant	0.56

Table A3
Summary of results in the repeated matching model for $\lambda = 1$ (otherwise identical to Table 6)

Average Game	Initial condition of learning	Learning rate	Simulated q^*
Prisoner's Dilemma	All individuals prefer Cooperation	Dynamic	0.04
		Constant	0.02
	All individuals prefer Defection	Dynamic	0.1
		Constant	0.02
Hawk-Dove Game	TR prefers Hawk, HR prefers Dove	Dynamic	0.82
		Constant	0.65
	TR prefers Dove, HR prefers Hawk	Dynamic	0.07
		Constant	0.25
Coordination Game	All individuals prefer Left	Dynamic	0.02
		Constant	0.25

Table A4
Summary of results in the repeated matching model for $\lambda = 100$ (otherwise identical to Table 6)

Average Game	Initial condition of learning	Learning rate	Simulated q^*
Prisoner's Dilemma	All individuals prefer Cooperation	Dynamic	0.36
		Constant	0.26
	All individuals prefer Defection	Dynamic	0.47
		Constant	0.24
Hawk-Dove Game	TR prefers Hawk, HR prefers Dove	Dynamic	0.16
		Constant	0.15
	TR prefers Dove, HR prefers Hawk	Dynamic	0.08
		Constant	0.09
Coordination Game	All individuals prefer Left	Dynamic	0.01
		Constant	0.02

Table A5
Summary of results in the one-shot matching model for $\lambda = 1$ and $T = 50$ (otherwise identical to Table 5)

Average Game	Initial condition of learning	Predicted q^*	Learning rate	Simulated q^*
Prisoner's Dilemma	Basin of (Cooperate, Cooperate) of TRvsTR	1	Dynamic	0.02
			Constant	0.02
	Basin of (Defect, Defect) of TRvsTR	1/2	Dynamic	0.06
			Constant	0.02
Hawk-Dove Game	Basin of (Hawk, Dove) of TRvsHR	1	Dynamic	0.99
			Constant	0.99
	Basin of (Dove, Hawk) of TRvsHR	0	Dynamic	0.01
			Constant	0.01
Coordination Game	Basin of (Left, Left)	1/2	Dynamic	0.55
			Constant	0.74

Table A6
Summary of results in the one-shot matching model for $\lambda = 10$ and $T = 50$ (otherwise identical to Table 5)

Average Game	Initial condition of learning	Predicted q^*	Learning rate	Simulated q^*
Prisoner's Dilemma	Basin of (Cooperate, Cooperate) of TRvsTR	1	Dynamic	0.99
			Constant	0.98
	Basin of (Defect, Defect) of TRvsTR	1/2	Dynamic	0.96
			Constant	0.60
Hawk-Dove Game	Basin of (Hawk, Dove) of TRvsHR	1	Dynamic	0.99
			Constant	0.98
	Basin of (Dove, Hawk) of TRvsHR	0	Dynamic	0
			Constant	0
Coordination Game	Basin of (Left, Left)	1/2	Dynamic	0.37
			Constant	0.60

Table A7
Summary of results in the one-shot matching model for $\lambda = 100$ and $T = 50$ (otherwise identical to Table 5)

Average Game	Initial condition of learning	Predicted q^*	Learning rate	Simulated q^*
Prisoner's Dilemma	Basin of (Cooperate, Cooperate) of TRvsTR	1	Dynamic	0.99
			Constant	0.99
	Basin of (Defect, Defect) of TRvsTR	1/2	Dynamic	0.98
			Constant	0.98
Hawk-Dove Game	Basin of (Hawk, Dove) of TRvsHR	1	Dynamic	0.97
			Constant	0.93
	Basin of (Dove, Hawk) of TRvsHR	0	Dynamic	0.01
			Constant	0.01
Coordination Game	Basin of (Left, Left)	1/2	Dynamic	0.53
			Constant	0.70

Table A8
Summary of results in the repeated matching model for $\lambda = 1$ and $T = 50$ (otherwise identical to Table 6)

Average Game	Initial condition of learning	Learning rate	Simulated q^*
Prisoner's Dilemma	All individuals prefer Cooperation	Dynamic	0.01
		Constant	0.02
	All individuals prefer Defection	Dynamic	0.06
		Constant	0.02
Hawk-Dove Game	TR prefers Hawk, HR prefers Dove	Dynamic	0.91
		Constant	0.67
	TR prefers Dove, HR prefers Hawk	Dynamic	0.05
		Constant	0.07
Coordination Game	All individuals prefer Left	Dynamic	0.03
		Constant	0.05

Table A9
Summary of results in the repeated matching model for $\lambda = 10$ and $T = 50$ (otherwise identical to Table 6)

Average Game	Initial condition of learning	Learning rate	Simulated q^*
Prisoner's Dilemma	All individuals prefer Cooperation	Dynamic	0.11
		Constant	0.05
	All individuals prefer Defection	Dynamic	0.21
		Constant	0.05
Hawk-Dove Game	TR prefers Hawk, HR prefers Dove	Dynamic	0.17
		Constant	0.10
	TR prefers Dove, HR prefers Hawk	Dynamic	0.04
		Constant	0.04
Coordination Game	All individuals prefer Left	Dynamic	0.02
		Constant	0.02

Table A10
Summary of results in the repeated matching model for $\lambda = 100$ and $T = 50$ (otherwise identical to Table 6)

Average Game	Initial condition of learning	Learning rate	Simulated q^*
Prisoner's Dilemma	All individuals prefer Cooperation	Dynamic	0.15
		Constant	0.22
	All individuals prefer Defection	Dynamic	0.23
		Constant	0.24
Hawk-Dove Game	TR prefers Hawk, HR prefers Dove	Dynamic	0.15
		Constant	0.09
	TR prefers Dove, HR prefers Hawk	Dynamic	0.04
		Constant	0.04
Coordination Game	All individuals prefer Left	Dynamic	0.01
		Constant	0.01

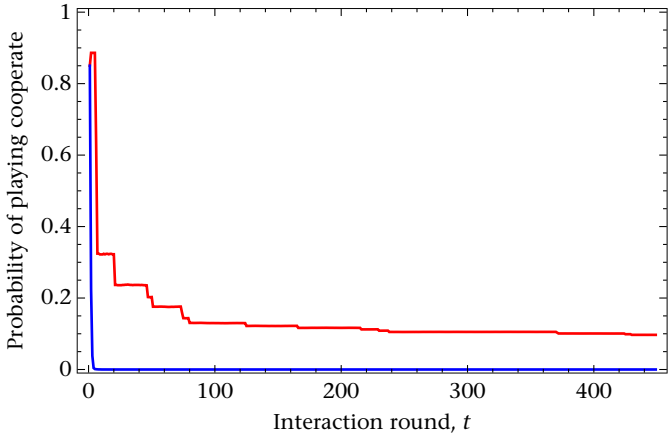

Figure A4. Behavioural dynamics of cooperation in the average Prisoner's Dilemma for a typical pair of HR (blue line) versus TR (red line) ($\varphi_{i,t} = (1/t)+1$) in the one-shot matching model with initial preference for cooperation ($p_{i,1} = 0.85$ for both players). Parameter values: $T = 500$, $\lambda = 1$ for both players, $B = 5$, $C = 3$, $\mu(\omega) = 1/3$ for all $\omega \in$ {PD, HD, CG}.

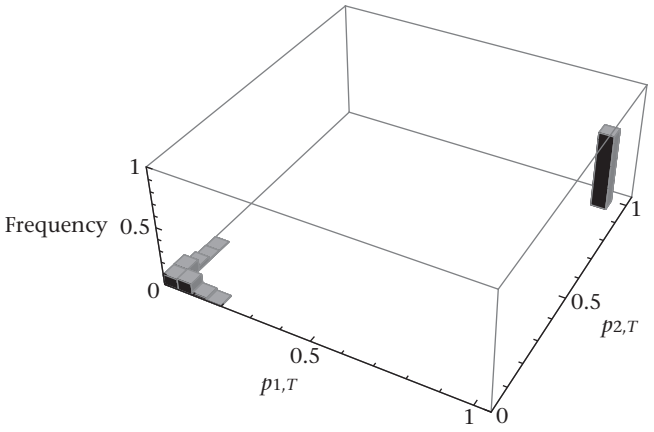

Figure A5. Distribution of behaviour at equilibrium of learning in the average Prisoner's Dilemma for the one-shot matching model with dynamic learning rate for pairs of TR individuals. Same as Fig. 3a but with $\lambda = 1$.

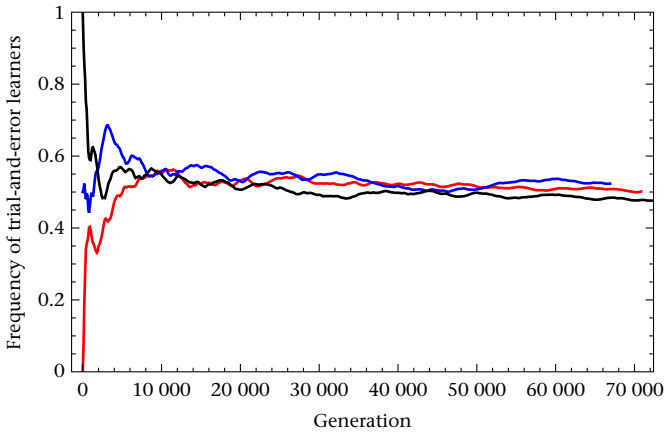

Figure A6. Evolutionary dynamics of the frequency of reinforcement learners in the one-shot matching model when the game is constant and corresponds to a Coordination Ggame. We plot the time average of the frequency of reinforcement learners at each generation. The three different lines represent distinct simulations where we used different initial compositions of the population. In red, the population is initially composed of HR only, in black of TR only and in blue of half TR and half HR.